Practice Final

Name: ____

You may use a calculator and three 8.5” by 11” sheets of notes (front and back), which you **will** turn in with your exam. This means you have to go to an **open book** room to take the test. However, you may **not** use the text book. You will have to bring your own calculator.

Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Cell phones and any other electronic devices (aside from your calculator) are not permitted. No interaction of any sort is allowed with your classmates.

I have tried to be very clear in my statements of all questions on this exam. If there are any questions where it is not clear what I am asking, please write down your best guess at what I am asking and answer that.

I Conceptual Questions

Please answer the following in no more than 1-2 sentences each.

1. (4 points) Comment on the following statement attributed to statistician George Box:

   > All models are wrong but some are useful.

   How does this statement relate to what we have learned about linear regression?

2. (4 points) What does it mean for data to be **paired**?

3. (4 points) What is a sampling distribution?

4. (6 points) Suppose I conduct a hypothesis test about the average amount of tea in a Twinings tea bag. The null hypothesis is that the population mean amount of tea is (less than or) equal to 2.5 grams, and the alternative hypothesis is that the population mean amount of tea is larger than 2.5 grams. In this context, what would a Type I error be? If I change the significance level of the test, α, from 0.05 to 0.01, how does that affect the probability of making a Type I error in this test?

5. (3 points) A survey of 500 households concluded that 82% of the population uses coupons at the grocery store. Describe what is meant by the poll having a margin of error of 3%.

II Applied Problems

1. (18 points) We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

   (a) (2 points) Describe the population parameter of interest.

   (b) (4 points) Check if the conditions for constructing a confidence interval and conducting a hypothesis test based on these data are met.

   (c) (4 points) What is a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university? Interpret the interval **in the context of this problem**.

(d) (4 points) According to the National Center for Education Statistics, the proportion of all 20-to-24 year olds with college degrees who were employed as of 2015 is 0.89. Write down a statement of the null and alternative hypotheses for a test that the employment rate for the current graduating class is different from the national average among 20 to 24 year olds.

(e) (4 points) I used R to compute a p-value for this test, and I got a p-value of 0.2. **In the context of this problem**, what is your conclusion? Use a significance level of $\alpha = 0.05$.

2. (16 points) Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 micrograms/liter and a SD of 37.74 micrograms/liter; a previous study of a large number of individuals (not all police officers) from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 micrograms/liter.

   (a) (3 points) Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a higher concentration of lead than their neighbors in the suburbs. You may treat the value of 35 micrograms/liter from the suburban study as a fixed, known constant (i.e., we are not structuring this as a test to compare the means of two groups, but rather as a test about the mean value for the group of urban police officers).

   (b) (4 points) Explicitly state and check all conditions necessary for inference on these data. If you don't have enough information, say what you would want to know. If you would want to look at any plots, describe what plot(s) you would want to make and what you'd be looking for.

(c) (4 points) Calculate a p-value for the test that the downtown police officers have a higher lead exposure than the group in the previous study. **For full credit, you must show all of your work!** Points are assigned to correct set-up. In doing this calculation, you may use the following facts:

- If $T \sim t_{51}$, then $P(T > 23.754) < 10^{-16}$
- If $T \sim t_{51}$, then $P(T > 17.067) < 10^{-16}$
- If $T \sim t_{51}$, then $P(T > 3.294) = 0.001$
- If $T \sim t_{51}$, then $P(T > 2.367) = 0.011$

(d) (3 points) Interpret your results in context.

(e) (2 points) Suppose that you rejected the null hypothesis. Would this prove that there was a causal relationship between exposure to car exhaust and increased concentrations of lead in the blood?

3. (19 points total) An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. They recruited 50 people and randomly assigned them to two groups: 68 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.

		Group		
		Treatment	Control	Total
Result	Yawn	10	4	14
	Not Yawn	24	12	36
	Total	34	16	50

(a) (2 points) The MythBusters did not conduct a formal statistical analysis of these experimental results. Reproduce their analysis here by calculating the proportion of subjects in the treatment group who yawned, and the proportion of subjects in the control group who yawned.

(b) (2 points) Which group had a higher proportion of subjects who yawned? Could this result have occurred by chance even if on average the two groups had the same chances of yawning?

(c) (3 points) Check the necessary assumptions for obtaining confidence intervals and conducting hypothesis tests with these data.

You may use the following R output in answering the questions on the following page (note that the only difference between the following commands is the specification of the alternative hypothesis).

```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "two.sided")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.1
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.08266  0.46502
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "greater")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.05
## alternative hypothesis: greater
## 99 percent confidence interval:
##  -0.05837  1.00000
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

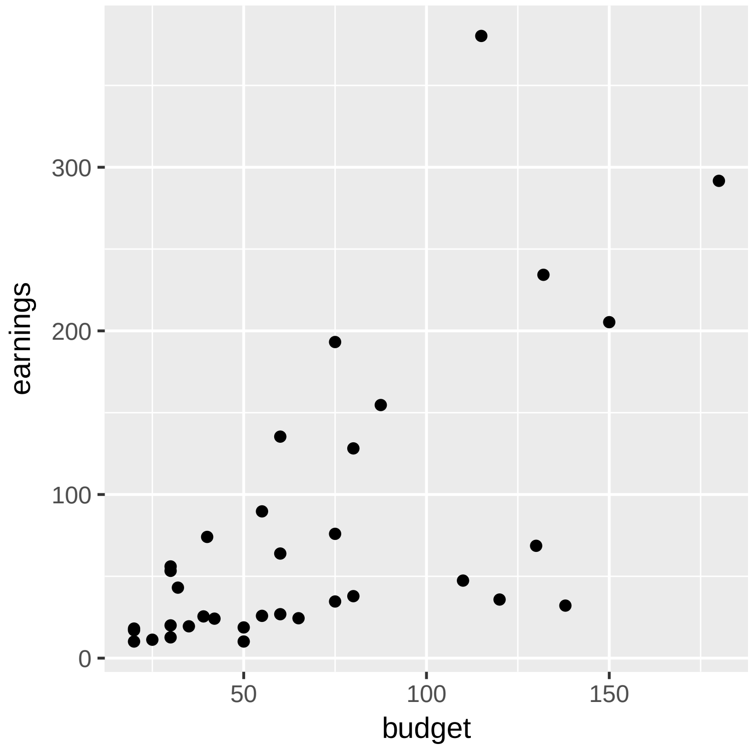
```
prop.test(x = c(30, 8), n = c(68, 32), conf.level = 0.99, alternative = "less")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(30, 8) out of c(68, 32)
## X-squared = 2.6, df = 1, p-value = 0.9
## alternative hypothesis: less
## 99 percent confidence interval:
##  -1.0000  0.4407
## sample estimates:
## prop 1 prop 2
## 0.4412 0.2500
```

(d) (8 points) Regardless of your answer to part (c), let's go ahead with inference. State and interpret a 99% confidence interval that is relevant to answering the question of whether a person can be subconsciously influenced into yawning if another person near them yawns. As part of your answer, discuss what it means to be "99% confident".

(e) (8 points) Suppose you'd like conduct a hypothesis test to see if yawning is contagious. Carry out a suitable hypothesis test at an $\alpha = 0.05$ significance level. Clearly state your null and alternative hypotheses, defining all population parameters that are involved. State your conclusions in context. Do you agree with the MythBusters' conclusion that these data provide strong evidence that people are more likely to yawn if someone near they yawns?

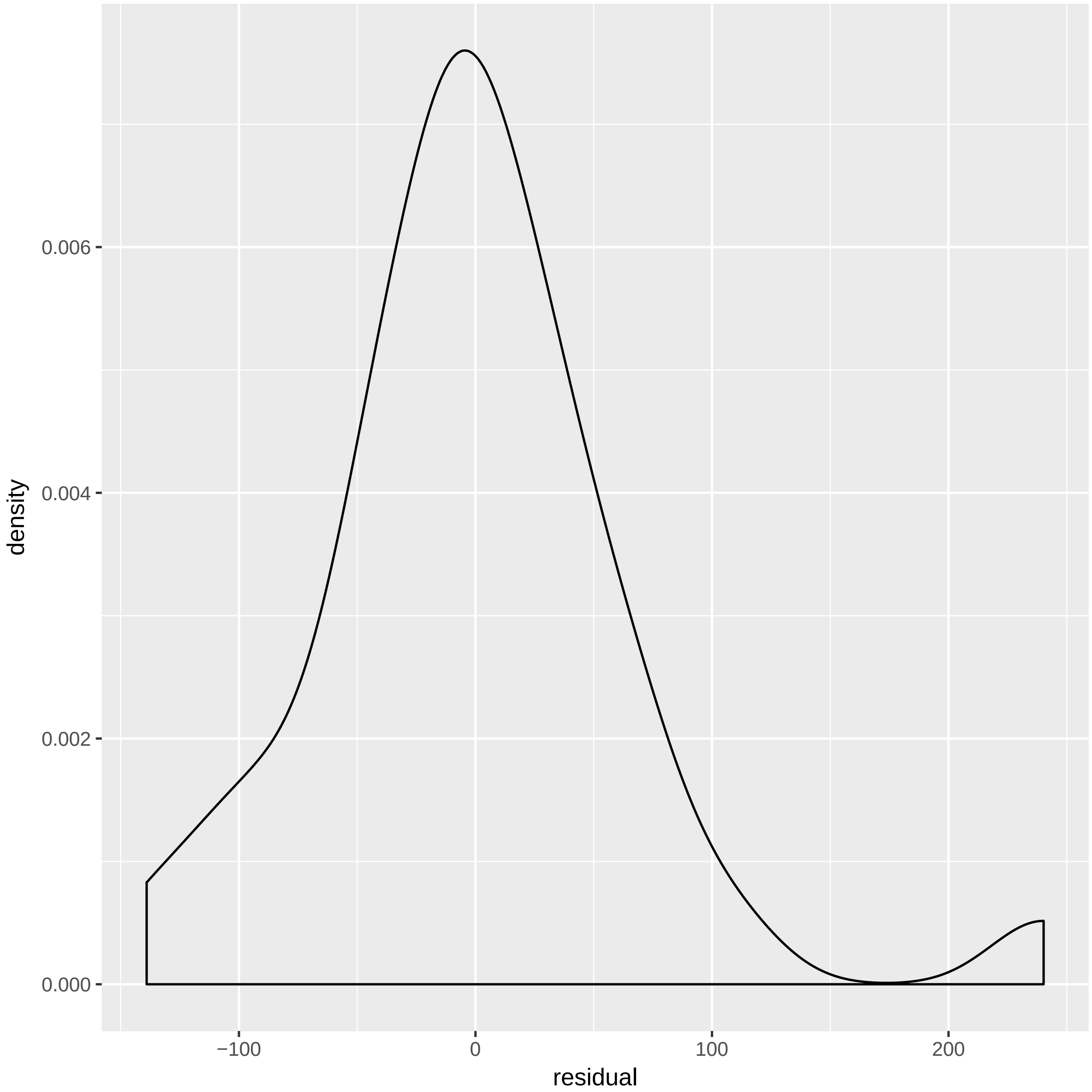
4. (30 points) What is the relationship between a movie's budget and its earnings? Let's look at data for 35 recent Action and Adventure movies.

```
## A first look
head(movies)

##                      title budget earnings     genre
## 1                  Elektra     65    24.41    Action
## 2 Assault on Precinct 13     30    20.04    Action
## 3 Pooh's Heffalump Movie     20    18.08 Adventure
## 4              Constantine     75    75.98    Action
## 5                  Hostage     75    34.64    Action
## 6                   Robots     80   128.20 Adventure

nrow(movies)

## [1] 35

ggplot() +
  geom_point(mapping = aes(x = budget, y = earnings), data = movies)
```



```
## A model fit
earnings_model <- lm(earnings ~ budget, data = movies)
summary(earnings_model)

##
## Call:
## lm(formula = earnings ~ budget, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.02  -36.17   -5.18   30.78  240.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.314     22.273   -0.69      0.5
## budget         1.351      0.278    4.85  2.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.2 on 33 degrees of freedom
## Multiple R-squared:  0.416,Adjusted R-squared:  0.399
## F-statistic: 23.5 on 1 and 33 DF,  p-value: 2.85e-05
```

```
## A plot of residuals
movies <- mutate(movies, residual = residuals(earnings_model))
ggplot() +
  geom_density(mapping = aes(x = residual), data = movies)
```



(a) Check the assumptions for the linear model. Extra Credit: If any assumptions are violated, suggest something you could do to address those limitations.

Regardless of your answer to part (a), let's proceed using the results from this model.

(b) (2 points) Write down the population model equation.

(c) (1 point) Write down the model equation again, filling in the estimated coefficients from the R output.

(d) (2 points) Interpret the model's slope coefficient in context.

(e) (2 points) The budget for "Wallace Gromit: The Curse of the Were-Rabbit" (classified as an Adventure movie) was 30 million dollars. Based on this model, what is the predicted earnings for this movie?

(f) (2 points) The actual earnings for "Wallace Gromit: The Curse of the Were-Rabbit" was 56 million dollars. What is the residual for this movie?

(g) (2 points) Give the value of the residual standard deviation and its interpretation, using the 95 part of the 68-95-99.7 rule.

(h) (2 points) Give the value of R^2 for this model and its interpretation.

(i) (6 points) Is there a statistically significant linear association between movie budget and earnings? State a null and alternative hypothesis in terms of the coefficients you set up in part (b), write down the p-value from the R output, and draw a conclusion in the context of the problem.

(j) (3 points) Draw a picture of the sampling distribution that's relevant for the test you conducted in part (i), locate the value corresponding to the test statistic, and shade in the area corresponding to the p-value for the test.

(k) (6 points) Write down a 95% confidence interval for the coefficient in the model describing the relationship between a movie’s budget and its earnings. Interpret the interval in context, including a discussion of what you mean by ”95% confident”. You will need to use one of the numbers in the following R output to construct your interval:

```
qt(0.95, df = 34)

## [1] 1.691

qt(0.95, df = 33)

## [1] 1.692

qt(0.975, df = 34)

## [1] 2.032

qt(0.975, df = 33)

## [1] 2.035
```