PH142 Fall 2021 Final Exam

The exam is open book. This means you can use electronic or hard copies of all class materials and datahub if you wish. You may not use the internet to search for the answers or inform your answers. Using the internet is strictly prohibited and any evidence of this may result in a 0 on the exam.

While you take the exam, you are prohibited from discussing the test with anyone. If you are taking the test after your classmates, you are also prohibited from talking to them about the test before you take it. Evidence of cheating may result in a 0 on the exam and may be reported to the Student Conduct Board.

Type your initials to affirm that you have read and agree to the above statements:

Berkeley's code of conduct is here: https://sa.berkeley.edu/code-of-conduct. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct (in particular the sections on cheating and plagiarism).

NOTE: Question 1 is the affirmation statement. The actual exam begins with question 2. Be sure to put all of your answers Gradescope - we will not be grading the Rmd for the datahub questions.

Questions 2-16: 15 points

Question 17: 2 points

Question 18: 5 points

Question 19: 8 points

Question 20: 6 points

Question 21: 12 points

Question 22: 13 points

Question 23: 8 points

Question 24 (Datahub): 13 points

Question 25 (Datahub): 8 points

Total: 90 points

Questions 2-16 [15 points total]

2. [1 point] We use a one-sample z test instead of the one-sample t test when _____.

(a) μ is known
(b) μ is unknown
(c) σ is known
(d) σ is unknown

```
## SOLUTION: (c) sigma is known
```

3. [1 point] Which of the following are possible null hypotheses for a chi-square goodness of fit test? Select all that apply.

(a) $H_0 : p_{asian} = 17.1\%,\ p_{black} = 10.3\%,\ p_{latine} = 25.1\%,\ p_{white} = 42.2\%,\ \ p_{other} = 5.3\%$

(b) $H_0 : p_{asian} = p_{black} = p_{latine} = p_{white} = p_{other} = 20\%$

(c) $H_0 : p_{white} = p_{black} = 50\%,\ p_{latine} = p_{asian} = p_{other} = 50\%$

(d) $H_0 : p_{white} = 25\%,\ p_{black} = 25\%,\ p_{latine} = 25\%,\ p_{asian} = 25\%,\ p_{other} = 25\%$

```
## SOLUTION: (a),(b)
# (d) is not correct because it sums to > 100%

# 0.5 point per correct answer
```

4. [1 point] When we are performing a permutation test, it is possible that in the first permutation we have 9 data points with treatment labels and 1 with a control label; in the second permutation we have 1 data point with a treatment label and 9 with control labels.

(a) True
(b) False

```
## SOLUTION:
# (b) False
```

5. [1 point] Suppose that you are comparing the mean values of a quantitative variable for 15 groups. Instead of using Tukey's HSD test, you decide to do a two-sample t-test for every possible pair of two variables to see which are statistically different. What is the problem you will likely encounter?

(a) You will find some low p-values just by chance
(b) The computation time will likely be an issue (unless you use a supercomputer)
(c) You will violate the assumptions of a two-sample t-test
(d) You can only check for differences in means when we have exactly 2 groups, not 15.

```
## SOLUTION:
# (a) You will find some low p-values just by chance
```

6. [1 point] The chi-square goodness of fit test and the chi-square test of independence both test the same null hypothesis.

(a) True
(b) False

```
## SOLUTION:
# (b) False
# The goodness of fit test tests "how well do the expect counts 'fit' the observed counts"
# and the independence test tests if the explanatory and response variable are independent.
```

7. [1 point] Under what conditions is it appropriate to apply the plus 4 method for inference of a proportion? Let n denote the sample size, and α denote the significance level.

(a) $n \geq 10, \alpha \leq 0.05$
(b) $n \geq 10, \alpha \leq 0.10$
(c) $n \geq 30, \alpha \leq 0.05$
(d) $n \geq 30, \alpha \leq 0.10$
(e) $n \geq 30$, any α

```
## SOLUTION:
# (b) n $\geq$ 10, $\alpha \leq 0.10$.
# The plus 4 method can be used when n is at least 10 and the confidence level is at least 90%.
# This means that the significance level alpha is no more than 0.10.
```

8. [1 point] The F statistic is a ratio of the _____.

(a) Probability of the null hypothesis being true and the probability of the alternative hypothesis being true
(b) Largest mean among the groups and the smallest mean among the groups
(c) Largest standard deviation among the groups and the smallest standard deviation among the groups
(d) Variation between each of the group means and the variation between individuals in the same group

```
## SOLUTION:
# (d) Variation between each of the group means and the variation between
# individuals in the same group
```

9. [1 point] A Public Health PhD student is working with data on hospital insurance premium rates from patients in Los Angeles (LA) County. He wants to determine if there is a difference in median premium rates (per individual) across different cities within LA County, but is struggling with what test he should use. The following table provides information on the cities he is including in his analysis.

County	City	n	Median Premium ($)
Los Angeles	Los Angeles	15	352
Los Angeles	Santa Monica	10	459
Los Angeles	Culver City	10	273
Los Angeles	Pasadena	9	574
Los Angeles	Beverly Hills	12	597
Los Angeles	Van Nuys	14	295

Note that these data were collected from a simple random sample.

Assuming that you were also given the premium rate data for each individual, select the test is most appropriate for analyzing the differences in median premium rates between cities in LA County.

(a) Two-Sample T-Test
(b) Paired T-Test
(c) ANOVA
(d) Wilcox Rank-Sum
(e) Wilcox Sign-Rank
(f) Kruskal Wallis

```
## SOLUTION:

# (f) Kruskal Wallis

# Note that we are only given the medians of the premium rates, not the mean.
# Therefore, Kruskal Wallis is more appropriate to use than ANOVA.
```

10. [1 point] Suppose you want to make inference about a population proportion and need to use a test that is statistically conservative. Which method should you use?

(a) "Exact" or Clopper Pearson
(b) Wilson Score
(c) Large Sample Confidence Interval
(d) None of the above

```
## SOLUTION:
# (a) "Exact" or Clopper Pearson
# most statistically conservative, meaning it has the best coverage, of the three options.
```

11. [1 point] A student is conducting a Wilcoxon Sign-Rank test manually. For one of their observations, if the two variables they are comparing have the same value (i.e. difference = 0), the student should still assign that observation a rank.

(a) True
(b) False

```
## SOLUTION:
# (b) False
# The student should just drop it instead.
```

12. [1 point] Suppose `sample_median` is a vector that contains 100 medians calculated from bootstrap samples. To obtain the lower bound of a 95% confidence interval for the median, we would use `quantile(sample_median, 0.05)`.

(a) True
(b) False

```
## SOLUTION:
# (b) False
```

13. [1 point] When we are estimating the standard error based on a sample, $s/\sqrt{n}$ estimates the variation between individuals.

(a) True
(b) False

```
## SOLUTION:
# (b) False
# $$s/\sqrt{n}$$ estimates how much sample means vary if we take many multiple samples
# instead of the variation between individuals
```

14. [1 point] We cannot use the Bootstrap to construct a confidence interval for the minimum value because the Central Limit Theorem does not apply to minimums.

(a) True
(b) False

```
## SOLUTION:
# (b) False
```

15. [1 point] Given a fitted model `fit0`, we want to make inference about a new observation's response. If the value of the new observation is stored in a data frame called `new_obs`, which of the following is the correct way to put bounds on this value?

(a) `predict(fit0, newdata = new_obs, interval = "confidence")`
(b) `predict(fit0, newdata = new_obs, interval = "predict")`
(c) `predict(fit0, newdata = new_obs)`
(d) `predict(fit0, newdata = new_obs, interval = "response")`

```
## SOLUTION:
# (b) `predict(fit0, newdata = new_obs, interval = "predict")`
```

16. [1 point] In the linear regression $y = \alpha + \beta x$, we want to test $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$. If we have a sample with 40 observations, the test statistic will follow a t-distribution with _____ degrees of freedom.

(a) 42
(b) 41
(c) 39
(d) 38

```
## SOLUTION:
# (d) 38
```

Question 17 [2 points total]

You are given paired data for a one sample t test. When analyzing the data, you accidentally set `paired = F`.

17.1 [1 point] How would the p-value for the test you conducted compare to the actual p-value from the paired test?

(a) the p-value of your test would be larger
(b) the p-value of your test would be smaller

```
## SOLUTION:
# (a) the p-value of your test would be larger
```

17.2 [1 point] How would the degrees of freedom for the test you conducted compare to the degrees of freedom in the paired test?

(a) the degrees of freedom of your test would be larger
(b) the degrees of freedom of your test would be smaller

```
## SOLUTION:
# (a) degrees of freedom of your test would be larger
```

Question 18 [5 points total]

Fill in the blanks. In a clinical trial that evaluates the effectiveness of a glycemic control drug, 125 patients were randomized to the treatment group, and 180 patients were randomized to the control group. Consider the following steps to construct a bootstrap confidence interval for the median glucose level for patients in the *treatment group*.

1. Compute the median glucose level for the original sample.
2. Resample **A** (with/without) replacement from the original sample. The size of the new sample should be **B**.
3. Compute the **C** of the new sample.
4. Repeat steps **D** 1000 times.
5. Suppose the medians of samples are stored in a vector called `sample_median`, the R code to compute the **upper** bound of an **80%** confidence interval is: **E**.

A:

B:

C:

D:

E:

```
## SOLUTION:
# A: with
# B: 125
# C: median
# D: 2 and 3
# E: quantile(sample_median, 0.9)
```

Question 19 [8 points total]

For the following scenario questions, determine which test/method you would use to analyze the data and explain your choice in 1 sentence.

19.1 [3 points] You interviewed 300 people in the Central Valley and classified each person's COVID-19 status as never having had the disease, having had mild infection/symptoms, or having had severe infection/symptoms. You also asked them about their occupational status, which you classified as essential worker, remote worker, or non-worker. What parametric *and* non-parametric test would you use to test if COVID-19 status and occupational status are related?

Tests:

Explanation:

```
## SOLUTION:
# Chi square test of independence
# Appropriate because both variables are categorical
# Non-parametric test you could also perform is a permutation test
# Appropriate b/c permutation tests can be used for categorical variables (mixing up labels)

# 1 point for correct parametric test, 1 point for correct non-parametric test
# 1 point for explanations
```

19.2 [3 points] You are conducting a study to determine if there is a relationship between the average number of days new mothers breastfeed their babies and their race/ethnicity (categorized as non-Latine Asian, non-Latine Black, Latine, non-Latine White). List two methods you could use to analyze these data and provide and explanation for why you chose these methods.

Tests:

Explanation:

```
## SOLUTION:

# solution a: ANOVA: one cat variable, one continuous variable

# solution b: regression with categorical predictor and continuous outcome

# 1 point for correctly saying regression and ANOVA
# 1 point for correct explanation
```

19.3 [2 points] You are studying a rare form of cancer in children. After collecting a sample of 15 children in the U.S. who have this cancer, you decide to test whether there is a difference in the age distributions

of children who have a particular gene expressed vs. those who do not have the gene expressed. What test/method would you use to analyze these data?

Test:

Explanation:

```
## SOLUTION:
# Wilcoxon Rank Sum
# sample is small and we want to see the difference in age distributions
# between 2 independent samples (gene expressed and gene not expressed)

# 1 point for correctly saying Wilcoxon Rank Sum
# 1 point for correct explanation
```

Question 20 [6 points total]

Epileptic patients in a randomized control trial were either assigned to a ketogenic diet or a control diet during a week-long study. The following table represents whether patients in each group experienced one or more seizures during the week-long period.

	Seizure	No Seizure	Total
Control	80	20	100
Ketogenic	65	35	100

20.1 [1 point] In each diet group calculate the proportion of those who experienced one or more seizures. What is the sample estimate for the difference in proportions between these groups? Let p_1 denote the proportion for the control group and p_2 denote the proportion for the keto group.

```
## SOLUTION:
# $p_1$ - $p_2$ = 0.8 - 0.65 = 0.15
# -0.15 also accepted
```

20.2 [1 point] Calculate the margin of error for a 95% confidence interval using the large sample method. Show your work and express your final answer as a single value rounded to four decimal places.

```
## SOLUTION:
# moe = 1.96*sqrt((0.8*0.2/100) + (0.65*0.35/ 100))
# moe = 0.1220

# 0.5 point for correct work
# 0.5 point for correct answer
```

20.3 [1 point] Provide the lower and upper bounds for the 95% confidence interval using the large sample method. Express your answers as percentages rounded to two decimal places.

```
## SOLUTION:
# lower = 0.15 - 0.1220 = 0.028 = 2.80%
# upper = 0.15 + 0.1220 = 0.2720 = 27.20%

# 0.5 point each for lower/upper bound
```

20.4 [3 points] In addition to the number of seizures, you are interested in conducting a study to determine whether the ketogenic vs. control diet has an effect on the maximum heart rate. For this next study, you're considering using a paired design. Define what a carryover effect is, and explain how you would investigate if a carryover effect is present.

```
## SOLUTION:

# carryover effect - when the effect of the first treatment is still detected
# when patient has switched to the second treatment.
# Could happen here for patients randomized to ketogenic diet first.

# Could make a plot for people randomized to ketogenic diet first and look at
# their maximum heartrate for the two treatment conditions.
# Then make the same plot for people randomizeed to control diet first and
# look at the same thing.
# If there is a carry over effect, then the average (or median) heart rate
# among the control group will be different based on whether they had the keto diet first.

# 1 point for mentioning carryover effect
# 2 points for explaining how to tell if it's present
```

Question 21 [12 points total]

A team of researchers asks you to assist them in their study on the prevalence of human papillomavirus (HPV) among different age groups. They collected a simple random sample of individuals from California and are trying to generalize their findings to individuals across the entire U.S. They aggregated the data on HPV status vs. age group below.

21.1 [2 points] Fill in the missing observed counts. Also fill in the missing observed percentages when indicated by parentheses. Round the percentages to two decimal places. Using scratch paper to work out this question will be helpful.

Age Group	HPV +	HPV -	Row total
14-19	160	**A**	**B ( )**
20-24	85	104	**C ( )**
25-29	48	126	174 (9.06%)
30-39	90	**D**	328 (17.07%)
40-49	82	242	324 (16.87%)
50-59	50	204	**E ( )**
Col total	**F ( )**	**G ( )**	**H**

A:

B:

C:

D:

E:

F:

G:

H:

```
## SOLUTION:

# |Age Group    | HPV +         | HPV -           | Row total     |
# |:-----------:|:-------------:|:---------------:|:-------------:|
# |14-19        | 160           | 492             | 652 (33.94%)  |
# |20-24        | 85            | 104             | 189 (9.84%)   |
# |25-29        | 48            | 126             | 174 (9.06%)   |
# |30-39        | 90            | 238             | 328 (17.07%)  |
# |40-49        | 82            | 242             | 324 (16.87%)  |
# |50-59        | 50            | 204             | 254 (13.22%)  |
# | Col total   | 515 (26.81%)  | 1406 (73.19%)   | 1921          |


# A: 492
# B: 652 (33.94%)
# C: 189 (9.84%)
# D: 238
```

```
# E: 254 (13.22%)
# F: 515 (26.81%)
# G: 1406 (73.19%)
# H: 1921

# 0.25 point per correct observed count
```

21.2 [3 points] Compute the expected counts for HPV status vs. age group. Round to two decimal places.

Age Group	HPV +	HPV -
14-19	A	B
20-24	C	D
25-29	E	F
30-39	G	H
40-49	I	J
50-59	K	L

A:

B:

C:

D:

E:

F:

G:

H:

I:

J:

K:

L:

```
## SOLUTION:

# |Age Group    | HPV +         | HPV -           | Row total     |
# |:-----------:|:-------------:|:---------------:|:-------------:|
# |14-19        | 174.80        | 477.20          | 652 (33.94%) |
# |20-24        | 50.68         | 138.32          | 189 (9.84%)  |
# |25-29        | 46.66         | 127.34          | 174 (9.06%)  |
# |30-39        | 87.91         | 240.09          | 328 (17.07%) |
# |40-49        | 86.88         | 237.12          | 324 (16.87%) |
# |50-59        | 68.09         | 185.91          | 254 (13.22%) |
# | Col total   | 515 (26.81%) | 1406 (73.19%)  | 1921          |

# A: 174.80
# B: 477.20
# C: 50.68
# D: 138.32
# E: 46.66
# F: 127.34
# G: 87.91
# H: 240.09
# I: 86.88
# J: 237.12
```

```
# K: 68.09
# L: 185.91

# 0.25 points per correct expected count
```

21.3 [1 point] State the null and alternative hypotheses to test whether HPV and age group are related.

H_0:

H_A:

```
## SOLUTION:

# $H_0$: HPV status and age group are independent.
# $H_A$: HPV status and age group are not independent.

# 0.5 points for each hypothesis
```

21.4 [2 points] Compute the test statistic for these data. Show how you arrived at your answer and round to two decimal places. You can use and show code, or provide calculations completed by hand.

```
## SOLUTION:

# chi square test stat = 40.55

# WORK OPTION 1: Code
library(tibble)

hpv <- tribble(~hpv, ~no_hpv,
               160, 492,
               85, 104,
               48, 126,
               90, 238,
               82, 242,
               50, 204)

chisq.test(hpv, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  hpv
## X-squared = 40.554, df = 5, p-value = 1.155e-07
```

```
# WORK OPTION 2: by hand

# [(174.80 - 160)^2/174.80] + [(477.20 - 492)^2/477.20] + [(50.68 - 85)^2/50.68] +
# [(50.68 - 85)^2/50.68] + [(138.32 - 104)^2/138.32] + [(46.66 - 48)^2/40.66] +
# [(127.34 - 126)^2/127.34] + [(87.91 - 90)^2/87.91] + [(240.09 - 238)^2/240.09] +
# [(86.88 - 82)^2/86.88] + [(237.12 - 242)^2/237.12] + [(68.09 - 50)^2/68.09] +
# [(185.91 - 204)^2/185.91] = 40.55

# 1 point for correct test stat
# 1 point for correct code or work
```

21.5 [1 point] Write the code used to calculate the p-value.

```
## SOLUTION:
# df = (r-1)(c-1) = (6-1)(2-1) = 5
pchisq(40.5, df = 5, lower.tail = FALSE)
```

```
## [1] 1.183843e-07
```

```
# OR

1 - pchisq(40.5,df = 5)
```

```
## [1] 1.183843e-07
```

```
# 1 point for showing code (give point if use incorrect test stat but
# correct df and lower.tail = F)
# 1 point for correct p-value
```

21.6 [1 point] You obtain a p-value of 1.18e-7. Interpret this value in the context of this problem.

```
## SOLUTION:

# The p-value is very small, less that 0.001%; therefore, we reject the
# null hypothesis that age group and HPV status are independent.

# OR

# There is a 1.18e-5% chance of seeing the test statistic we saw or more extreme
# if the null were true - this is very unlikely so we reject the null.
```

21.7 [2 points] Suppose that the researchers later learned that there is no difference in HPV prevalence among age groups in the U.S. What kind of error occurred in the researchers' study? Explain in 1-2 sentences.

```
## SOLUTION:

# Type I error
```

```
# Type I error = Pr(rejecting the null | null is true).
# There was really no difference in HPV prevalence in the whole population (null is true)
# but the researchers rejected the null in the study.

# 1 point for correct error
# 1 point for correct explanation
```

Question 22 [13 points total]

A new dating app for UC Berkeley students (CalLove) recruits you to analyze differences in gender identity on their platform. They created a dataframe called `dating` that contains data on each participant's gender (`gender`), hours spent on the app (`hours_spent`), number of matches (`num_matches`), and number of likes received (`num_likes_received`) for 90 individuals.

22.1 [1 point] You wish to test whether there is a difference between the mean amount of time spent on the app among the three gender identities on the platform (Male, Female, Non-Binary/Other). Which of the following is the best procedure to answer this research question?

(a) ANOVA
(b) Chi Squared Goodness of Fit
(c) Chi Squared Test for Independence
(d) Wilcoxon Sign-Rank

```
## SOLUTION:
# (a) ANOVA
```

22.2 [1 point] State the null and alternative hypotheses in the context of this question.

H_0:

H_A:

```
## SOLUTION:

# $H_0$: The mean amount of time spent on the app is the same for
# females, males, and non-binary/other folks

# OR

# $H_0$: $\mu_f$ = $\mu_m$ = $\mu_nb$



# $H_A$: At least one of the means differs.
```

22.3 [3 points] List the assumptions for the test you chose and in one sentence (total) describe how you would check if each assumption is satisfied.

```
## SOLUTION (1 point for each):
# SRS - cannot check with plot, need to know how data is collected
# Normality - qqplot
# same SD - calculate SD for each group and see if largest is < 2 times the smallest
```

22.4 [2 points] Write a line of code to run the statistical test you chose in 22.1. In the code, assign the test to an object called `dating_hyp_test`. Then write the line of code that gives you the output below.

```
## SOLUTION:
# dating_hyp_test <- aov(hours_spent ~ gender, dating)
# summary(dating_hyp_test)

# 1 point for each correct line of code
```

22.5 [1 point] Use the output above to interpret the p-value in the context of this question.

```
## SOLUTION:
# The p-value is 0.00155. Thus there is strong evidence to reject the null hypothesis
# and conclude that at least one of the means is different than the others.
```

22.6 [2 points] Write the line of code that gives you the output below. What can you conclude from this output? Note that the gender variable has the three levels: Female, Male, and Non-Binary/Other.

Code:

Conclusion:

```
## SOLUTION:

# TukeyHSD(dating_hyp_test)

# The mean number of hours spent on the app is significantly different for males
# compared to females and non-binary/other.

# 1 point for correct code
# 1 point for correct conclusion
```

22.7 [1 point] Why is the test you performed in the previous question better than performing pairwise tests for each group? Explain in one sentence.

```
## SOLUTION:
# It maintains an overall/familywise error rate of 5% (corrects for multiple testing)
```

22.8 [1 point] Now suppose your research question is to determine whether the representation of genders on CalLove is different from the distribution of gender identities of all students registered at Cal. Which of the following is the best procedure to answer this research question?

(a) One-Sample/Two-Sample T test
(b) Chi Square Goodness of Fit
(c) Regression
(d) ANOVA

```
## SOLUTION:
# (b) Chi Square Goodness of Fit
```

22.9 [1 point] Now suppose your research question is to determine whether there is a relationship between the number of likes received vs. the number of hours spent on the app. Which of the following is the best procedure to answer this research question?

(a) One-sample/Two-sample proportion test
(b) One-sample/Two-sample t-test
(c) Permutation test
(d) Regression

```
## SOLUTION:
# (d) Regression
```

Question 23 [8 points total]

Researchers at the Lawrence Berkeley National Laboratory are working with a hydrogen-oxidizing bacterium called *Cupriavidus necator*, which is classified as a chemolithotroph. This organism has the amazing ability to naturally consume carbon dioxide and convert it to a biodegradable plastic-like molecule called PHB.

One researcher wants to see whether or not light affects *Cupriavidus necator*'s ability to convert carbon dioxide to PHB. She took 5 different cultures of *Cupriavidus necator* (each culture having the same number of organisms) to test this theory. For each culture, she measured the amount of PHB produced under normal conditions, left it under a light for one hour, then measured the amount of PHB produced after the light exposure. The following chart shows the PHB produced (in weight) before and after exposure to light:

(Note: Although the bacterium is real, the data provided is simulated.)

PHB Produced Before Light (μg)	PHB Produced After Light (μg)	Difference	Sign	Rank
300	305	**A**	**B**	1
273	255	-18	-	**C**
289	301	12	+	2
265	242	-23	+	**D**
321	303	-18	-	**C**

Sum of Positive Sign Ranks	Sum of Negative Sign Ranks
E	**F**

23.1 [3 points] Fill in the table with the difference, sign, and rank for each data point and the total sum for each sign.

A:

B:

C:

D:

E:

F:

```
## SOLUTION:
# A: 5
# B: +
# C: 3.5
# D: 5
# E: 8
# F: 7

# 0.5 points for each
```

```
## Note: there was a typo in this question. The difference in row 4 should
# actually be -23 and the corresponding sign negative. Updated solutions below:
# A: 5
# B: +
# C: 3.5
# D: 5
# E: 3
# F: 12
```

23.2 [1 point] Calculate the μ_T and σ_T statistics. Do not round your answers.

$\mu_T =$
$\sigma_T =$

```
## SOLUTION:
# mu_T = 5 * (5 + 1) / 4 = 7.5
# sigma_T = sqrt(5 * (5 + 1) * (2 * 5 + 1) / 24) = 3.708099

# NOTE: Despite the typo, the answer for 23.2 is still the same.
```

23.3 [1 point] Calculate the Z_T statistic. Show your work and do not round your answer.

$Z_T =$

```
# SOLUTION:
# Z_T = (3 - 7.5) / 3.708099
# Z_T = -1.21356

# 0.5 point for work
# 0.5 point for correct answer
```

23.4 [1 point] What distribution does Z_T approximately follow?

$Z_T \sim$

```
## SOLUTION:
# Either N(0, 1) or standard normal distribution
# Half credit if they simply mention normal distribution
```

23.5 [1 point] Write the line of code used to calculate the p-value.

```
# SOLUTION:
# 2 * pnorm(-1.21356) = 0.2249158
```

23.6 [1 point] You obtain a p-value of 0.89. Interpret your p-value in the context of this question. Would you reject the null hypothesis?

```
## SOLUTION:
# Given the null hypothesis that there is no difference in PHB production between
# exposure to light and no light, there is a 89% chance that we see
# the observations that we saw or more extreme.
# Since our p-value is very large, we fail to reject the null hypothesis.

# 0.5 point: Interpreted the p-value correctly
# 0.5 point: Fail to rejected the null hypothesis
# -0.5 Points: didn't answer in the context of the question
# (e.g. didn't mention difference in PHB production, exposure to light vs no light, etc.).
```

```
# NOTE: Actual p-value is 0.22, but left it as 0.89 on the exam.
```

Question 24 [13 points total]

The FDA has recruited you to study the nutritional components of a sample of common US cereal brands. They ask you to analyze whether the grams of complex carbohydrates per serving (`carbo`) are linearly correlated with the total calories per serving (`calories`).

First, go to Datahub. Replace NULL with your numeric student ID. Do not put quotes around your SID. Run that line of code. Then, run the subsequent chunks to import the two datasets into the R environment. For this question, you will be working with the dataset called `cereal`.

24.1 [1 point] What are the null and alternative hypotheses to test whether there is a linear relationship between `carbo` and `calories`?

H_0:

H_A:

```
## SOLUTION:
# $H_0$: slope = 0 (no linear relationship between carbohydrates and total calories)
# $H_A$: slope does not = 0 (there is a linear relationship between carbohydrates
# and total calories)
```

24.2 [1 point] On Datahub, perform a linear regression of `calories` and `carbo` using the `cereal` data and assign this to an object called `cereal_fit`. Tidy your results. What is the slope coefficient of this model? Do not round your answer. (You do not need to include your code here).

```
## SOLUTION:

# y = calories, x = carbo
cereal_fit <- lm(formula = calories ~ carbo, data = cereal)
tidy(cereal_fit)

# slope = 5.812818
```

24.3 [2 points] Calculate the 95% confidence interval for this slope coefficient using the tidied results you generated in the previous question. Show how you calculated the CI and do not round your answer.

```
## SOLUTION:

# lower bound: 4.672149
5.812818 - 0.570808 * qt(p = 0.975, df = 65-2)
```

```
## [1] 4.672149
```

```
# upper bound: 6.953487
5.812818 + 0.570808 * qt(p = 0.975, df = 65-2)
```

```
## [1] 6.953487
```

```
# 0.5 points for lower bound, 0.5 points for upper bound
# 1 point for correct work shown (CI equation, correct critical value)
```

24.4 [2 points] Interpret *both* the slope and 95% confidence interval in the context of this problem.

Slope interpretation:

95% CI interpretation:

```
## SOLUTION:

# For every 1 gram increase in the complex carbohydrates, there is an increase of
# 5.81 total calories. If we repeat this process many times, we would expect the
# true slope for all cereal brands in the US to be between
# 4.672149 and 6.953487 95% of the time.

# 1 point for correct slope interpretation
# 1 point for correct 95% CI interpretation
```

24.5 [1 point] Calculate the R^2 value. Write the code you used to determine the R^2 value and interpret this value in the context of the problem.

Code:

Interpretation:

```
## SOLUTION:

# glance(cereal_fit)

## The variation in complex carbohydrates explains about 62.2% of the variation
# in total calories for this sample of cereal brands.
```

24.6 [1 point] What can you conclude about the relationship between `carbo` and `calories`?

```
## SOLUTION:

# p < 0.05 and the confidence interval does not contain 0 so we reject the
# null hypothesis that there is no linear relationship between `carbo` and `calories`.
```

24.7 [2 points] Calculate the mean response when the cereal brand has a carbohydrate component of 15 grams and report the estimate with its 95% confidence interval. Do not round your answer. Show the code you used to calculate this value. How would this compare to the prediction interval for an individual cereal brand with a carbohydrate component of 15 grams?

Code:

Estimate and 95% CI:

Comparison:

```
## SOLUTION:

# newdata = data.frame(carbo = 15)
# predict(cereal_fit, newdata, interval = "confidence")
```

```
# 120.5324 ; 95% CI [109.3976 131.6671]

# The prediction interval for an individual would be wider
# intervals are narrowest near the mean value since the line of best fit
# passes through the mean values of x and y.


# 0.5 point for correct code
# 0.5 point for correct estimate and 95% CI
# 0.5 point for saying the prediction interval would be wider than the CI
# 0.5 point for correct explanation
```

The cereal brand companies also reported data on the vitamins contained in each cereal. The corresponding variable is called `vitamins`. You are encouraged to examine this variable to help you with the next question.

24.8 [2 points] The FDA would like you to consider the relationship between the cereal's type of vitamin (`vitamins`) in relation to the total calories per serving. Make a new dataframe called `cereal2` where "none" is the referent category for the `vitamins` variable and perform a linear regression to help you answer the FDA's question. Show the code you used to order the `vitamins` variable and the code used to run the model. Interpret the coefficient for cereals with enriched vitamins.

Code:

Interpretation:

```
## SOLUTION:

# cereal_ordered <- cereal %>% mutate(vitamins_reordered = fct_relevel(vitamins, "none"))

# vitamin_fit <- lm(formula = calories ~ vitamins_reordered, data = cereal_ordered)
# tidy(vitamin_fit)

# Cereal brands with enriched vitamins have an average of 47.40 more
# total calories per serving compared to cereal brands with no vitamins.

# 1 point for code that reorders variable + fits the linear regression
# 1 point for correct interpretation
```

24.9 [1 point] Calculate the mean response and 95% CI of the total calories for the cereal brands with enriched vitamins. Do not round your answer.

```
## SOLUTION:

# newdata_vitamin = data.frame(vitamins_reordered=c("none", "enriched", "100%"))

# predictions_vitamin <- data.frame(predict(vitamin_fit, newdata_vitamin,
# interval="confidence"))

# predictions_vitamin$vitamins_reordered <- c("none", "enriched", "100%")
# predictions_vitamin
```

```
## 153.6178 ; 95% CI [137.13586, 170.0997]
```

Question 25 [8 points total]

Researchers are conducting a phase II trial to test the efficacy of a drug's ability to help reduce systolic blood pressure among hypertensive individuals. They obtain a simple random sample of 100 individuals that is representative of the population of US adults with hypertension. Each person's systolic blood pressure was measured before beginning the drug treatment (`pretreat_bp`) and 6 months after undergoing treatment (`posttreat_bp`).

For this question you'll be working with the `drug_bp` dataset on Datahub. You should already have this dataset loaded into Datahub when you ran the code to set up the data for the previous question. If not, load it in now.

```
drug_bp <- read.csv("drug_bp.csv")
```

25.1 [1 point] State the null and alternative hypotheses in the context of this question.

H_0:

H_A:

```
## SOLUTION:

# $H_0$: There is no difference in the systolic blood pressure before
# the drug treatment and after the drug treatment $\mu_d$ = 0
# $H_A$: The systolic blood pressure before the drug treatment is greater than
# the blood pressure after the treatment $\mu_d$ > 0.
# They may say < 0 if their difference is post - pre.
```

25.2 [1 point] Calculate the p-value. Show the code used to calculate the p-value for this test and interpret this exact value in the context of this question.

Code:

Interpretation:

```
## SOLUTION:

# OPTION 1:
# drug_bp_diff <- drug_bp %>% mutate(diff = pretreat_bp - posttreat_bp)

# t.test(drug_bp_diff %>% pull(diff), data = drug_bp_diff, alternative = "greater")

# OPTION 2:
```

```
# t.test(drug_bp %>% pull (pretreat_bp), drug_bp %>% pull(posttreat_bp),
# data = drug_bp, paired = T, alternative = "greater")
# p-value: 2.2e-16
# Assuming the null is true, there is a nearly 0% (2.2e-16%) chance of seeing
# the test statistic we saw or smaller.
# We reject the null hypothesis that the drug has no effect on blood pressure.

# 0.5 point for code
# 0.5 point for correct interpretation
```

25.3 [2 points] How would the variability (measured by the standard error) have been different if the researchers had instead measured and compared the systolic blood pressures of two independent samples of individuals? (i.e., one sample's blood pressure was used as the "before" measurement while a separate sample underwent the treatment and their blood pressure used as the "after" measurement).

```
## SOLUTION:

# The variability would increase. There is more variation between separate individuals
# compared to the variation within a single individual at two points in time
# (difference in variability between a paired t test vs. a 2 sample t test).

# 1 point for correctly saying the variability would increase
# 1 point for correct explanation
```

The researchers are interested in testing whether a higher dose of the drug further reduced blood pressure compared to the lower dose of the drug assessed in the previous question. They plan to conduct another experiment with a different set of 45 individuals in which they measure the participants' blood pressure before and after receiving a high dose of the drug vs. a lower dose of the drug. Their null hypothesis is that the difference between doses is equal to 8 mm Hg, and their alternative is that the mean difference (e.g., low dose - high dose) is larger than 8 mm Hg.

25.4 [2 points] How certain are you that the test will reject the null that the sample mean difference in blood pressure is equal to 8 mm Hg in the case that the true population mean is equal to 12 mm Hg and the true population standard deviation is equal to 9 mm Hg? Assume a 5% type I error rate and that the systolic blood pressure measurements are normally distributed in the population. Show your work.

```
## SOLUTION:

# population mean based on the phase II trial: mean(drug_bp_diff$diff) = 11.59783
# population sd based on the phase II trial: sd(drug_bp_diff$diff) = 9.907486

# qnorm(p = 0.05, mean = 8, sd = 9/sqrt(45), lower.tail = F)

# pnorm(10.2068, mean = 12, sd = 9/sqrt(45), lower.tail = F)

# power = 90.9%

# 1 point for correct work
# 1 point for correct power
```

25.5 [2 points] The researchers go on to conduct the test. They fail to reject the null that the mean difference in systolic blood pressure before and after drug treatment is 8 mm Hg. Given the information in question 25.4, what kind of error (if any) occurred? Explain in 1-2 sentences.

```
## SOLUTION:
# type II error
# type II error = Pr(fail to reject null | null is false).
# From the information above, we know that the population mean difference
# in blood pressure before and after treatment is 11.59783 mm Hg.
# We mistakenly reject the null (that the pop mean difference before and after
# treatment is 8 mm Hg), so we are making a Type II error.
```

```
# 1 point for correct error named
# 1 point for explanation
```