# MATH 338

# FINAL EXAM - LAB PORTION

# DUE THURSDAY, DECEMBER 20, 2018

## Your name: ______________________

Your scores (to be filled in by Dr. Wynne):

| | |
|---|---|
| Problem 1: | ____/4 |
| Problem 2: | ____/6 |
| Problem 3: | ____/4 |
| Problem 4: | ____/4 |
| Total: | ____/18 |

You have one week to complete this exam, save your answers (as a .docx or .pdf file) and upload the file to Titanium.

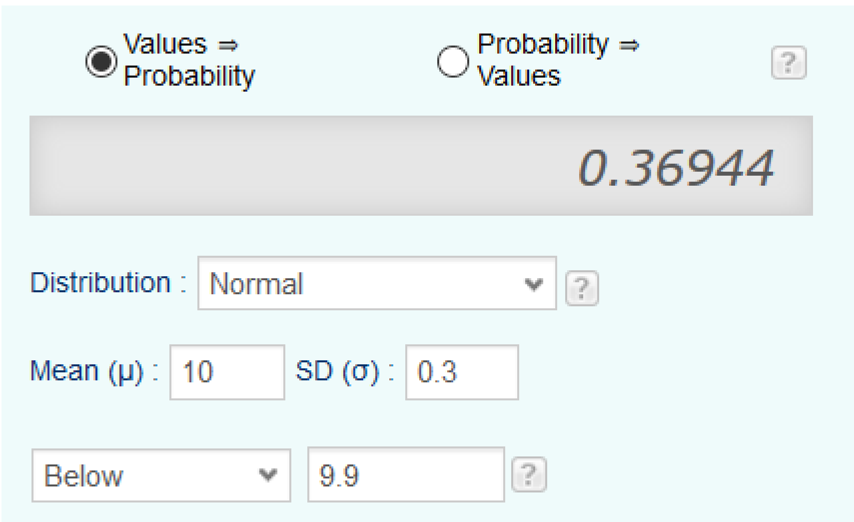
You may refer to your notes, your textbook, and any pre-existing online reference (eBook, R/Rguroo help, anything on Titanium). You may search for help online, but you must cite any source found through the search. You may ask Dr. Wynne to clarify what a question is asking for. You may not ask other people for help or use any other resources.

For full credit, show all work except for final numerical calculations (which can be done using a scientific/graphing calculator or R).

1. If you're as bad at chem lab as I was, you will have a lot of measurement error affecting your results. Suppose that I am terrible at pipetting and, when I attempt to pipette 10 mL of water, the actual amount I pipette is normally distributed with mean 10 mL and standard deviation 0.3 mL.

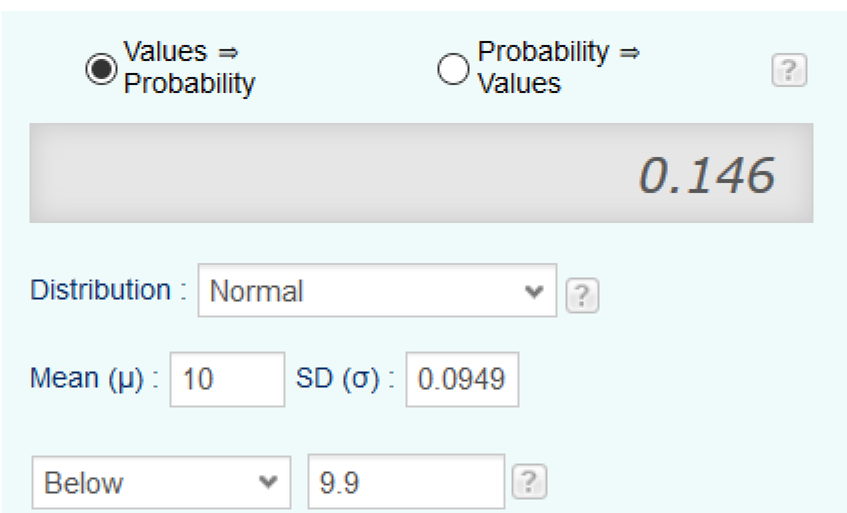
For parts (a) and (b) below, round your answers to 3 decimal places. Also, include some way of indicating to me how you got software to compute your answer (Rguroo screenshots and/or R code).

a) [1.5 pts] What is the probability that I pipette less than 9.9 mL of water?




The probability that I pipette less than 9.9 mL of water is approximately 0.369.

b) [2 pts] What is the probability that, over 10 independent attempts, I average less than 9.9 mL of water in the pipette?



The probability that I average less than 9.9 mL of water in the pipette over 10 independent attempts is approximately 0.146.

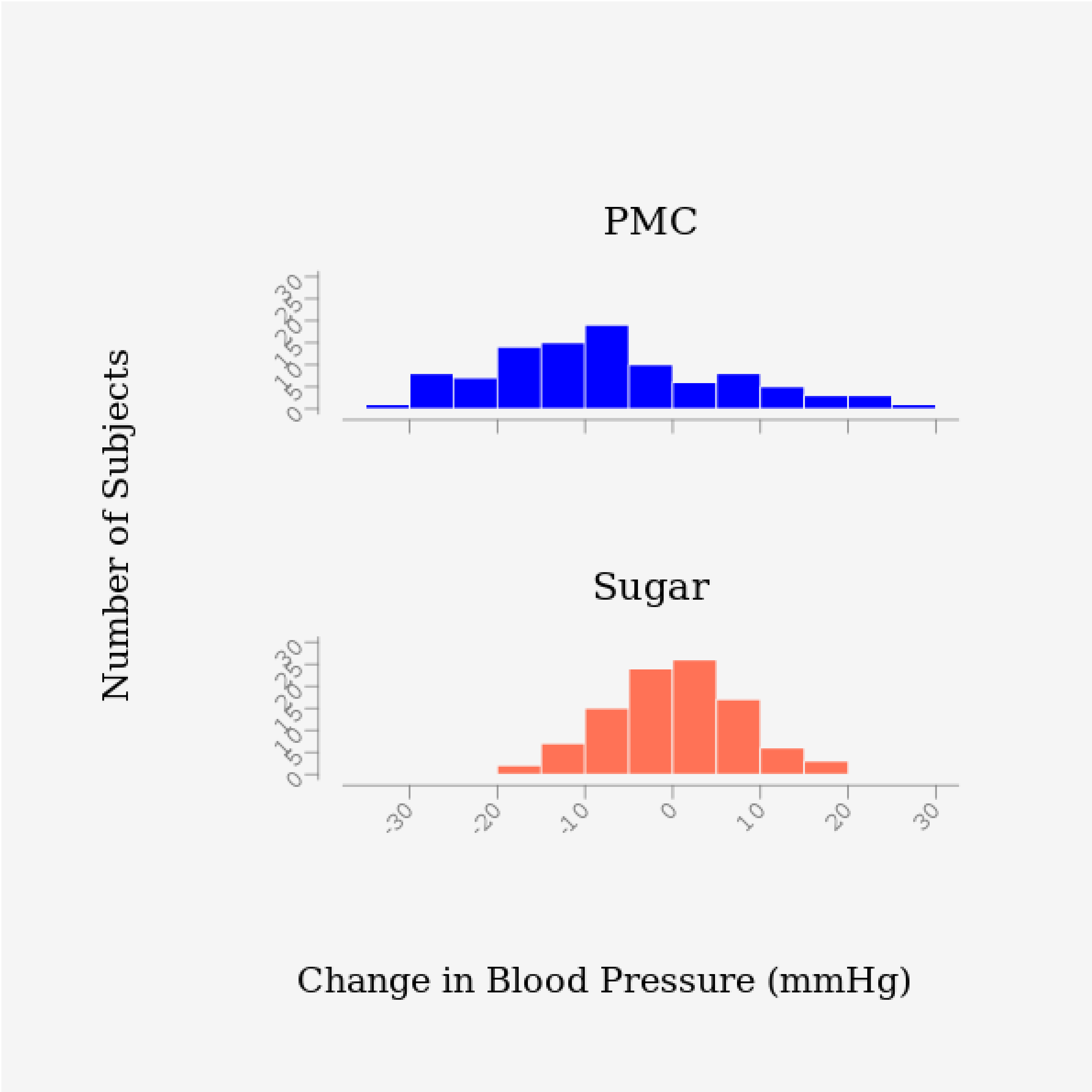
c) [0.5 pts] Go to the following site: http://fullerton.qualtrics.com/jfe/form/SV_8wUu57lprlFPH1P. Write your name and the date in the boxes at the top of the survey and take a screenshot. Paste the screenshot below. Then read the rest of the page and either accept or decline participation in this research study. If you do participate in the study, fill out the rest of the survey. (I cannot give credit for participating because I'm not supposed to know who is in the study)

2. Having prolonged high blood pressure can have many negative effects on general health and can potentially lead to stroke and/or death. There are many drugs available that are known to help lower blood pressure, however, these drugs often come with negative side effects. A clinical trial was set up to explore a more natural treatment of taking a potassium, magnesium, and calcium (PMC) complex tablet. These minerals are known to counteract the blood pressure raising effects of high sodium intake.

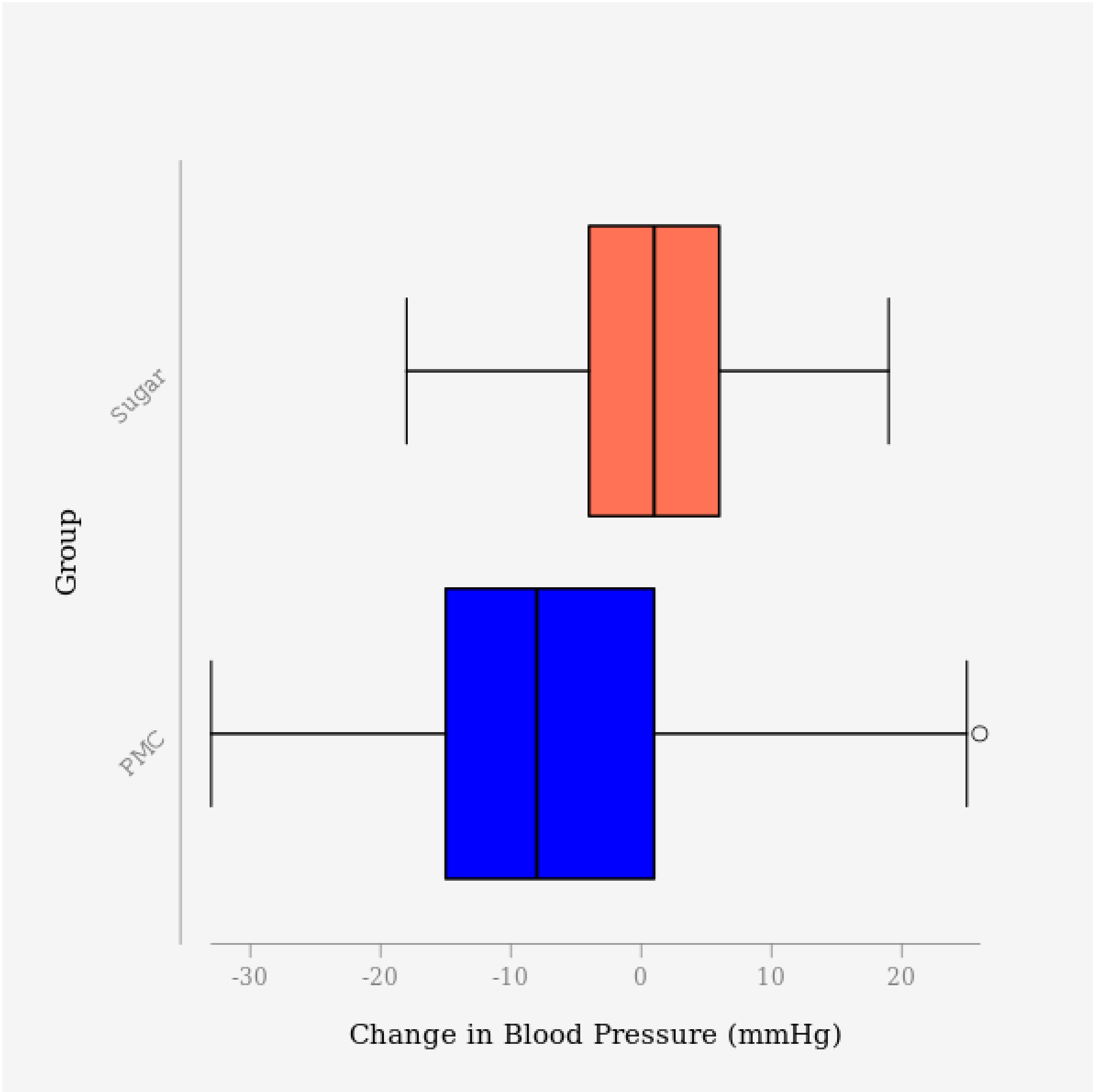
A random sample of 200 adult males (29-49 years of age) currently diagnosed with hypertension (high blood pressure) were obtained from a pool of subjects who volunteered to be part of the study. The subjects were then randomly assigned to one of two groups: one group received the PMC complex tablet and the other received a sugar tablet.

The *PMC.csv* and *PMC1.csv* files contains the change in blood pressure, to the nearest mmHg, over the course of the study in both the PMC and Sugar groups. (These files contain the same data, presented in two different forms. You may use either form you are comfortable with) Note that positive values mean that the blood pressure went up over the course of the study, and negative values mean that the blood pressure went down.

a) [1.5 pts] Perform exploratory analysis on this dataset – describe the distribution of the change in blood pressure in each group in the observed sample of 200 adult males and compare the two distributions.




According to the histograms above, the distribution of the change in the PMC group is skewed right with a peak around -10, while the distribution of the change in the Sugar group is roughly symmetric with a peak around 5. The variability in the PMC group is considerably larger than the variability in the Sugar group.




The boxplots confirm the difference in variability, with one outlier identified in the PMC group, and the difference in center, with the median in the PMC group being negative and the median in the Sugar group being positive. The boxplots confirm that the PMC group is skewed right and that the Sugar group is roughly symmetric.

b) Perform an appropriate statistical hypothesis test to draw conclusions about the effectiveness of the PMC complex. Write a short paragraph describing your conclusions. Full credit will be given in this part for a writeup that:

[0.5 pts] Uses background analysis and the exploratory analysis from part (a) to inform your decision about the correct hypothesis test to use

[1 pt] Specifies an appropriate null hypothesis, alternative hypothesis, and significance level

[1 pt] Contains relevant, necessary and correct (R code and) software output

[2 pts] Makes a statistically justified conclusion and interprets that conclusion <u>in the context of the problem</u>.

Based on our background analysis and the exploratory analysis in part (a), we have two independent samples with quite different variation; although there is an outlier in one of the groups, we have a reasonably large sample size (100 in each group), we should be okay to do a two-sample t-test.

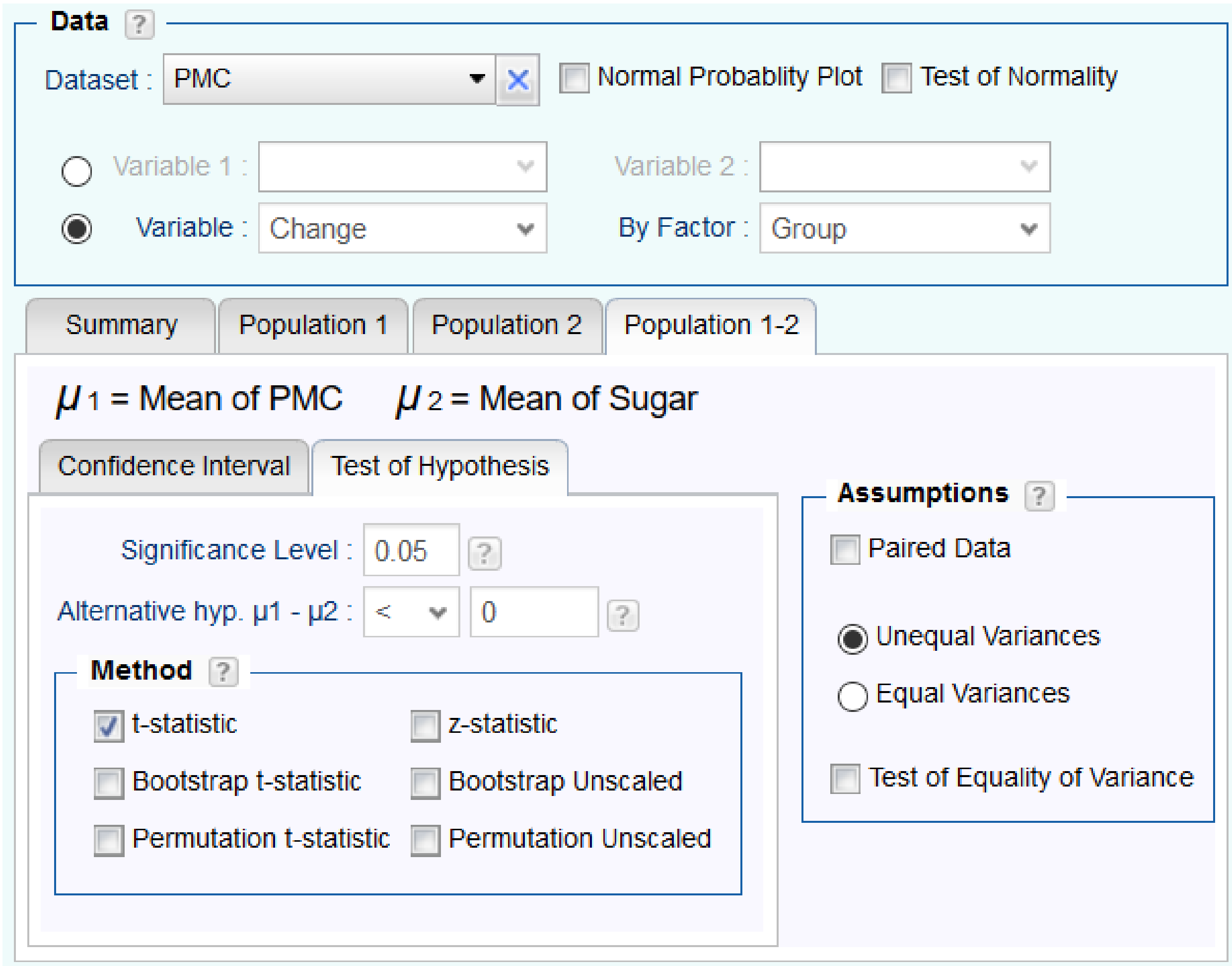
$H_0$: the two groups have the same population mean, $\mu PMC = \mu_{Sugar}$; or, the PMC tablet has no effect on the mean change in blood pressure (compared to placebo)

$H_a$: the PMC complex is effective at reducing blood pressure, $\mu_{PMC} < \mu_{Sugar}$

Appropriate alternative Ha: the PMC tablet has an effect on the mean change in blood pressure (compared to placebo), $\mu_{PMC} \neq \mu_{Sugar}$

Significance level: $\alpha = 0.05$

We perform a two-sample t-test:




And obtain the following output:

*Test of Hypothesis: t-test*
*Change (PMC) - Change (Sugar)*

Research Hypothesis Ha: Mean of 'Change (PMC) - Change (Sugar)' is less than 0
5% lower critical value in units of data = -2.52136
Unequal population variances was assumed.

| Diff of Means | Std Error | Obs t Stat | DF | 5% t-Lower Critical | P-value |
|---|---|---|---|---|---|
| -7.44000 | 1.52352 | -4.88342 | 151.791 | -1.65495 | 1.30887e-06 |

*Test is significant at 5% level.*

Because the p-value of 0.0000013 < 0.05, we reject the null hypothesis. We found evidence that the PMC tablet does significantly decrease blood pressure (as compared to placebo); or, if you used a two-sided alternative, we found evidence of a significantly different effect of the PMC tablet on blood pressure as compared to placebo.

3. A 2018 study investigated whether macaque monkeys can perform statistical reasoning. In the study, 11 macaques watched a researcher select food from two buckets. On each trial, the researcher drew a piece of food from a bucket containing 80% grapes with one hand and drew from a bucket containing only 20% grapes with the other hand. After looking at the drawing, the monkeys indicated which hand they wanted food from. Each monkey selected a hand on 12 trials.

The *monkeysee-monkeydoinference.csv* file contains the name of the monkey (Individual) and the number of "correct" choices out of 12 that the monkey made (Score).

Estimate with 95% confidence the proportion of the time that Sophie would make the "correct" choice if presented with very many trials. Do you believe that Sophie is picking a hand entirely at random, or do you think that she is thinking about which hand is more likely to have the grapes?

Hint: You don't need to actually import this file into the software! You can just figure out what the appropriate numbers are!

Full credit will be given in this problem for a writeup that:

[1 pt] Includes appropriate background and/or exploratory analysis that informs your decision about the correct confidence interval procedure to use

[1 pt] Contains relevant, necessary and correct (R code and) software output

[0.5 pts] Correctly identifies the confidence interval from the output

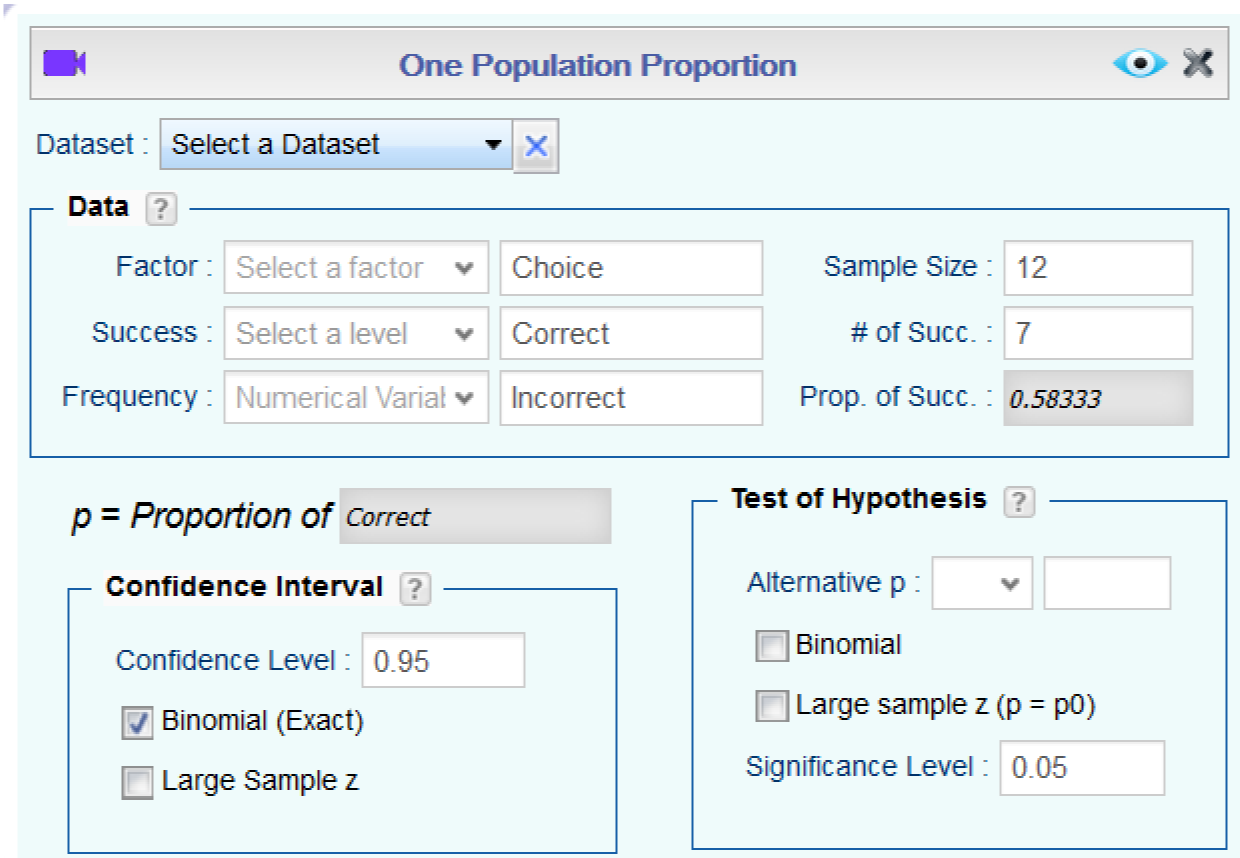
[1.5 pts] Correctly interprets the confidence interval in the context of the question asked.

Sophie is in row 2 of the dataset; she got 7 out of 12 correct.

We would like to perform a one-sample confidence interval for a population proportion here. The B and N assumptions of BINS are definitely met: "success" is when Sophie picks the "correct" hand and "failure" is when she does not, and there are 12 trials.

If you believe that Sophie is learning during these trials, then the I and S assumptions of BINS are not met. If you do not believe that she is learning, then the I assumption of BINS is met (the trials are independent) and it's reasonable to assume the S is too (she has the same chance of picking the "correct" hand on each trial).

Either way, we only have 7 successes and 5 failures, so we should do a binomial exact confidence interval:




*Confidence Interval for One Population Proportion*

Success = Correct
Sample Size = 12
Number of Successes = 7
Proportion of Success = 0.5833
Confidence level = 95%

| Method | Lower CL | Upper CL | Midpoint | Width |
|---|---|---|---|---|
| Binomial (Exact) | 0.276670 | 0.848348 | 0.562509 | 0.571678 |

If you believe that all four of the BINS assumptions are met, you can interpret as: We are 95% confident that Sophie will get between 27.7% and 84.8% correct in the long run. Since this interval includes 50%, it is certainly plausible that Sophie is picking a hand entirely at random.

If you believe that only two of the BINS assumptions are met, you can interpret as: The confidence interval we obtained for the proportion of the time Sophie is correct in the long run is (0.277, 0.848). However, because the assumptions necessary for the inference to be valid are not met, this interval has no real-world meaning. Sophie may be learning to not pick a hand entirely at random, and we do not have an inferential framework to account for this.

4. The *lumefantrine.csv* dataset contains the concentrations (in ng/mL) of the anti-malarial drug lumefantrine two hours after ingestion in 31 pregnant women. The variable Vc_2h contains the concentrations in venous plasma (veins) and the variable Cc_2h contains the concentrations in capillary plasma.

Researchers would like to predict the concentration of lumefantrine in capillary plasma (response variable) from the concentration of lumefantrine in venous plasma (predictor variable). When researchers make a scatterplot of the data for the 31 women, they notice that the points have an approximately linear relationship and that the slope of the regression line is approximately 1.

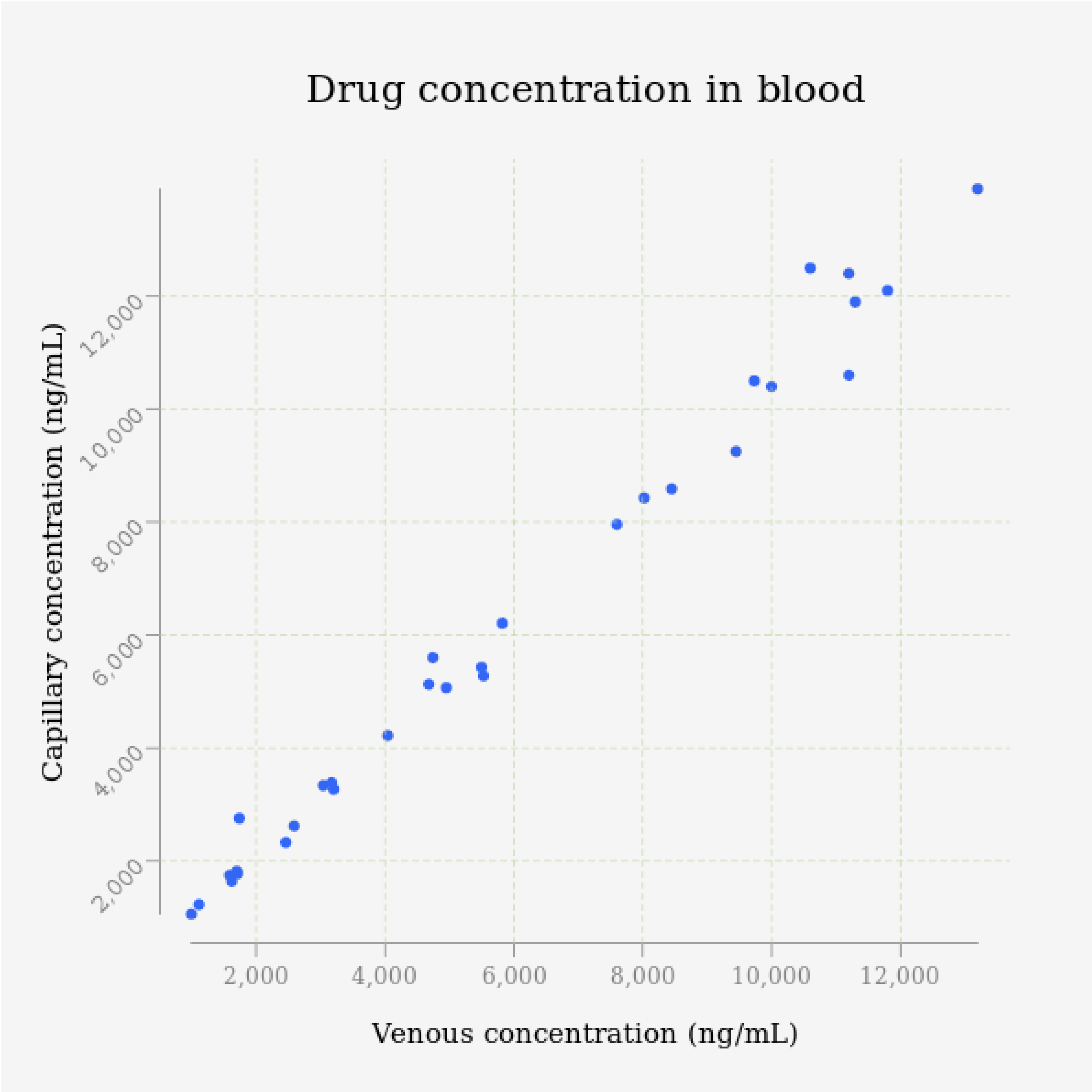
Suppose that a new pregnant woman is given a dose of lumefantrine and, two hours later, the concentration in her venous plasma is measured to be 1700 ng/mL. Predict with 95% confidence the concentration of lumefantrine in her capillary plasma. Do you have faith in this method of prediction, or are assumptions necessary for this prediction violated?

Full credit will be given in this problem for a writeup that:

[1 pt] Contains relevant, necessary and correct (R code and) software output

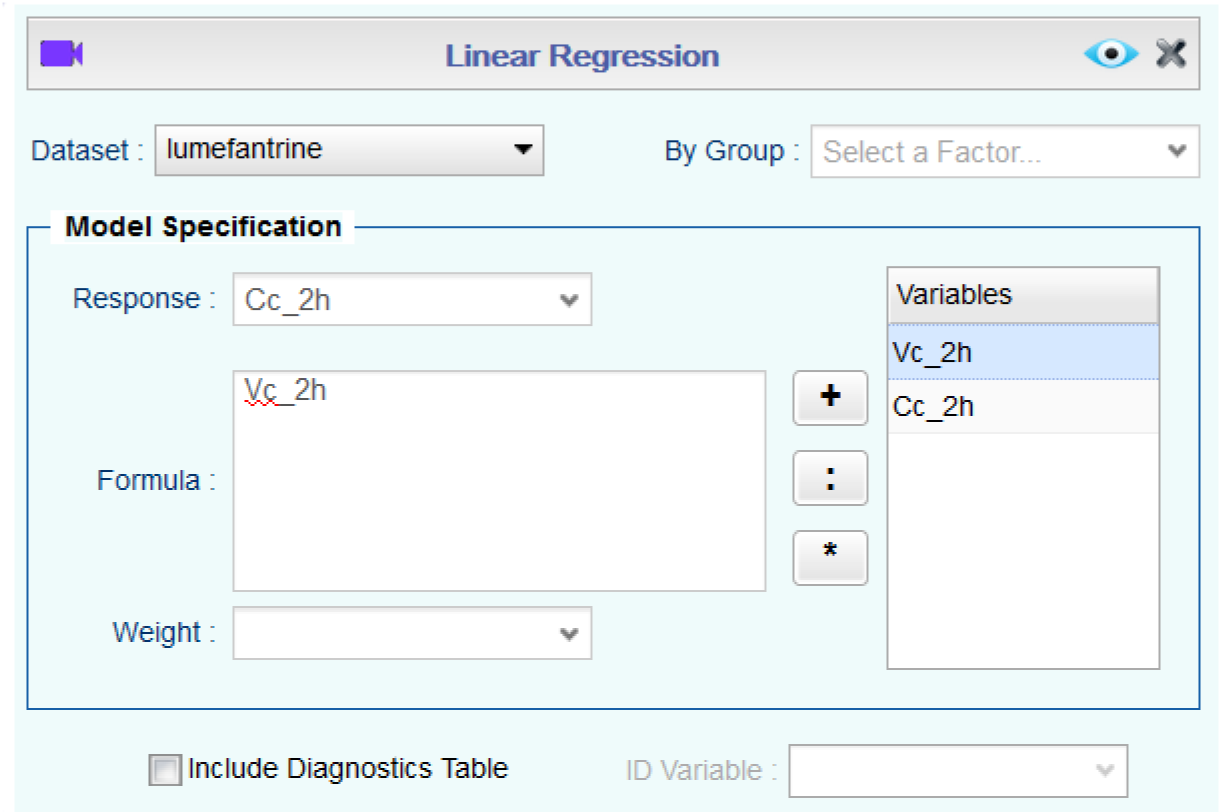
[1 pt] Provides a correct interval estimate from the output

[2 pts] Includes appropriate background and/or exploratory analysis and arrives at a reasonable conclusion about the appropriateness of the inference




A linear model certainly seems appropriate based on the scatterplot and the information in the problem.

We fit the model:




*Regression Coefficients Estimates*

Model: Cc_2h ~ Vc_2h

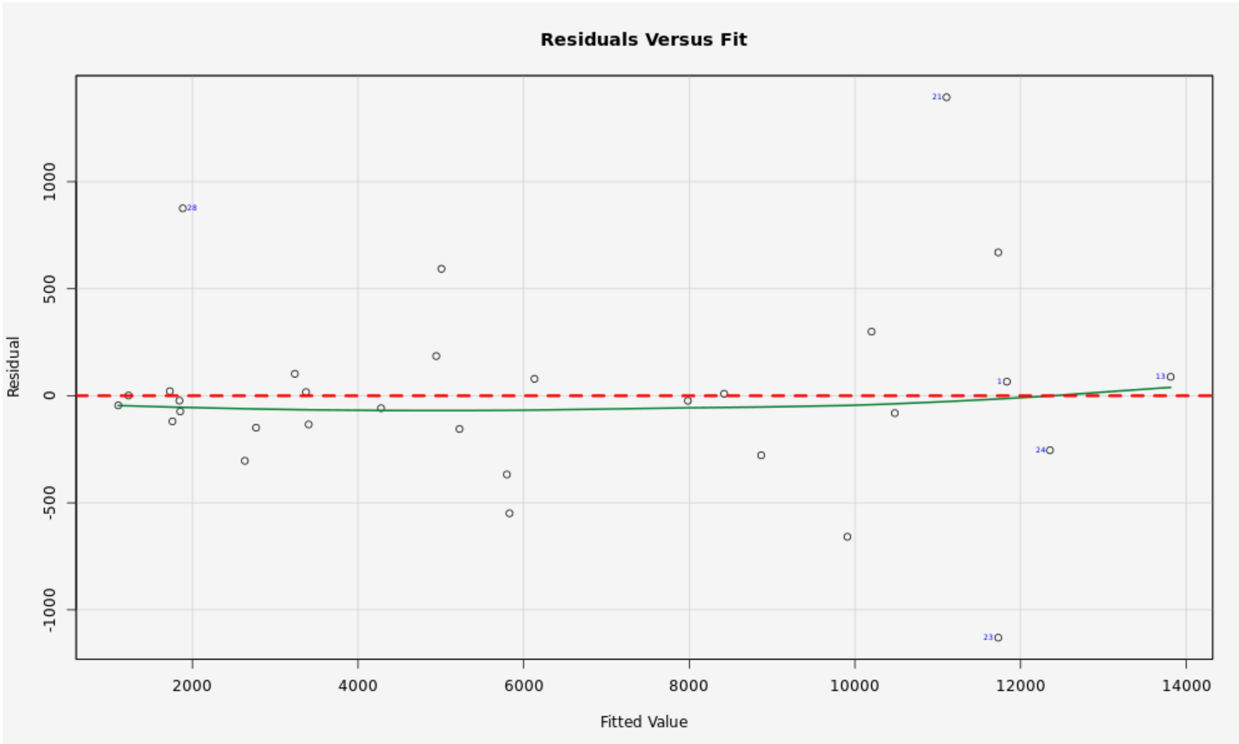
| Term | Coefficient Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| (Intercept) | 73.5420 | 153.129 | 0.480261 | 0.634645 |
| Vc_2h | 1.04077 | 0.0218665 | 47.5963 | 4.41191e-29 |

*Model Summary: R-Square*

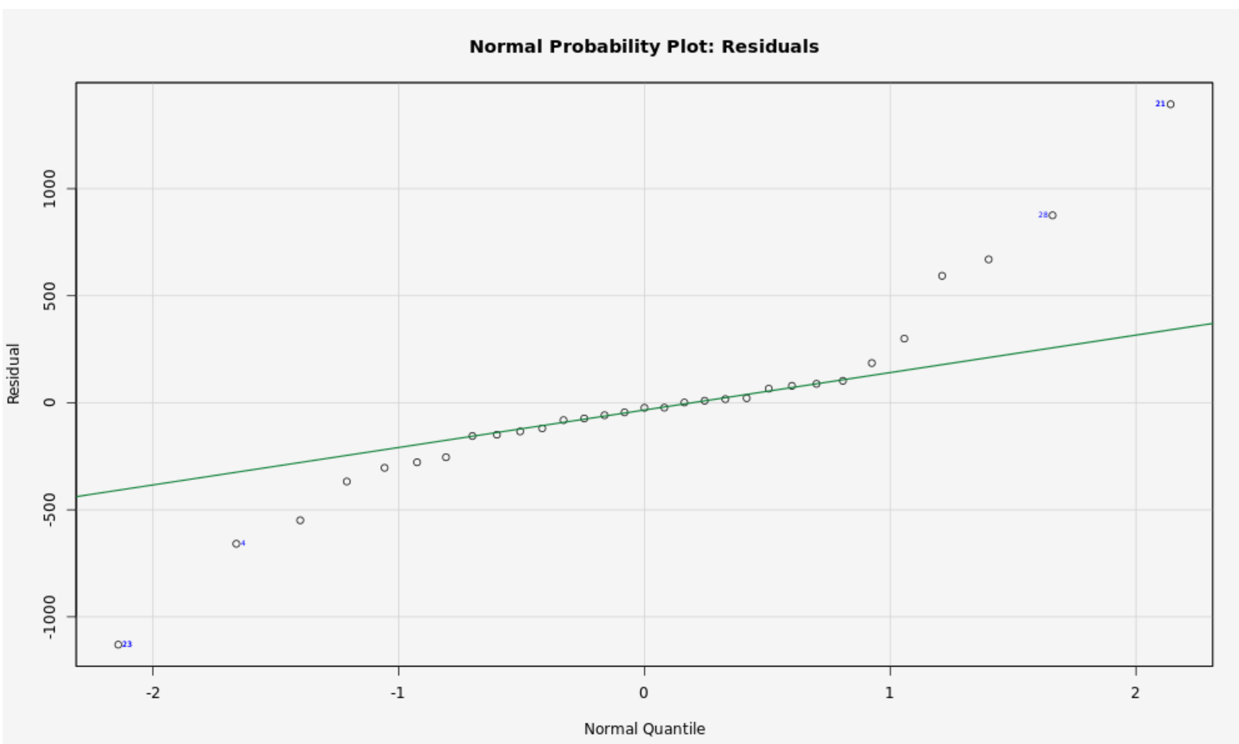
Model: Cc_2h ~ Vc_2h

| Residual Standard Error | DF | R-Square | Adjusted R-square |
|---|---|---|---|
| 460.314 | 29 | 0.987361 | 0.986925 |

We have a very good $R^2$ value of 0.9874. Let's see if we see any issues with zero mean, constant variance or normality:




Zero mean looks okay, but constant variance seems to be a bit of a problem; the residuals look like they tend to be larger at high fitted values.




Normality is definitely an issue.

The problem asks us to get a 95% prediction interval anyway:

## *Case-wise Estimates and Diagnostics*

Model: Cc_2h ~ Vc_2h

| Obs | Cc_2h | Vc_2h | Predicted Values | Residuals | Std. Error Mean Predict | Lower Mean 2.5% | Upper Mean 97.5% | Lower Pred 2.5% | Upper Pred 97.5% |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 11900 | 11300 | 11834.2 | 65.8031 | 144.243 | 11539.2 | 12129.2 | 10847.6 | 12820.8 |
| 2 | 2330 | 2460 | 2633.83 | -303.826 | 111.693 | 2405.39 | 2862.26 | 1665.06 | 3602.59 |
| 3 | 1820 | 1700 | 1842.84 | -22.8441 | 123.482 | 1590.30 | 2095.39 | 868.111 | 2817.58 |

The interval we want is in line 3. If the assumptions were anywhere close to met, we would claim with 95% confidence that the concentration of lumefantrine in the woman's capillaries was between 868 and 2818 ng/mL. However, our assumptions of constant variance and normality of the residuals are not met, and so we have little faith in our prediction.