STAT 240 Practice Final Exam

1st Letter of Last/Family Name	Last/Family Name as in Canvas	First/Given Name as in Canvas	Student ID

Instructor (Circle) **Bret Larget** **Bi Cheng Wu** **Hamna Hannan**

Lecture Time (Circle) MWF 8:50 - 9:40 MWF 1:20 - 2:10 MWF 2:25 - 3:15 TH 8:00 - 9:15

Discussion (Circle Name and Time)

TA	time 1	time 2	time 3
Shane Huang	T 7:45 am	T 8:50 am	T 9:55 am
Christian Varner	M 2:25 pm	M 3:30 pm	M 4:35 pm
Cameron Jones	M 2:25 pm	M 3:30 pm	M 4:35 pm
Ryan Yee	M 2:25 pm	M 3:30 pm	M 4:35 pm
Congwei Yang	T 7:45 am	T 12:05 pm	Tue 1:20 pm
Jingyang Lyu	T 7:45 am	T 12:05 pm	W 7:45 am
Nathaniel Pritchard	W 7:45 am	W 8:50 am	W 4:35 pm
Haoran Xiong	T 4:35 pm	W 7:45 am	W 4:35 pm

Instructions:

1. You may use both sides of two regular sheet of paper with self-prepared notes.
2. You may not consult other resources, your phone, a computer, online info, nor your neighbor's exam.
3. Do all of your work in the space provided. Use the backs of pages if necessary, indicating clearly that you have done so (so the grader can easily find your complete answer).

Multiple Choice.

Problem 1. Which aesthetic changes the transparency of a plotted point? **Circle one answer.**
(a) `alpha` (b) `shape` (c) `size` (d) `x`

Problem 2. Which command do you use to eliminate some rows from a tibble? **Circle one answer.**
(a) `arrange()` (b) `bind_rows()` (c) `filter()` (d) `select()`

Problem 3. Which lubridate command will convert the string "3/5/2021" into the date May 3, 2021? **Circle one answer.**
(a) `as_date()` (b) `date()` (c) `dmy()` (d) `mdy()`

Problem 4. Which command could you use to reshape a tibble by turning the unique values in one variable into names of new columns and the values in a second variable into the values in these new columns? **Circle one answer.**
(a) `group_by()` (b) `pivot_longer()` (c) `pivot_wider()` (d) `summarize()`

Problem 5. The expression `x %>% str_detect( "[:digit:]{2,3}-\\++//?$")` evaluates as `TRUE` for only one of the strings below if this string replaces `x` in the expression. Which one? **Circle one answer.**
(a) `"abc00-++"` (b) `"abc00-++/"` (c) `"123-\\+++//?"` (d) `"12-+++++///"`

Problem 6. What is the type of the output produced by the **purrr** command `map()`?
Circle one answer.
(a) character (b) data.frame (c) list (d) numeric

Problem 7. What is the output value of this expression after evaluation?

```
f = function(a, b=2, c=3) { return ( a + 2*b - c) }
f(1, c=2)
```

Circle one answer.
(a) 0 (b) 1 (c) 2 (d) 3

Problem 8. What numerical value is likely to be very close to the value produced by the command `sd(rbinom(10000, 100, 0.5))`? **Circle one answer.**
(a) 5 (b) 25 (c) 50 (d) 5000

Problem 9. One uses the Agresti-Coull method to construct a 95% confidence interval for an unknown population proportion p from data where the observed proportion $\hat{p} = 0.4$. The center of this confidence interval: **Circle one answer.**
(a) is less than 0.4 (b) is equal to 0.4 (c) is between 0.4 and 0.5 (d) cannot be determined

Problem 10. A two-sided Welch t-test test of the null hypothesis $H_0 : \mu_1 = \mu_2$ where μ_1 and μ_2 are two populations means and data is collected by random sampling independently from each population results in a t test statistic of $t = 2.36$ and degrees of freedom 67.3 when using the R function `t.test()`. The corresponding p-value is equal to the numerical value produced by which expression? Only one is correct. **Circle one answer.**

(a) `pt(2.36, 67.3)` (b) `2*pt(2.36, 67.3)` (c) `2*pt(2.36, 67.3, lower.tail = TRUE)`
(d) `2*pt(-2.36, 67.3)`

Problem 11. The calculated correlation coefficient between two quantitative variables is -0.97. **Circle each correct response, of which there may be more than one. Cross out responses which are incorrect.**

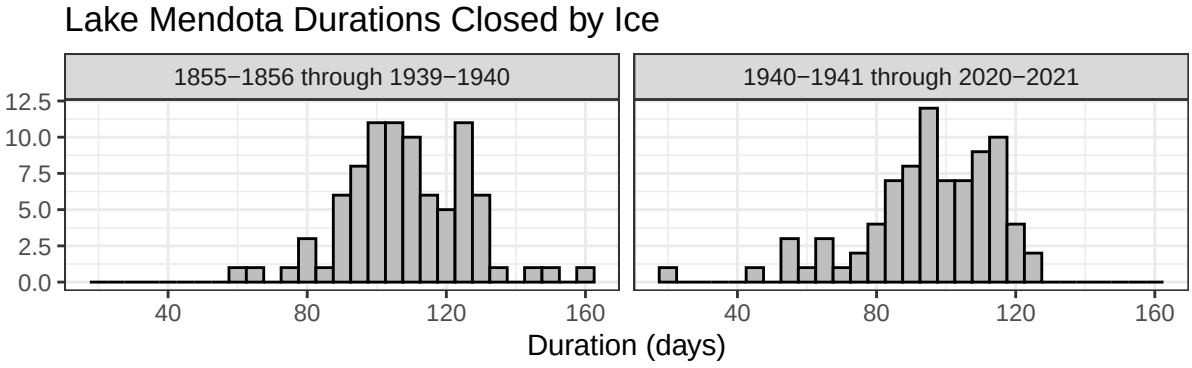
(a) The variables are negatively associated.
(b) A plot of the points shows they are tightly clustered around a line with a negative slope.
(c) There is an error in the calculation as the correlation coefficient cannot be negative.
(d) It is possible for a plot of the points to fall exactly on a smooth curve that is not straight, such as a parabola.

Problem 12. A simple linear regression line to predict y from x from data for these two quantitative variables is $\bar{y} = 9.7 - 4.3x$. **Circle each correct response, of which there may be more than one. Cross out responses which are incorrect.**

(a) the correlation between x and y is negative
(b) the regression line passes through the point $(\bar{x}, \bar{y})$
(c) if we solved this equation for x, this would match the regression coefficients we would obtain if we would use regression to estimate coefficients for the line $\hat{x} = a + by$
(d) the value $\hat{y} = 9.7$ will be a reasonable prediction for an observation where $x = 0$ in all situations.

Short Answer Problems.

Problem 13



The surface of Lake Mendota freezes each winter. The graph above plots the total durations in days for which the surface of Lake Mendota is at least half covered by ice. Use each of the values 19, 95, 109, 132 once to fill in the missing values in the table below.

period	mean	sd	95th percentile
1855-1856 through 1939-1940		18	
1940-1941 through 2020-2021			119

Problem 14

In a study of obesity among adult Wisconsin residents from 2015, the percentage of obese individuals is recorded in the following table by sex and age range. This data summary projects obesity rates from sampled individuals by age, sex, and zip code to the corresponding population of the same before summarizing by age and sex for the entire state.

```
prob14
```

```
## # A tibble: 8 x 3
## # Groups:   sex [2]
##   sex     age   pct_obese
##   <fct>   <chr>     <dbl>
## 1 Females 18-34      31.6
## 2 Females 35-54      44.7
## 3 Females 55-74      45.8
## 4 Females 75+        32.3
## 5 Males   18-34      29.2
## 6 Males   35-54      47.9
## 7 Males   55-74      48.5
## 8 Males   75+        33.7
```

Fix the block of code by crossing out the incorrect choice between braces `{` and `}` in order to reshape this data so that there are four rows, one for each age group, with a column `age` with the age range and columns `Females` and `Males` with the corresponding obesity percentages as values.

```
prob14 %>%
  pivot_{longer, wider}(
    names_{from, to} = {sex, age, pct_obese},
    values_{from, to} = {sex, age, pct_obese}
)
```

Problem 15

To construct a 90% confidence interval for a population parameter p from a sample of 500 trials,, calculate a point estimate plus or minus a margin of error. The margin of error is calculated as the quantile from a reference distribution times an estimate standard error based on the data.

A What reference distribution and quantile is most appropriate?

B Write an R expression which calculates this value.

Problem 16

Over a 150 year period of time, the surface of Lake Mendota is more than 50% covered by ice for $\bar{y} = 102.0$ days on average with a standard deviation of $s_y = 19.7$ days. The correlation coefficient of this variable y with the first year x of the corresponding winter is $r = -0.48$. In a simple linear regression model to predict y from x, write an expression showing how to calculate the predicted duration $\hat{y}$ in a year which is 1.2 standard deviations later than the mean first year of $\bar{x} = 1938$.

Problems

Problem 17

Double check the front page. Verify that:

- The first letter of your family name is printed clearly in the box.
- Your family and given names are printed clearly and legibly in the correct spaces.
- You students ID number is printed clearly
- You have **circled** the lecture and discussion section in which you are enrolled accurately.

Full credit if your information is complete, accurate, and legible. Partial credit if we have trouble reading your printing or if we cannot find your name easily in Canvas.

Problem 18

Treat the data from Problem 16 as a random sample of rallies played between Minnesota and Wisconsin from a population of all rallies which might have occurred had the matches occurred differently. The observed data saw Wisconsin winning 268 points from the 498 rallies played. Let p be a hypothetical probability that Wisconsin wins a point versus Minnesota during a rally in this volleyball season.

A What four assumptions are made to justify a binomial model for X, the total number of points won by Wisconsin versus Minnesota in the 498 rallies played?

B Using the Agresti-Coull method, write an expression for a 95% confidence interval for p if we were willing to assume that the binomial model was reasonable. Do not make any numerical calculations. For example, you might write something in a form similar to this.

$$\frac{100}{200} \pm 2.5 \times \left(\frac{10}{100}\right)^2$$

but with different numbers and perhaps different mathematical expressions.

C Assume that your expression from the previous problem can be correctly calculated as 0.538 ± 0.044, or the interval from 0.494 to 0.582. Write an interpretation of this confidence interval in context, following examples modelled during lectures.

Problem 19

Consider the same setting as the previous problem where the probability that Wisconsin scores a point from a rally in a match against Minnesota is p and the data is that Wisconsin scored 268 points from 498 rallies. For this problem, we will test the null hypothesis that $p = 0.5$ versus the alternative that p is something different.

A Use conventional notation as described in class and write the null and alternative hypotheses.

B A hypothesis test introduced in class uses the observed count of X, the number of points scored by Wisconsin, as the test statistic. For the purposes of this test, condition on the fact that there were a total of $n = 498$ trials this year.

Name the distribution of X if the null hypothesis is true, providing numerical values for all parameters, and write expressions for how to calculate the mean and standard deviation of this distribution.

C Write an R expression for how to calculate the p-value for this test

D Assume that the calculated p-value for this test is $p = 0.097$. Write a summary conclusion about this hypothesis test in the context of the problem, following examples as described in the lectures.

Problem 20

The Boston Marathon is a footrace of over 26 miles, a distance over 40 km. In addition to an overall time, competitors receive official splits times for various 5 km segments. This problem compares the completion times for male runners aged 18–34 during the first and last full 5 km segments of the race for a random sample of $n = 100$ runners in the 2010 race. We test if runners run these segments in the same time on average. Data for these runners are summarized below: `x` is the time in the first segment; `y` is the time in the second segment; `d` is the difference for each runner.

```
## # A tibble: 3 x 4
##   sample     n  mean    sd
##   <chr>  <int> <dbl> <dbl>
## 1 x        100 22.8   3.97
## 2 y        100 27.2   6.45
## 3 d        100 -4.44  4.65
```

Here is are two different calls to the `t.test()` function for this data.

```
## First t-test
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -5.8667, df = 164.71, p-value = 2.373e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.940464 -2.948736
## sample estimates:
## mean of x mean of y
##   22.7663   27.2109
```

```
## Second t-test
t.test(x, y, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -9.5584, df = 99, p-value = 1.006e-15
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.367248 -3.521952
## sample estimates:
## mean difference
##         -4.4446
```

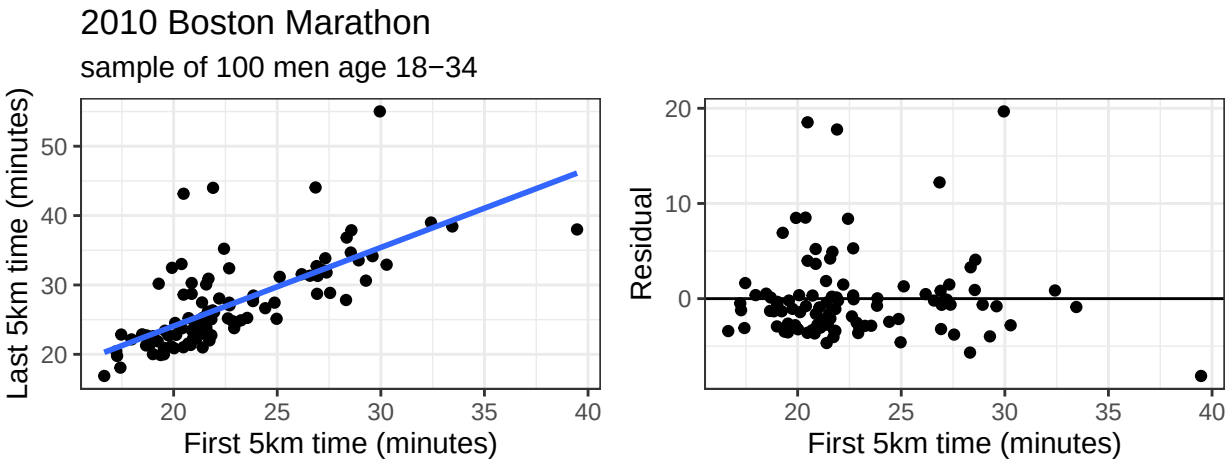
A Explain why one test is appropriate and the other is not

B Report the test statistic and the probability distribution used to calculate the p-value

C Write a summary of the conclusions of the appropriate hypothesis test in the style modeled in class with a statement in context and minimal jargon followed by a summary of the statistical evidence to justify this conclusion on parentheses.

Problem 21

This problem uses the same data from the previous problem. The scatter plot shows a graph of the two split times for the sample of 100 runners. The second plot shows residuals versus the first split time.



Here is a tidy summary of the fitted regression model.

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     1.43     2.71      0.528 5.99e- 1
## 2 x               1.13     0.117     9.64  7.20e-16
```

A Write an equation of the fitted regression line using `x` for the time in the first 5km and `y` for the last time.

B Which of the following assumptions about regression does the residual plot most cast doubt on? **Cross out all assumptions but one.**

(1) linearity (2) independence (3) normally distributed errors (4) equal variances

C Write a confidence interval for the slope of a true regression line between these variables in the population of all male Boston Marathon runners aged 18-34. Use numbers where you can and briefly explain how to find a number that is not provided. Do not simplify any calculations. Do not interpret.