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.

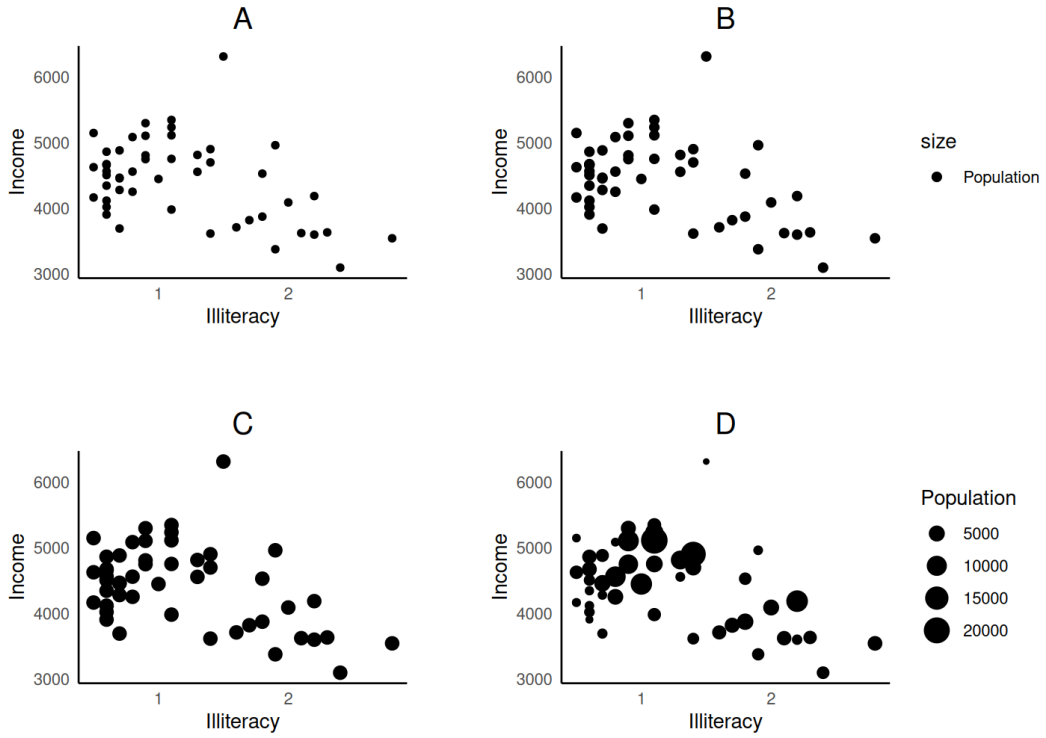
When the exam start time is called, you may turn the page and begin your exam.
If you need more room, we can give you some scratch paper.

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. (4 points) Below are four different `geom_point()` plots of illiteracy rate versus income for US states. Some plots also show the states' population.



Match the four plots (A, B, C, D) to the four different `geom_point()` calls below.

A `geom_point(aes(x = Illiteracy, y = Income))`

D `geom_point(aes(x = Illiteracy, y = Income, size = Population))`

B `geom_point(aes(x = Illiteracy, y = Income, size = "Population"))`

C `geom_point(aes(x = Illiteracy, y = Income), size = 3)`

2. (4 points) A data set `bm` has Boston Marathon data from the year 2010 with a row for each female runner who completed the race and variables `Age`, `Age_Range`, and `Time`. Identify code that calculates the mean time for all female runners between the ages of 35 and 39 (`Age_Range = "35-39"`). **Select ALL that apply**.

If an answer outputs other summaries in addition to the desiered value, it is still considered correct.

- [x] `bm %>% filter(Age_Range == "35-39") %>% summarize(mean = mean(Time))`
- [x] `bm %>% group_by(Age_Range) %>% summarize(mean = mean(Time))`
- [ ] `bm %>% mutate(mean = mean(Time)) %>% filter(between(Age, 35, 39))`
- [ ] `bm %>% select(Age_Range == "35-39") %>% summarize(mean = mean(Time))`

3. The plot above shows the densities of three different distributions, named Solid, Dashed, and Dotted. One of these distributions is $N(0,1)$, standard normal, another one is $t(5)$, the t distribution with 5 degrees of freedom, and the final one is $D(0,1)$, some other distribution with mean 0 and standard deviation 1.

   (a) (3 points) Identify each distribution by the corresponding line type.

   - $N(0,1)$: ● Solid ○ Dashed ○ Dotted
   - $t(5)$: ○ Solid ○ Dashed ● Dotted
   - $D(0,1)$: ○ Solid ● Dashed ○ Dotted

   (b) (3 points) Rank the three distributions using their line type names (Solid, Dashed, Dotted) in order from that with the smallest to the largest values of the 0.975 quantile.

   (Smallest) Dashed $<$ Solid $<$ Dotted (Largest)

   (c) (3 points) A random sample of size $n = 400$ is drawn from distribution $D(0,1)$ and we calculate the sample mean

   $$\bar{X} = \frac{1}{400}\sum_{i=1}^{400} X_i.$$

   What are the mean and standard deviation of the distribution of $\bar{X}$? **Select ONE**.

   ○ $\mu_{\bar{X}} = 0,\ \sigma_{\bar{X}} = 1$
   ○ $\mu_{\bar{X}} = 0,\ \sigma_{\bar{X}} = \frac{1}{400}$
   ○ $\mu_{\bar{X}} = 400,\ \sigma_{\bar{X}} = 400$
   ● $\mu_{\bar{X}} = 0,\ \sigma_{\bar{X}} = \frac{1}{\sqrt{400}}$

   (d) (3 points) Which of the three distributions will have a shape closest to that of the distribution of $\bar{X}$? (Note that the scale will be different.) **Select ONE**.

   ● $N(0,1)$ ○ $t(5)$ ○ $D(0,1)$

4. Define four different independent binomial random variables.

$X_1 \sim \text{Binom}(5, 0.6)$, $X_2 \sim \text{Binom}(25, 0.4)$, $X_3 \sim \text{Binom}(100, 0.1)$, $X_4 \sim \text{Binom}(100, 0.4)$

(a) (4 points) Fill in the blanks below with "less than", "greater than", or "equal to".

- The expectation of X_1 is ____greater than____ 2.5.
- The smallest possible value of X_1 is ____equal to____ the smallest possible value of of X_4.
- The expectation of X_3 is ____less than____ the expectation of X_4.
- The standard deviation of X_3 is ____less than____ the standard deviation of X_4.

(b) (3 points) Which of the resulting combinations below is also a binomial random variable? **Select ONE**.

- ○ $Y_1 = X_1 + X_2$
- ○ $Y_2 = X_2 + X_3$
- ● $Y_3 = X_2 + X_4$
- ○ $Y_4 = X_3 + X_4$

(c) (3 points) Which of the four variables is *least* well approximated as $N\left(np, \sqrt{np(1-p)}\right)$? **Select ONE**.

- ● X_1
- ○ X_2
- ○ X_3
- ○ X_4

5. A regression line is fit on 50 (x, y) pairs. Summaries of the data and the regression line are given below.

	X	Y
Mean	26.8	80.1
Range	[10.5, 39.8]	[29.8, 123.4]

Regression line: $\hat{y}_i = 4.06 + 2.84x_i$

(a) (5 points) A 90% prediction interval for y at $x = 30$ is reported as (73.6, 104.68). Which of the following intervals are narrower than this interval? **Select ALL that apply**.

- ● An 80% prediction interval for y at $x = 30$.
- ● A 90% confidence interval for the regression line height at $x = 30$.
- ● A 90% prediction interval for y at $x = 28$.
- ● A 90% confidence interval for y at $x = 28$.

(b) (3 points) A 95% prediction interval for y at $x = 50$ is reported as (129.15, 162.25). Why would this result be considered extrapolation? **Select ONE**.

- ● The value 50 is outside of the range of the original X values.
- ○ The lower bound of the interval is outside the range of the original Y values.
- ○ The interval fails to cover the mean of the original Y values.
- ○ None of the above; this is not considered extrapolation.

6. `diamonds` is a dataset that comes with `tidyverse`. There are 53,940 observations in `diamonds`, each of which represents a single diamond. `price` represents the price of a diamond in US dollars, and `carat` represents the weight in of a diamond in carats.

   Consider the following output of `summary()` on a linear model created by `lm()` which regresses price (Y) on carat (X).

```
Call:
lm(formula = price ~ carat, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-18585.3   -804.8    -18.9    537.4  12731.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2256.36      13.06  -172.8   <2e-16 ***
carat        7756.43      14.07   551.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1549 on 53938 degrees of freedom
Multiple R-squared:  0.8493, Adjusted R-squared:  0.8493
F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

   (a) (4 points) Using the information from the output above, fill in the blanks to create a 90% confidence interval for the true slope, β_1, of this model. You do not need to write in the critical value.

   $$\underline{\quad 7756.43 \quad} \pm \text{critical value} \times \underline{\quad 14.07 \quad}$$

   from the `carat` line of the summary table

   (b) (3 points) Let `n` be the number of rows in the `diamonds` dataset. Which line R code calculates the correct critical value for the interval in (a)? **Select ONE**.

   - ○ `qnorm(0.95)`
   - ○ `qnorm(0.975)`
   - ● `qt(0.95, df = n-2)`
   - ○ `qt(0.975, df = n-2)`

7. From 1999 to 2008, the U.S. Mint released a huge quantity of special commemorative quarters (25 cent coins), each of which had a unique design on the back celebrating one of the fifty states. Let p be the true proportion of all the commemorative quarters that have a Wisconsin design on the back. Consider the following hypothesis test:

$$H_0 : p = 1/50 \text{ vs. } H_\alpha : p > 1/50$$

You randomly collect 200 of these commemorative quarters and find that 11 have Wisconsin designs on the back.

In conducting this test, you may assume that any relevant assumptions are met and do not have to mention nor explain them.

(a) (3 points) What is the observed value of the test statistic for this test?
The test statistic is the count of "successes", which are Wisconsin quarters. The observed test statistic in this case is 11.

(b) (5 points) Assuming the null hypothesis is true, what is the sampling distribution of the test statistic?

$$X \sim \underline{\text{Binom}} \left(\underline{200} , \underline{\frac{1}{50}} \right)$$

The parameter of interest is p from Binom(n, p). The null $n = 200$ since we have 200 total observations and the null $p = \frac{1}{50}$ from the null hypothesis.

(c) (3 points) Which line of R code correctly calculates the p-value for this test? **Select ONE**.

- ○ `pbinom(11, 200, 1/50)`
- ○ `pbinom(11-1, 200, 1/50)`
- ○ `1 - pbinom(11, 200, 1/50)`
- ● `1 - pbinom(11-1, 200, 1/50)`

(d) (4 points) Assume the p-value is 0.003. **Select ALL** correct statements or justifiable conclusions

- ● The result is statistically significant at the $\alpha = 0.05$ level.
- ● There is strong evidence that the proportion of special commemorative state coins with Wisconsin on the back is greater than 1/50.
- ○ Based on the same data, a 95% confidence interval for the true proportion of Wisconsin quarters would contain the value 1/50.
- ○ The p-value from a two-sided test with the same data would also be equal to 0.003.

8. A 95% confidence interval for the difference in population proportions $p_1 - p_2$ is

$$-0.1 < p_1 - p_2 < 0.2.$$

(a) (3 points) What is true about the p-value of a two-sided hypothesis test of $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$? **Select ONE**.

- ○ The p-value is less than 0.05.
- ○ THe p-value is equal to 0.05.
- ● The p-value is greater than 0.05.

(b) (3 points) The interval was created using the Agresti-Coffe adjustment. If X_1, X_2 and n_1, n_2 are the number of successes and the number of trials from each group, what is the A-C adjusted point estimate for $p_1 - p_2$? **Select ONE**.

- ○ $\frac{X_1}{n_1} - \frac{X_2}{n_2}$
- ● $\frac{X_1 + 1}{n_1 + 2} - \frac{X_2 + 1}{n_2 + 2}$
- ○ $\frac{X_1 + 2}{n_1 + 4} - \frac{X_2 + 2}{n_2 + 4}$
- ○ $\frac{X_1 + X_2 + 2}{n_1 + n_2 + 4}$

9. Identify each of the following scenarios as having independent or paired data.

(a) (3 points) A teacher wants to compare the performance of 200 students on the practice SAT and actual SAT exam. Each student has their score recorded for both the practice and actual exam.

○ Independent ● Paired

(b) (3 points) A teacher wants to compare the performance of students in her 9:00 AM class with students in her 1:00 PM class. She selects 100 students from each class and records the SAT score for each one.

● Independent ○ Paired

10. (4 points) What lines of R code can be used to perform a paired T test on x and y? **Select ALL that apply**.

- ○ `t.test(x, y)`
- ● `t.test(x - y)`
- ● `t.test(x, y, paired = T)`
- ○ `t.test(x - y, paired = T)`

11. In a large bookstore, you are trying to determine if the average number of pages of a book in the sci-fi section is *different* from the average number of pages of a book in the mystery section.

- Let μ_S be the average page length of all books in the sci-fi section.
- Let μ_M be the average page length of all books in the mystery section.

(a) (3 points) State the appropriate null and alternative hypotheses for this test.
We want to know if the average is different, which implies a two-sided test for a difference in means.

$$H_0 : \mu_S - \mu_M = 0 \quad \text{versus} \quad H_A : \mu_S - \mu_M \neq 0$$

(b) (4 points) You take a random sample of 30 sci-fi books and find the average page length to be 236.4 with standard deviation 23.8. You take a random sample of 40 mystery books and find the average page length to be 231.3 with standard deviation 52.6.

Write a numerical expression that calculates the test statistic for your hypotheses. Do not simplify, evaluate, or round.
We are using the Welch T test statistic for a difference in means.

$$t_{obs} = \frac{\bar{x}_S - \bar{x}_M - 0}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_M^2}{n_M}}} = \frac{236.4 - 231.3 - 0}{\sqrt{\frac{23.8^2}{30} + \frac{52.6^2}{40}}}$$

(c) (4 points) Consider the following output of `t.test()` on this data. **Select ALL** true statements or justifiable conclusions.

```
        Welch Two Sample t-test

data:  scifi and mystery
t = 0.54033, df = 57.381, p-value = 0.5911
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.72092  23.86401
sample estimates:
mean of x mean of y
 236.3856  231.3140
```

- ○ There is strong evidence that the population mean length of sci fi books is larger than that of mystery books.
- ● The sample mean length of sci fi books is greater than the sample mean of mystery books.
- ○ The two-sided hypothesis test is significant at the $\alpha = 0.05$ level.
- ● There is a lack of evidence that the population mean lengths of either book genre differ from each other.