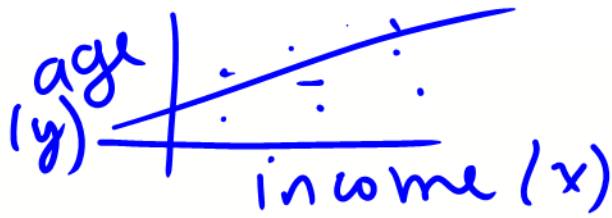
FALL 22 FINAL

Problem 6

In this question, we'll explore the relationship between the ages and incomes of credit card applicants.



Problem 6.1

The credit card company that owns the data in `apps`, BruinCard, has decided not to give us access to the entire `apps` DataFrame, but instead just a sample of `apps` called `small_apps`. We'll start by using the information in `small_apps` to compute the regression line that predicts the age of an applicant given their income.

For an applicant with an income that is $\frac{8}{3}$ standard deviations above the mean income, we predict their age to be $\frac{4}{5}$ standard deviations above the mean age. What is the correlation coefficient, r, between incomes and ages in `small_apps`? Give your answer as a **fully simplified fraction**.

Click to view the solution.

$= \frac{4}{5} * \frac{3}{8}$

$= 0.3$

$x_{su} = \frac{8}{3}$

predicted $y_{su} = \frac{4}{5}$

pred $y_{su} = r * x_{su} \rightarrow r = \frac{4/5}{8/3}$

Problem 6.2

Now, we want to predict the income of an applicant given their age. We will again use the information in `small_apps` to find the regression line. The regression line predicts that an applicant whose age is $\frac{4}{5}$ standard deviations above the mean age has an income that is s standard deviations above the mean income. What is the value of s? Give your answer as a **fully simplified fraction**.

y = income

x = age

$x_{su} = \frac{4}{5}$

ask for predicted $y_{su} = r * x_{su}$

$= \frac{3}{10} * \frac{4^2}{5}$

$0.24 = 6/25$

FALL 23 FINAL

Problem 11

On Reddit, Yutian read that 22% of all online transactions are fraudulent. She decides to test the following hypotheses:

- **Null Hypothesis**: The proportion of online transactions that are fraudulent is **0.22**.
- **Alternative Hypothesis**: The proportion of online transactions that are fraudulent is not **0.22**.

parameter

$= x$

$\neq x$

To test her hypotheses, she decides to create a **95%** confidence interval for the proportion of online transactions that are fraudulent using the Central Limit Theorem.

Unfortunately, she doesn't have access to the entire `txn` DataFrame; rather, she has access to a simple random sample of `txn` of size n. In her sample, the proportion of transactions that are fraudulent is **0.2** (or equivalently, $\frac{1}{5}$).

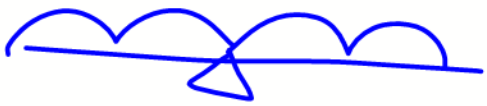
like human body temp example

Problem 11.1

95%

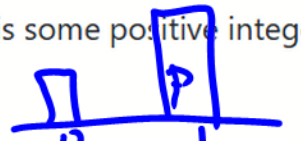
has SRS of size n in sample, 0.2 are fraudulent

The width of Yutian's confidence interval is of the form



$$\frac{c}{5\sqrt{n}}$$

where n is the size of her sample and c is some positive integer. What is the value of c? Give your answer as an integer.



Hint: Use the fact that in a collection of 0s and 1s, if the proportion of values that are 1 is p, the standard deviation of the collection is $\sqrt{p(1-p)}$.

$$4 * \frac{\text{SD of sample}}{\sqrt{\text{sample size}}} = 4 * \frac{\sqrt{0.2 * 0.8}}{\sqrt{n}} = \frac{4 * 0.4}{\sqrt{n}}$$

$$= \frac{1.6}{\sqrt{n}} = \frac{8/5}{\sqrt{n}} = \frac{8}{5\sqrt{n}}$$

$c = 8$

Problem 11.2

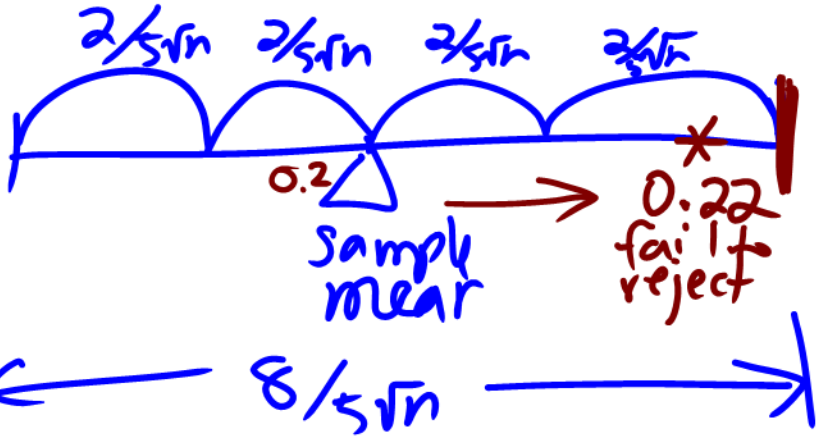
There is a positive integer J such that:

- If $n < J$, Yutian will fail to reject her null hypothesis at the **0.05** significance level.
- If $n > J$, Yutian will reject her null hypothesis at the **0.05** significance level.

question of does 0.22 fall in this interval

What is the value of J? Give your answer as an integer.

95% CI: sample mean $\pm$ 2 $\frac{\text{SD sample}}{\sqrt{\text{sample size}}}$



$$0.2 + 2 \cdot \frac{2}{5\sqrt{n}} < 0.2$$

$$\frac{4}{5\sqrt{n}} < 0.02$$

$$4 < 0.1\sqrt{n}$$

$$40 < \sqrt{n}$$

$$1600 < n$$