Answer Key (Complete)

Fall 2023 STAT 240 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**

Lecture (Circle) MWF 8:50–9:40 MWF 9:55–10:45 MWF 2:25–3:15 MWF 3:30–4:20

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).

Scoring

Question	Name/Course	1–3	4–8	9–12	13	14	15	16	Total
Points									
Possible	2	12	20	16	12	13	11	14	100
		14	34	50	62	75	86	100	

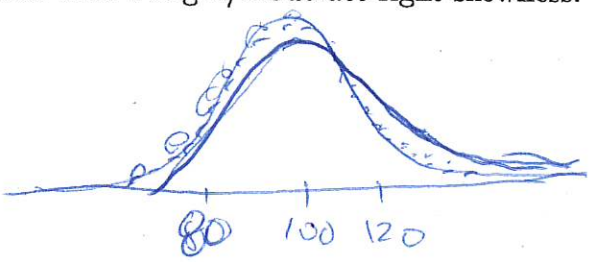
Multiple Choice and Short Answer. (4 points each)

For each multiple choice problem, circle the letter for all correct answers and cross out the letter for all incorrect answers. Answer briefly for other problems.

Problem 1. Sketch the density plot of a normal distribution, including labeling the x axis, where the mean and median are 100 and the standard deviation is about 20, using a dotted or dashed line. Over this, sketch another density plot using a solid line with the same median and about the same standard deviation, but with a slight/moderate right skewness.

2 for normal
2 for other — wide right tail

4



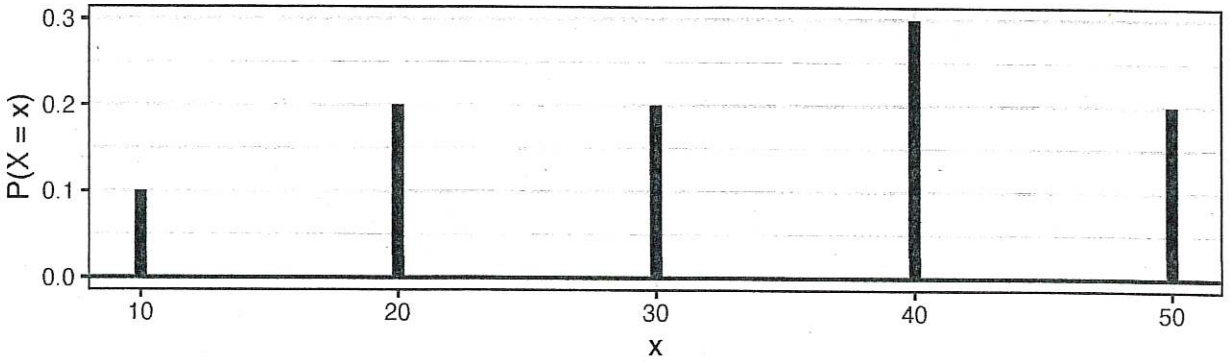
Problem 2. A data set `bm` has Boston Marathon data from the year 2010 with a row for each runner who completed the race and variables `Age`, `Age_Range`, and `Time`. Identify code that calculates the mean time for all runners between the ages of 35 and 39. Note: **`Age_Range` equals "35-39" when `Age` is in this range and code which calculates the desired mean along with other things should be circled as correct.**

Circle correct answers and cross out incorrect answers.

1 each

(a) `bm %>% filter(Age_Range == "35-39") %>% summarize(mean = mean(Time))`
(b) `bm %>% group_by(Age_Range) %>% summarize(mean = mean(Time))`
(c) `bm %>% mutate(mean = mean(Time)) %>% filter(between(Age, 35, 39))`
(d) `bm %>% select(Age_Range == "35-39") %>% summarize(mean = mean(Time))`

Problem 3. The probability mass function of a discrete random variable X is plotted here. It has a mean μ and a standard deviation σ.



1 each

Write the following four numbers under their corresponding values: 12.7, 33, 40, 70

μ	σ	0.6 quantile	$100 \times P(X > 20)$
33	12.7	40	70

Problems 4 and 5.

A data frame `matches` contains 4746 rows, one for each match between two teams, with the variables `index`, `W`, and `L`, where `index` is the row number and `W` and `L` have the names of one of 332 teams that won and lost a match, respectively. Each of the 332 teams appears at least once in columns `W` and `L`. The data frame `ncaa` has 64 rows and columns `Team` and `Conference` where each value in `Team` is distinct and is one of the same the 332 team names in `matches`. `Conference` is another categorical variable with 32 distinct values.

The data frame `df` is created by the following code.

```
df = matches %>%
  pivot_longer(W:L, names_to = "Result", values_to = "Team")  %>%
  count(Team, Result) %>%
  pivot_wider(names_from = Result, values_from = n) %>%
  semi_join(ncaa, by = "Team")
```

Problem 4. How many rows are in `df`?

(a) 1 (b) 32 (c) 64 (d) 332 (e) 4746

Problem 5. List the column names (in any order) in `df`.

Team, L, W

Problem 6. A random variable X is created by adding together the number of heads in five tosses of a fair coin plus the number of tails in a different set of five tosses of the same coin. **Circle correct answers and cross out incorrect answers.**

(a) X has a binomial distribution
(b) X is not binomial because the number of trials is not fixed
(c) X is not binomial because the trial success probability changes
(d) X is not binomial because the trials are not independent.

Problem 7. When constructing a 95% confidence interval for a single population proportion from sample of size $n = 105$, the margin of error is some quantity a times an estimated standard error. How is the value of a determined? **Circle all correct answers and cross out incorrect answers.**

(a) `qnorm(0.95)` (b) `qnorm(0.975)` (c) `qt(0.975, 104)` (d) `qt(0.975, 103)`

Problem 8. In the test of a hypothesis test for a population proportion p with $H_0 : p = 0.5$ versus $H_a : p > 0.5$, the p-value is equal to 0.043. **Circle correct answers and cross out incorrect answers.**

(a) We have proven that $p > 0.5$.
(b) The probability that $p > 0.5$ is more than 95%.
(c) There is evidence that $p > 0.5$.
(d) If we had tested with the two-sided alternative hypothesis, the test would have been statistically significant at the $\alpha = 0.05$ level.

Problem 9. Put the following four quantities in order from smallest to largest.

(a) `qnorm(0.1)` (b) `qt(0.1, 5)` (c) `qt(0.1, 10)` (d) `qt(0.1, 100)`

qt(0.1,5) < qt(0.1,10) < qt(0.1,100) < qnorm(0.1)

4 -2 if opposite order | -2 if qnorm < qt(), but qt correct
-2 if qt() order wrong

Problem 10. The correlation coefficient between the average height in inches (plotted on the x axis) and weight in pounds (plotted on the y axis) of a sample of 100 people is $r = 0.68$. **Circle correct answers and cross out incorrect answers.**

4 1 each

(a) If height were measured in feet instead of inches, the value of r would be $r = 0.68/12$.
(b) In this sample, relatively tall people tend to weigh more than relatively short people do.
(c) 68% of the points fall exactly on a straight line.
(d) If we switched the axes, the new correlation coefficient would be equal to -0.68.

Problem 11. A linear regression model predicts weight from height from a sample of 100 people with a mean height of 67 inches. The correlation coefficient is $r = 0.68$. How much taller than average is the predicted height of a person who is 73 inches tall if the standard deviation of heights in the data is 4 inches? *(Note that 73 is 6 inches above the mean height of 67 inches.)*

4

73 is $z = 1.5$ sd above the mean

weight is (0.68)(1.5)(30) ~~inches~~ pounds above average

Fix

2 **Problem 12a.** Using the same setting as the previous problem: Which of the following intervals is the widest? **Circle the correct answer and cross out the incorrect answers.**

(a) A 95% confidence interval for the mean weight of all people who are 65 inches tall.
(b) A 95% confidence interval for the mean weight of all people who are 73 inches tall.
(c) A 95% prediction interval for the weight of a single individual who is 65 inches tall.
(d) A 95% prediction interval for the weight of a single individual who is 73 inches tall.

2 **Problem 12b.** Using the same setting as the previous problem: Which of the following intervals is the narrowest? **Circle the correct answer and cross out the incorrect answers.**

(a) A confidence interval for the mean weight of all people who are 65 inches tall.
(b) A confidence interval for the mean weight of all people who are 73 inches tall.
(c) A prediction interval for the weight of a single individual who is 65 inches tall.
(d) A prediction interval for the weight of a single individual who is 73 inches tall.

Problem 13 (12 points)

In a certain genetics experiment with fruit flies and simple recessive traits a and b, the probability of offspring with the double recessive genotype *aabb*, call the probability p, is expected to be $1/16 = 0.0625$ if the traits are *unlinked*. If the traits are *linked*, then $p < 0.0625$.

In the experiment, there are 500 offspring. Let X be the number of offspring with genotype *aabb*. Assume that the genotypes of all offspring are independent of one another.

4 **(a)** Is it reasonable to assume that $X \sim \text{Binomial}(500, p)$ for some p? Briefly explain.

Yes – BINS
binary (1 pt)
independent (1 pt)
fixed sample size (1 pt)
same probability (1 pt)

4 **(b)** Assume that $X \sim \text{Binomial}(500, 0.0625)$. Write an R expression to calculate the exact probability $P(X \leq 28)$.

pbinom(28, 500, 0.0625)

-1 if wrong order

4 **(c)** If you wanted to approximate the probability in (b) with an area under a normal curve using the function `pnorm(x, m, s)`, write expressions to calculate the values of `x`, `m`, and `s` to do this calculation accurately. You do not need to simplify these expressions.

x = 28.5

m = 500 * 0.0625

s = $\sqrt{500 * 0.0625 * (1 - 0.0625)}$

Problem 14 (13 points)

Assume the same setting as in Problem 13. Suppose that in the genetic cross, 28 out of 500 offspring have the genotype *aabb*. For context, $28/500 = 0.056$.

4 **(a)** Write an expression for a 95% confidence interval for p, the probability of the genotype *aabb* in the genetic cross using the Agresti-Coull method. You do not need to simplify any numerical expressions.

$\tilde{p} = \frac{30}{504}$

$\frac{30}{504} \pm 1.96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{504}}$

1 right $\tilde{p}$
1 1.96, not t or other# qnorm(0.975) ok
1 est ± mult * SE
1 SE correct

3 **(b)** In a test of the hypothesis $H_0 : p = 0.0625$ versus the one-sided alternative $H_a : p < 0.0625$, write an R expression to calculate the p-value.

pbinom(28, 500, 0.0625)
or sum(dbinom(0:28, 500, 0.0625))
-1 if normal, but correct otherwise

3 **(c)** Suppose that the calculated p-value is equal to 0.313. **Circle the single letter label (A–C, below) of all appropriate conclusions from statistical inference in context and cross out those not supported by the data analysis.**

1 each Recall that if the traits are unlinked, the probability of genotype *aabb* is exactly equal to 0.0625 and if the traits are linked, then this probability is smaller than 0.0625.

~~A.~~ There is strong evidence that the genetic traits are unlinked.

~~B.~~ There is strong evidence that the genetic traits are linked.

(C.) The observed data is consistent with the genetic traits being unlinked.

3 **(d)** Without doing any numerical calculations, do you think that the upper limit of the 95% confidence interval from the correct solution to (a) above is larger than or smaller than 0.0625? Briefly justify your response by referring to your answer in part (c).

Larger than 0.0625. 1 for answer, 2 for justification
Large p-value in (c) implies data consistent with H_0: $p = 0.0625$, so we expect 0.0625 to be in the interval, meaning right endpoint > 0.0625
1 if backwards reasoning

Problem 15 (11 points)

Treat the calendar years from 1870–1899 as a sample of 30 years from a time period in the late 1800s and the calendar years from 1990–2019 as a sample of 30 years in a more recent time period. Consider the expected monthly population mean temperature in December for each of these years, with μ_1 representing a late 1800s population mean December temperature and μ_2 representing a recent population mean December temperature, in each case ignoring the effects of random annual temperature fluctuations, but instead representing unobserved climate conditions. The following table summarizes the sampled temperature data.

period	n	mean	sd
late 1800s	30	22.71	7.00
recent	30	25.17	5.25

Here is the output of the function `t.test()` using these two samples.

```
    Welch Two Sample t-test

data:  late_1800s and recent
t = -1.5399, df = 53.757, p-value = 0.1295
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.6636964  0.7432663
sample estimates:
mean of x mean of y
 22.71344  25.17366
```

3 (a) Using values in the summary data table and other R functions (such as `qnorm()` or `qt()`) if needed, write an expression which shows how to compute the upper endpoint of the confidence interval, 0.7432663. Do not simplify or evaluate the expression.

$(22.71 - 25.17) + qt(0.975, 53.757)\sqrt{\frac{7.00^2}{30} + \frac{5.25^2}{30}}$

1 pt 1 pt 1 pt

3 (b) Using values in the summary data table and other R functions (such as `qnorm()` or `qt()`) if needed, write an expression which shows how to compute the value $t = -1.5399$.

$$t = \frac{22.71 - 25.17}{\sqrt{\frac{7.00^2}{30} + \frac{5.25^2}{30}}}$$

-3 if qt(...)
-1 if num wrong
-1 if SE in denom wrong

This problem continues on the next page.

5

(c) Circle the single letter label (A–E, below) of all appropriate conclusions and **cross out those not supported by the data analysis.**

A. There is strong evidence that the observed difference in average December temperatures, which is about 2.5 degrees Fahrenheit higher in recent years than in the late 1800s, cannot plausibly be explained by random annual fluctuations in weather, providing evidence that a changing climate is making December warmer in Madison.

B. There is strong evidence that the mean temperature in December 2023 will be higher than what the mean temperature in Madison was in December 1900.

C. There is strong evidence that the observed mean temperature in December between 1870 and 1899 is exactly equal to that observed between 1990 and 2019.

D. The observed data is consistent with random annual temperature variation alone explaining the observed difference in mean December temperature between the 1800s and more recently.

E. A change in climate could, in part, explain part of the observed differece in mean December temperatures between the two time periods.

Problem 16 (14 points)

Researchers have a theory that mammals sleep to heal brain cell damage. This theory suggests a power law relationship between the ratio of the average daily sleep time versus awake time and body mass, or

$$(\text{sleep ratio}) = C \times (\text{body mass})^{-\theta}$$

where $0.16 \leq \theta \leq 0.19$ (for reasons argued in the paper). In contrast, if sleep's primary function is for whole body cellular repair, the researchers expect that θ will be closer to 0.25. Taking natural logs on both sides of the equation results in the equation

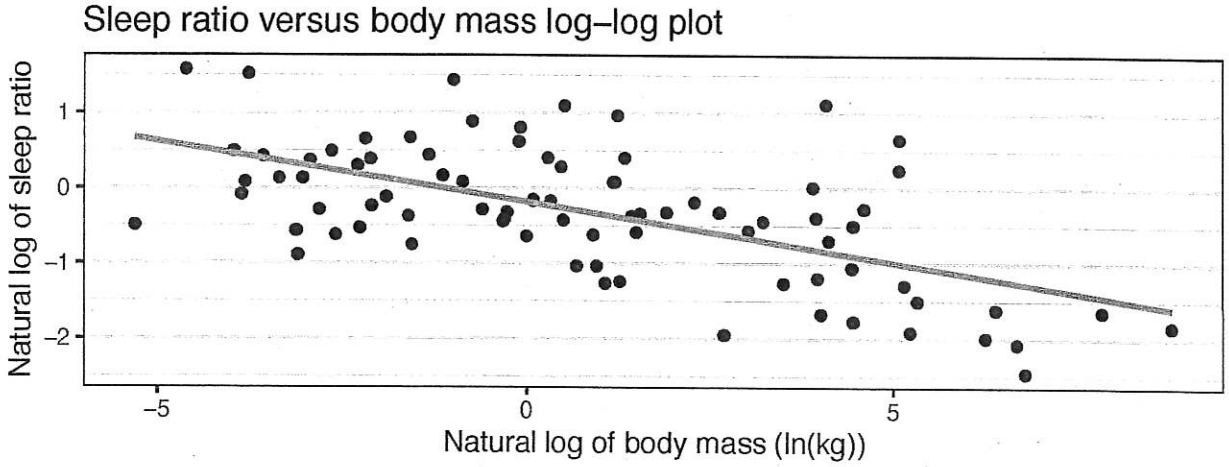
$$\ln(\text{sleep ratio}) = \ln C + (-\theta) \times \ln(\text{body mass})$$

(Note that the slope of this equation is $-\theta$.)

Sleep ratio is a positive number and has no units. For example, an animal that sleeps 16 hours and is awake 8 hours per day has a sleep ratio of 2. Body mass is measured in kilograms. All values are averages for entire species or other animal group.

Study data is from 83 animal groups that span several orders of magnitude in body sizes, ranging from small shrews (average body mass equals 0.005 kg or 5 grams) to African elephants (average body mass equals 6654 kg). Small animals tend to sleep much longer (have larger sleep ratios) than larger animals.

The following graph shows the relationship between the natural logs of this sleep ratio (x) versus the natural log of the body mass (y) for these 83 animal groups.




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17977    0.08046  -2.234   0.0282 *
x           -0.16071    0.02405  -6.681 2.76e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7093 on 81 degrees of freedom
Multiple R-squared:  0.3553,    Adjusted R-squared:  0.3473
F-statistic: 44.64 on 1 and 81 DF,  p-value: 2.764e-09
```

Answer questions on the following page

3 (a) Write an R expression to calculate the upper and lower limits of a 95% confidence interval for θ. Use numbers from the regression summary on the previous page when possible and code when needed (such as using either `qnorm()` or `qt()`).

0.16071 ± c(-1,1) * qt(0.975,81) * (0.02405)

3 (b) Write an expression to calculate a test statistic for the hypothesis test $H_0 : \theta = 0.25$ versus the alternative $H_a : \theta \neq 0.25$.

$$t = \frac{0.16071 - 0.25}{0.02405}$$

3 (c) Assume that the value of the test statistic in (b) equals -3.71. Write an R expression to calculate the p-value of this hypothesis test.

2 * pt(-3.71, 81)

-2 if not t
-1 if wrong df
-1 if not doubled

5 (d) Assume that the p-value for this hypothesis test calculated in (c) is $p = 0.00038$ and that the numerical limits of the confidence interval calculated in (a) are 0.113 and 0.209. **Circle the single-letter label (A–E, below) of all appropriate conclusions from statistical inference in context and cross out those not supported by the data analysis.**

1pt each

X A. The observed data is consistent with $\theta = 0.25$, implying that sleep's primary function in mammals could be whole body cellular repair.

X B. There is strong evidence that $\theta = 0.25$ and that the primary function of sleep in mammals is whole body cellular repair.

(C) There is strong evidence that $\theta < 0.25$, suggesting that whole body cellular repair is not the primary function of sleep in mammals.

X D. There is strong evidence that $0.16 \leq \theta \leq 0.19$, implying that the primary function of sleep in mammals is to heal brain cell damage.

(E) The observed data is consistent with $0.16 \leq \theta \leq 0.19$, consistent with the biological hypothesis that the primary function of sleep in mammals is to heal brain cell damage.