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**
Miranda 8:50	001
Sahifa 12:05	002
Miranda 9:55	004

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. The fictitious dataset `state_employees` contains columns for the ID, department, number of years of service, and annual salary for 100000 state employees.

```
state_employees %>%
  sample_n(5)
```

	employee_id	department	years_service	salary
	<int>	<chr>	<dbl>	<dbl>
1	10783	agriculture	3	23660
2	69923	public_safety	7	33917
3	35260	corrections	8	28847
4	3336	agriculture	4	25772
5	45582	social_services	5	26632

The graph below shows the average annual salary for the five departments in the dataset.

(a) (3 points) In the plot above, which aesthetics are variable aesthetics? **Select ALL that apply**.

- [x] `x`
- [x] `y`
- [ ] `color`
- [ ] `fill`
- [x] `text`
- [ ] `line`

(b) (3 points) Was the plot above made using `geom_bar()` or `geom_col()`? Briefly explain how you can tell.

This plot must have been made with `geom_col()` after calculating the average annual salary since we need to provide both an x and a y to specify the height of the bars. `geom_bar()` would give a count of each department which is not what we want in this case.

2. Continue working with the employee salary data from problem 1.

```
state_employees %>%
  sample_n(5)
```

	employee_id	department	years_service	salary
	<int>	<chr>	<dbl>	<dbl>
1	10783	agriculture	3	23660
2	69923	public_safety	7	33917
3	35260	corrections	8	28847
4	3336	agriculture	4	25772
5	45582	social_services	5	26632

(a) (4 points) Consider the following code which creates a new dataframe, `employee_summary`.

```
employee_summary <- state_employees %>%
 filter(department != "agriculture") %>%
 group_by(department) %>%
 summarize(avg_salary = mean(salary), employees = n())
```

Which of the following statements are TRUE about `employee_summary`? **Select ALL that apply**.

- [x] `employee_summary` has 4 rows.
- [ ] `employee_summary` has 5 rows.
- [x] `employee_summary` has a column called "`department`".
- [ ] If were any `NA` values in the `salary` column of the original data, all the values of `avg_salary` in `employee_summary` will be `NA`.

(b) (4 points) Continue working with the 100000 rows of employee salary data. Another dataset called `employment_type` has 74033 rows and two columns: numeric `employee_id` and categorical `type` which labels each employee as either full-time or part-time. The 74033 employee IDs are all also part of the larger `state_employees` dataset.

How many rows and columns result from following join? Which columns will contain `NA` values?

```
left_join(state_employees, employment_type)
```

The joined data will have 100000 rows, just like the original `state_employees` data. There will be 5 columns: the four original columns from `state_employees` and the column for employment type from the `employment_type` dataframe.

This `type` column will have `NA` values for the rows that appear in `state_employees` but not in `employment_type`.

3. The distribution of the number of children in a household (X) in a certain neighborhood is given below.

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

(a) (3 points) 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) (3 points) What is the median number of children per household? **Select ONE**.

- ○ 0
- ● 1
- ○ 2
- ○ Cannot be determined from the information given

4. You are interested in comparing the performance of Student A at University A with Student B at University B.

Grade point averages at University A are normally distributed with mean 2.8 and standard deviation 0.4. Grade point averages at University B are normally distributed with mean 2.5 and standard deviation 0.5.

(a) (4 points) Student A has a GPA of 3.2 at University A. Which R code below calculates the GPA of Student B who has the same percentile GPA at University B? **Select ONE**.

- ● `pnorm(3.2, 2.8, 0.4) %>% qnorm(2.5, 0.5)`
- ○ `pnorm(3.2, 2.5, 0.5) %>% qnorm(2.8, 0.4)`
- ○ `qnorm(3.2, 2.8, 0.4) %>% pnorm(2.5, 0.5)`
- ○ `qnorm(3.2, 2.5, 0.5) %>% pnorm(2.8, 0.4)`

(b) (3 points) Consider taking a random sample of five students from University A. What is the distribution of $\bar{X}$, the average of their GPA's? **Select ONE**.

- ○ $\bar{X} \sim N(2.8, 0.4)$
- ○ $\bar{X} \sim N(2.8, 0.4\sqrt{5})$
- ● $\bar{X} \sim N(2.8, \frac{0.4}{\sqrt{5}})$
- ○ Cannot be determined from the information given

5. 74 individuals were asked to give their annual income and rate their life satisfaction (out of 100). A linear model is fit on life satisfaction (y) in terms of income (x).

   The predicted satisfaction for an annual income of $x^* = \$70,000$ is $\hat{y} = 61.2$ and a 95% prediction interval at $x^* = \$70,000$ is given by

   $$(38.4, 84.0)$$

   (a) (4 points) Which of these changes will result in a narrower interval? **Select ALL that apply**.

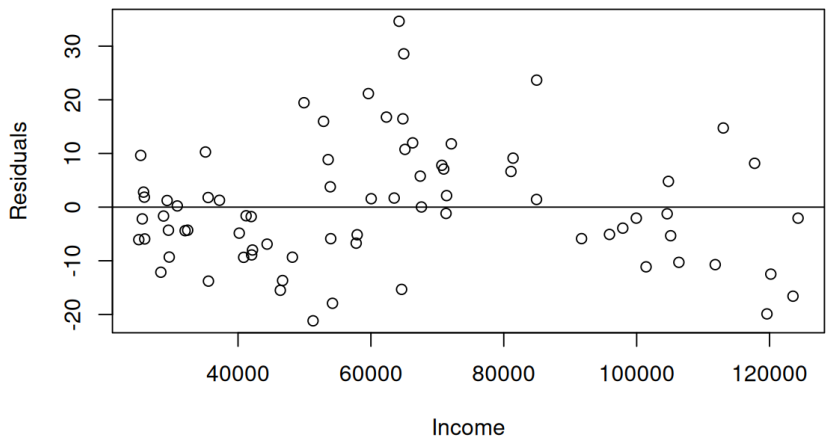
   - [x] Using a confidence level of 90%.
   - [ ] Using a confidence level of 98%.
   - [x] Building a confidence interval instead of a prediction interval.
   - [x] Building a prediction interval at $\bar{x} = 62900$ instead of $x = 70000$.

   (b) (4 points) The y-intercept of the linear model is 22.8. Write a numerical expression for the slope of the linear model. You do not need to solve or simplify.

   We know that the linear model passes through (70000, 61.2) and also passess through (0, 22.8). So, the slope of the model is given by

   $$\hat{\beta}_1 = \frac{61.2 - 22.8}{70000 - 0}$$

   (c) (3 points) A scatterplot of income (x) versus the model residuals is given below.



   Based on the residual plot, which of the linear modeling assumptions could be violated for this data? **Select ONE**.

   - [x] X and Y have a linear relationship.
   - [ ] The residuals have mean 0.
   - [ ] The residuals have constant variance.
   - [ ] None of the above.

6. The `trees` has 31 observations with the Girth, Volume and Height of cherry trees. Below is the coefficient table of a linear regression model for volume vs girth on the `trees` data fit with `lm()`.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use the R functions `pt(q, df, lower.tail)` and `qt(p, df, lower.tail)` to answer parts (a) and (b) below.

(a) (3 points) Build a 90% confidence interval for the slope β_1. Use one of the R functions above to write the expression for the critical value.

The estimated slope is 5.0659 with standard error 0.2474. We need to use the 95th percentile on the $T(n-2)$ distribution as our critical value. So, the CI is given by

$$5.0659 \pm \texttt{qt(0.95, df = 31-2)} \times 0.2474$$

(b) (4 points) Based on the output of the table, identify the test statistic for testing $H_A : \beta_1 < 0$. Use one of the R functions above to write the expression for the correct p-value for this test.

The test statistic is 20.48, the same as the test statistic given in the output for testing $\beta_1 \neq 0$. The p-value for our one-sided test is the area below our test statistic on the null $T(n-2)$ distribution. This is

```
pt(20.48, df = 31-2)
```

(c) (3 points) The p-value in (b) is approximately equal to: **Select ONE**.

- ○ $< 2 \times 10^{-16}$
- ○ $< 1 \times 10^{-16}$
- ○ 0.05
- ● 1

7. Consider trying to estimate the proportion of UW-Madison students that will graduate at the end of the semester, p. You ask 50 random students, and 7 of them will graduate at the end of the semester.

   Consider using this information to construct a confidence interval or perform a hypothesis test for p.

   (a) (4 points) 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).
   - ● If we were to calculate a Wald confidence interval for p, its center would be 7/50.
   - ● The upper bound of the A-C confidence interval would be greater than the upper bound of the Wald confidence interval.
   - ○ As we increase the confidence level toward 100%, our interval would wide an approach [0, 100].

   (b) (3 points) Consider performing a hypothesis test of whether **greater than 15%** of students will graduate at the end of the semester. Write appropriate hypotheses and identify the null distribution for this test.

   The hypotheses are

   $$H_0 : p = 0.15 \quad \text{versus} \quad H_A : p > 0.15$$

   where p is the true proportion of students graduating at the end of the semester.

   We can do this test with either a binomial or a normal null. So, the null would be $Binom(50, 0.15)$ or $N(0, 1)$.

8. (4 points) We are interested in comparing the proportion of individuals with bachelor's degrees among adult men 25 and older in 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 25 and older, the number of individuals with bachelor's degrees is $x_M = 90$ and $x_W = 100$.

   Without simplification, write a numerical expression using the provided data for the standard error of the test statistic Z in a hypothesis test for the equality of proportions p_M and p_W.

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

   We need to calculate a pooled proportion since we are testing for an equality of proportions. This is $\hat{p} = \frac{90+100}{300+400} = \frac{190}{700}$. The standard error for the difference in proportions is

   $$\hat{se}(\hat{p}_M - \hat{p}_W) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_m} + \frac{1}{n_W}\right)} = \sqrt{\frac{190}{700}\left(1 - \frac{190}{700}\right)\left(\frac{1}{300} - \frac{1}{400}\right)}$$

9. The output of R `t.test()` on two sets of data, `shopOne` and `shopTwo`, is given below.

```
        Welch Two Sample t-test

data:  shopOne and shopTwo
t = 0.019454, df = 91.984, p-value = 0.4923
alternative hypothesis: true difference in means is greater than -10
99 percent confidence interval:
 -12.20095       Inf
sample estimates:
mean of x mean of y
 139.7905  149.7723
```

(a) (4 points) Write a numerical expression for the standard error for the difference of means of Shop One and Shop Two. You do not need to solve or simplify.

From the test statistic formula, we have

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\text{se}}$$

$$0.019454 = \frac{139.7905 - 149.7723 + 10}{\text{se}}$$

$$se = \frac{139.7905 - 149.7723 + 10}{0.019454}$$

(b) (3 points) How can we write the pair of hypotheses is being tested in the `t.test()` output above? **Select ALL that apply**.

- ○ $H_0 : \mu_X - \mu_Y \leq -10$ and $H_\alpha : \mu_X - \mu_Y \geq -10$
- ● $H_0 : \mu_X - \mu_Y \leq -10$ and $H_\alpha : \mu_X - \mu_Y > -10$
- ○ $H_0 : \mu_X - \mu_Y \neq -10$ and $H_\alpha : \mu_X - \mu_Y = -10$
- ● $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y > -10$
- ○ $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y \neq -10$
- ○ $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y \geq -10$

(c) (3 points) Draw the correct statistical conclusion for the above hypothesis test. You answer should reference the `t.test()` output.

We have a large p-value, which means we have a non-significant result. We do not have evidence that the difference $\mu_X - \mu_Y$ is greater than -10.