

DSCI354-351m-451 Final Exam (CWRU, Pitt, UCF, UTRGV)

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

19 December, 2022

Contents

0.1 Final Exam (worth 20 pts)

- Will be held Monday 12/19
 - From 12pm to 3pm, in Nord 356, or remote
- Comprehensive overview of the course

0.1.1 Academic Integrity Policy

All students in this course are expected to adhere to University standards of academic integrity.

Cheating, plagiarism, misrepresentation, and other forms of academic dishonesty will not be tolerated.

This includes, but is not limited to, consulting with another person during an exam, turning in written work that was prepared by someone other than you, making minor modifications to the work of someone else and turning it in as your own, or engaging in misrepresentation in seeking a postponement or extension.

- Ignorance will not be accepted as an excuse.
- If you are not sure whether something you plan to submit
 - would be considered either cheating or plagiarism,
 - it is your responsibility to ask for clarification.

For complete information, please go to [CWRU Academic Integrity Policy](#).

0.1.2 Final Exam Format

- The exam will appear in the prof repo
- In /assignments/exam-final folder
- Done as .Rmd file to turn in as .pdf report
- Submit Final Exam .Rmd, .pdf to the Canvas Assignment Page

0.1.3 Types of Questions

- 8 questions total
- OI Stats questions to do
- Data Wrangling: Tidying, EDA,
- 5 Paragraph Essay Question with cites: about Data Science
 - Citations to literature supporting your discussion
 - * These are done as footnotes
 - * Format: Author, Title, Source:Journal,Magazine, Page, Year, URL link
- Data Analysis: Modeling using Linear Regression

0.1.4 Points per question

- 1. OIS 1 pt
- 2. OIS 1 pt
- 3. OIS 1 pt
- 4. Tidy data wrangling 2 pt
- 5. EDA, Summary Stats & Visualization 3 pts
- 6. 5 paragraph Essay 4 pts
- 7. EDA on Real Dataset problem 4 pts
- 8. Linear Regression on a dataset 4 pts
 - Do an exploratory data analysis on Degradation of Transparent Conductive Oxides

You have a pdf of OIStats book in your readings folder of your Repo

- this is open book, open resource test

If the answer to a question part, like a), is in your code block,

- put # a) to show it in your code
-

1 1. Hypothesis Test: Car Insurance (1 pt)

1.0.1 OIStats v2 4.30

A car insurance company advertises that customers switching to their insurance save, on average, 432 on their yearly premiums.

A market researcher at a competing insurance discounter is interested in showing that this value is an overestimate so he can provide evidence to government regulators that the company is falsely advertising their prices.

He randomly samples 82 customers who recently switched to this insurance and finds an average savings of 395 USD, with a standard deviation of 102.

1.a) Are conditions for inference satisfied?

1.b) Perform a hypothesis test and state your conclusion.

1.c) Do you agree with the market researcher

- that the amount of savings advertised is an overestimate?
- Explain your reasoning.

1.d) Calculate a 90% confidence interval

- for the average amount of savings
 - of all customers who switch their insurance.

1.e) Do your results from the hypothesis test

- and the confidence interval agree?
- Explain.

1.1 ANSWER 1.a)

Size = 82 > 30, meaning Central Limit Theorem can be applied, therefore conditions for inference are satisfied

1.2 ANSWER 1.b)

μ = average insurance savings

Null Hypothesis $H_0: \mu \geq 432$ Alternative Hypothesis $H_A: \mu < 432$

Lower Tailed Test:

```
library(stats)
act <- 432
exp <- 395
sd <- 102
n <- 82
df <- n - 1
sig_lvl <- 0.05
number <- sd / sqrt(n)

t <- (exp - act) / (number)
pt(t, df, lower.tail = TRUE)
```

```
## [1] 0.0007544927
```

p-value = 0.0008 < sig-level = 0.05 => reject null hypothesis in favor of alternative, meaning false advertisement is happening

1.3 ANSWER 1.c)

sample size consisted of only people who switched to the insurance, meaning it was biased. Not worth agreeing

1.4 ANSWER 1.d)

```
conf_int <- 0.9
sig_lvl <- (1 - conf_int) / 2
z <- qnorm(sig_lvl, lower.tail = FALSE)

e <- z * number

interval <- c(exp - e, exp + e)
interval
```

```
## [1] 376.4723 413.5277
```

actual (432) is NOT within bounds of the confidence interval (376-413) meaning false advertising is happening

1.5 ANSWER 1.e)

The results from the hypothesis test and confidence interval do match:

- hypothesis test states false advertisement (alternative hypothesis)
 - confidence interval does not reach 432, meaning false advertisement
-

2. Speed Reading (1 pt)

2.0.1 OIStats v3 4.23

A company offering online speed reading courses

- claims that students who take their courses
 - show a 5 times (500%) increase in
 - the number of words they can read in a minute without losing comprehension.

A random sample of 100 students yielded

- an average increase of 415%
 - with a standard deviation of 220%.