Lecture Section: _____ Name:_____

Read the following directions carefully. DO NOT turn to the next page until the exam has started.

Write your name and section number at the top right of this page:

Class	**Section Number**
Bret 8:50	001
Sahifa 1:20	003
Miranda 9:55	004
Sahifa 3:30	005
Sahifa 8:50	006
Cameron 1:20	007

As you complete the exam, write your initials at the top right of each other page.

When the exam start time is called, you may turn the page and begin your exam.
If you need more room, there is a blank page at the end of the exam, or we can give you some scratch paper.

Some multiple choice questions are "Select ONE" while others are "Select ALL that apply". Pay attention to the question type and only mark one option if it says "Select ONE". Fill in the circles completely.

If you finish early, you can hand your exam to your instructor or TA and leave early.

Otherwise, stop writing and hand your exam to your instructor or TA when the exam stop time is called.

1. A dataframe `lions` has 32 rows, each representing a unique lion.

   It has two numeric columns: `age`, each lion's age in decimal years, and `proportion.black`, the percentage of that lion's nose which is black.

   Finally, it has a logical column `adult`, which is `TRUE` if `age` is greater than 3 and `FALSE` otherwise.

   Consider the following code which attempts to make a scatter plot of `age` on the X axis vs. `proportion.black` on the y axis.

````
```{r}
lions = read_csv("../../data/lions.csv") %>%
  mutate(adult = age > 3)
```

```{r, echo = TRUE, eval = FALSE}
ggplot(lions, aes(age, proportion.black)) +
  geom_point(aes(fill = adult, alpha = 0.5), size = 2)
```
````

Which of the following statements are true about errors in the above code? Select ALL that apply.

- ● `fill = adult` must be changed to `color = adult` to color the points differently.
- ● `alpha = 0.5` must be moved outside of `aes()`, like `size = 2` currently is.
- ○ `size = 2` must be moved within `aes()`, like `alpha = 0.5` is.
- ○ `fill = adult` must be moved to the `aes()` in the `ggplot()` call, not `geom_point()`.

2. Using the same dataframe as Question 1, consider the following code which creates a new dataframe, `lions_summary`.

```
```{r, echo = TRUE, eval = FALSE}
lions_summary = lions %>%
  group_by(adult) %>%
  summarize(averageAge = mean(age),
  averagePropBlack = mean(proportion.black))
```
```

Which of the following statements are true about `lions_summary`? Select ALL that apply.

- [x] `lions_summary` has 2 rows.
- [ ] `lions_summary` has 2 columns. (*False: adult, averageAge, and averagePropBlack.*)
- [x] `lions_summary` has a column called 'adult'.
- [ ] If there were any `NA` values in the `age` column of `lions`, all the values of `averageAge` in `lions_summary` will be `NA` (*Sneakily false: If there are only NA values in one category of adult, the other can still function!*)

3. A random viewer's ideal movie length, in minutes, is approximately $X \sim N(120, 15)$. Actual movie lengths are given by $Y \sim N(143, 19)$.

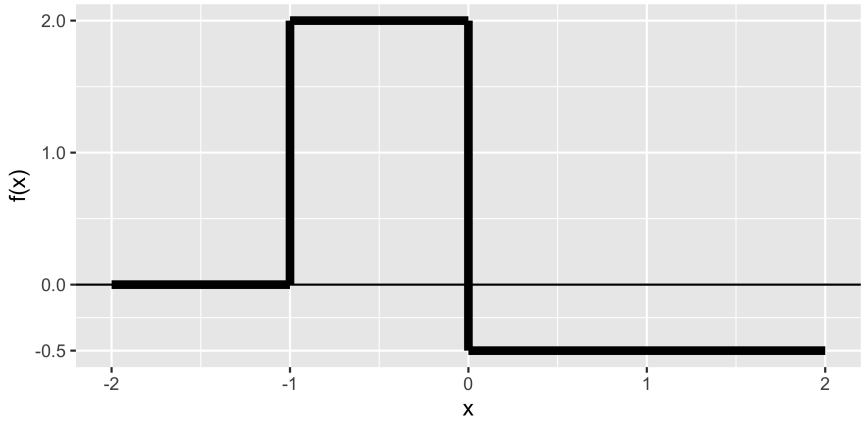
   (a) Which R code below calculates the probability that a random movie is shorter than the 40th percentile for ideal movie length? **Select ONE**.

   - [x] `qnorm(0.4, 120, 15) %>% pnorm(143, 19)`
   - [ ] `qnorm(0.4, 143, 19) %>% pnorm(120, 15)`
   - [ ] `pnorm(0.4, 120, 15) %>% qnorm(143, 19)`
   - [ ] `pnorm(0.4, 143, 19) %>% qnorm(120, 15)`

   (b) What movie length y corresponds to a z-score of 1? **Select ONE**.

   - [ ] 19
   - [ ] 124
   - [ ] 143
   - [x] 162 $(143 + 19 = 162)$

4. Your friend claims that a random variable, X, follows the following probability distribution:



Which of the following are true about your friend's claim? Select ONE.

- ○ This distribution has negative values of x, therefore X is not a random variable.
- ● This distribution has negative values of $P(X = x)$, therefore X is not a random variable.
- ○ The area under this distribution is not 1, therefore X is not a random variable.
- ○ X is a valid random variable.

5. Consider ordering a "footlong" (12 inch) sub from Subway.

   As your sub is prepared, it may not end up being exactly 12 inches.

   Consider trying to estimate μ, the average length of the sub you receive when you order a 12 inch sub.

   You do this by ordering 40 footlong subs and measuring their lengths in inches. (What a great problem this is.) You observe that the average length of your 40 subs is 11.5 inches, and the standard deviation of their lengths is 0.4 inches.

   Consider testing $H_0 : \mu = 12$ vs. $H_\alpha : \mu < 12$.

   Write a numeric expression (only containing numbers and artihmetic symbols like + and ×) for the test statistic of this test.

$$\frac{\bar{x} - \mu_{\text{null}}}{s_x/\sqrt{n}}$$

Answer:

$$\frac{\bar{x} - \mu_{\text{null}}}{s_x/\sqrt{n}} = \frac{11.5 - 12}{0.4/40}$$

6. The fictitious distribution of the number of children in a household (X) is given below.

x	0	1	2	3
$P(X = x)$	?	0.4	0.25	0.15

(a) Which of the following statements about X is true? **Select ONE**.

- ● X is a discrete RV.
- ○ X is a binomial RV.
- ○ X is a continuous RV.
- ○ X is a normal RV.

(b) What value of $P(X = 0)$ makes this a valid probability distribution? **Select ONE**.

- ○ 0.1
- ● 0.15
- ○ 0.2
- ○ 0.25

(c) What is the median number of children per household? **Select ONE**.

- ○ 0
- ● 1
- ○ 2
- ○ Not enough infomration to determine

7. Data is collected on the duration each winter that a lake's surface is frozen over a period of 103 consecutive years.

The correlation coefficient between the variables is $r = -0.4$.

The year variable has a mean $\bar{x} = 1950$ and a standard deviation $s_x = 30$.

The freeze duration variable has $\bar{y} = 90$ and a standard deviation $s_y = 17$.

Consider fitting a simple linear regression model to this data.

Write a numerical expression (using only numbers and arithmetic symbols like + or ×) for the predicted freeze duration in the year 1890. No need to simplify.

Answer: (1890 is two x standard deviations BELOW the mean of x, so our predicted value will be r * two y standard deviations below the mean of y.)*

$$90 + (-0.4) \times (-2) \times 17$$

You could also take the long way:

$$\hat{\beta}_1 = -0.4 * \frac{17}{30} \text{ and } \hat{\beta}_0 = 90 - \hat{\beta}_1 * 1950, \text{ so } \hat{y_1} = \hat{\beta}_0 + \hat{\beta}_1 * 1890$$

8. We continue to consider the lake freezing data from Problem 7.

   A 95% confidence interval for the freeze duration of the true regression line at $x = 1890$ has the form:

   $$\hat{y}_1 \pm c \times SE_C$$

   A 95% prediction interval for the duration that the lake was frozen in the single year 1890 has the form

   $$\hat{y}_1 \pm c \times SE_P$$

   where c is a critical value from some distribution and SE_C, SE_P are some positive values. *Reminder: There are 103 years in the dataset.*

   Which R code calculates the numeric critical value c? Select ONE.

   - ○ qnorm(0.95)
   - ○ qnorm(0.975)
   - ○ qt(0.95, 101)
   - ● qt(0.975, 101)
   - ○ qt(0.95, 102)
   - ○ qt(0.975, 102)

9. Using the expressions from the previous problem, circle the true relationship between SE_C and SE_P and briefly explain why.

   $SE_C < SE_P$ $SE_C = SE_P$ $SE_C > SE_P$

   Answer: A confidence interval for the height of the regression line is narrower than a prediction interval for at the same x point; but they have the same center and critical value. Therefore, it must be that $SE_C < SE_P$.

10. Consider trying to estimate the proportion of UW-Madison undergraduate students who will graduate at the end of this semester, p.

    You ask 50 random students, and 7 of them will graduate at the end of the semester.

    Consider using this information to calculate a confidence interval for p.

    Which of the following statements are true? Select ALL that apply.

    - ○ If we were to calculate an Agresti-Coull confidence interval for p, its center would be $(7+1)/(50+2)$. *(False: It would be $(7+2)/(50+4)$.)*
    - ● If we were to calculate a Wald confidence interval p, its center would be $7/50$.
    - ● The upper bound of the Agresti-Coull confidence interval would be greater than the upper bound of the Wald confidence interval. *(it's shifted towards 0.5)*
    - ● As we increase the confidence level of our interval towards 100%, our interval would widen and approach [0, 1].

11. We are interested in comparing the proportions of individuals with bachelor degrees among of adult men aged 25 and older in the states of Minnesota and Wisconsin. Let p_M and p_W represent these two population proportions. In random samples of $n_M = 300$ Minnesota men and $n_W = 400$ Wisconsin men aged 25 and older, the numbers of individuals with bachelor degrees are $x_M = 90$ and $x_W = 100$.

    Without simplification, write a numerical expression using provided data for the **standard error** SE_{ci} used in a 95% confidence interval for $p_M - p_W$ of the form

    $$(\text{point estimate}) \pm z_{\text{crit}} \times \text{SE}_{\text{ci}}$$

    when using the Agresti-Coull method.

    Answer:

    $$\text{SE}_{\text{ci}} = \sqrt{\frac{\frac{90+1}{300+2} * (1 - \frac{90+1}{300+2})}{300+2} + \frac{\frac{100+1}{400+2} * (1 - \frac{100+1}{400+2})}{400+2}}$$

12. In the setting of the problem above, write a numerical expression using provided data for the **standard error $\mathbf{SE_{ht}}$ of the test statistic** Z:

$$Z = \frac{\hat{p}_M - \hat{p}_W}{\text{SE}_{\text{ht}}} \sim N(0,1)$$

in a hypothesis test for the equality of the proportions of individuals with college degrees in the two states populations.

Answer:

$$\text{SE}_{\text{ht}} = \sqrt{\frac{\frac{90+100}{300+400} * (1 - \frac{90+100}{300+400})}{300} + \frac{\frac{90+100}{300+400} * (1 - \frac{90+100}{300+400})}{400}}$$

13. Suppose that the test statistic from the previous problem has the value $z = 1.47$ and that the p-value from a two-sided test is 0.14.

Which of the following statements are true? Select ALL that apply.

- [x] In the sample data, the proportion of Minnesota men with a college education is larger than the proportion of Wisconsin men with college education. (*Test stat is positive so $pM - pW$ is positive so $pM > pW$.*)
- [ ] The test is statistically significant at an $\alpha = 0.05$ level.
- [ ] There is proof beyond a reasonable doubt that Minnesota has a higher proportion individuals with college degrees among men aged 25 and older than Wisconsin does.
- [x] There is insufficient evidence to conclude that there is a difference in the proportions of individuals with college degrees among men aged 25 and older between Minnesota and Wisconsin, using a $\alpha = 0.05$ statistical significance level.
- [x] A 95% confidence interval for $p_M - p_W$ will contain the value 0. (*insignificant p-value*)

14. You ask a group of 30 people from Country A to each privately give you a random number between 1 and 1000.

    You then ask a group of 40 people from Country B to each privately give you a random number between 1 and 1000.

    Let μ_A be the true average value that all members of country A would give upon this request, and similarly for μ_B. You are interested in the quantity $\mu_A - \mu_B$.

    Which of the following statements are true about how to approach this problem through statistical inference? Select ONE.

    ○ Two-sample means inference is impossible to conduct because the sample sizes are different.

    ○ Two-sample means inference is impossible to conduct because the random variables are technically discrete.

    ● Two-sample inference is more appropriate than paired-sample inference in this case.

    ○ Paired-sample inference is more appropriate than two-sample inference in this case.

15. Assume the correct value of the test statistic from problem 14 is -50. **Which R code correctly calculates the p-value for this test? Select ONE.**

    ● `pt(-50, df = 40-1)`

    ○ `2*pt(-50, df = 40-1)`

    ○ `2*pt(abs(-50), df = 40-1)`

    ○ `1 - pt(-50, df = 40-1)`