Statistical Inference Critical Thinking Application

PLEASE READ CAREFULLY: Introduction

A company wants to know how manufacturing temperature (in F) affects the lifespan of its lightbulbs. Every day, for 200 days, they take a random bulb, record its manufacturing temperature, plug it in, and let it run until it dies (months later, of course). Eventually, they end up with a sample of 200 lightbulbs, with their manufacturing temperature and how long they lived. A linear regression analysis has been conducted to determine if the temperature at which bulbs are manufactured affects their longevity. The objective is to critically analyze the provided regression outputs, discuss potential biases, and evaluate the precision of the estimate.

Regression Analysis

First, let's look at the regression analysis that has been performed. The analysis investigates how manufacturing temperature influences bulb lifespan.

Linear Regression Model

The theoretical model is given by:

$$\text{Lifespan} = \beta_0 + \beta_1(\text{Temperature}) + \epsilon$$

Where:

- β_0 is the intercept,
- β_1 is the regression coefficient for the manufacturing temperature, our **estimand**
- ϵ is the error term, which contains everything that affects lifespan that is not temperature.

It is our assumption that temperature has no relationship with lifespan, until we demonstrate otherwise. In fact,

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

In plain English:

H_0: There is no relationship between temperature and lifespan in the population.

H_1: There is a relationship between temperature and lifespan in the population.

EXERCISE 1: Explain in words, what our estimand is in this problem.

The estimand is the effect manufacturing temperature has on the lifespan of the company's lightbulbs.

EXERCISE 2: Explain in words, what our population is in this problem.

The population are all the lightbulbs the company manufactures.

Exercise 3: Explain in words, what our sample is in this problem.

The sample are the 200 lightbulbs selected at random.

The R code used to generate this model and its outputs is as follows:

```r
data <- read.csv("data.csv")
data <- data[,-1]
model <- lm(Lifespan ~ Temperature, data = data)

# Display the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Lifespan ~ Temperature, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -270.316  -67.286   -3.725   64.373  244.860
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2032.4898    44.0032  46.190   <2e-16 ***
## Temperature   -0.3854     0.1877  -2.053   0.0414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.82 on 198 degrees of freedom
## Multiple R-squared:  0.02084,    Adjusted R-squared:  0.0159
## F-statistic: 4.214 on 1 and 198 DF,  p-value: 0.0414
```

Exercise 4: Identify and interpret the value of our estimate (as it pertains to our sample), our "best guess" at the estimand.

The estimate of the effect of maufacturing temperature on the lifespan of the lightbulbs is the coefficient associated with the Temperature variable. The estimate suggests that a one degree increase, in farenheit, of the manufacturing temperature is associated with a decrease in the the lifespan of the lightbulb by 0.3854 hours.

Exercise 5: Do we have evidence that there is genuinely a relationship between temperature and lifespan in the population. How do we know?

\textcolor{red}{Yes, there is evidence that there is a genuine relationship between temperature and lifespan because of the standard error and p-value associated with the estimate. The standard error, theoretically, given an unbiased estimator measures the distance of the estimate from the estimand. Because our standard error is small we know that our estimate is close to the estimand. The p-value, is the probability we would observe the estimate of temperature in terms of distance from the null hypothesized estimand of 0. With a low standard error indicating distance from the estimand and a p-value of 4.14% we can conclude that our estimand is not equal to zero. Therefore, we reject the null hypothesis in favor of the alternative that the estimand is not equal to zero. Meaning there is a relationship between temperature and lifespan.}

Exercise 6: Discuss potential biases that could have affected the results.

A potential bias is we are not accounting for other factors that may have an affect of the lifespan of bulbs. We are only considering that temperature is the only factor that affects lifespan and not controlling for other factors. Such a factor could be voltage fluctuations or the humidity of the environment.

Exercise 7: Evaluate the precision of our estimator. Do we expect to get similar (negative) estimates if we performed the exact same sampling process, assuming no new biases were introduced.

The precision of the model is the ability to yield similar estimates using the same sampling techniques. Given the low standard error of the estimate we can conclude that the model is proucing precise estimates. And because the estimate is statistically significant we can cocnlude that the negative estimate is improbable given an estimand of 0. Therefore, because the estimand is below 0 and the model has a low standard error, we can exxpect to get similar negative estimates.

Exercise 8: Interpret the p-value of our estimate in literal terms, as I explained it in the notes from last class.

\textcolor{red}{Assume temperature had no effect on the lifespan of the lightbulbs. We then take a random sample of lightbulbs an infinite numbers of times to the point where we know the population distribution of the lightbulbs. We would observe an estimate of -0.3854 or larger, in terms of absolute value, 4.14% of the time.}

Please knit the PDF and submit one per group of 2-3