Assignment 4

PAI 721: Introduction to Statistics

Prof. Jack Reilly

F2025

Instructions

For this assignment, you must turn in two documents: 1) your answers (as a PDF) and 2) a plain-text `.do` file containing the Stata code you used to arrive at your answers.

> Note: For Questions 1–5, you may calculate by hand or using Stata. For Questions 6–12, you **must** use Stata.

Problem 1–2: DMV Processing Times

The Commissioner of the NYS Department of Motor V ehicles is analyzing the impact of a new software for processing new Drivers License applications. To do so, he divided DMV employees at the Syracuse branch into two groups:

- **GROUP A** uses the current system
- **GROUP B** uses the new software

Below is a table showing the time it takes (in minutes) to people in each group to process applications.

Group	Times (minutes)
A	14, 18, 16, 22, 25, 12, 32, 16, 15, 18
B	12, 10, 13, 14, 9, 17, 11, 10, 8, 11

1. Compute **mean**, **median**, **first quartile (Q1)**, **third quartile (Q3)**, and **range** for each group.

2. Based on these summaries, which software should the DMV use? Explain briefly.

Problem 3: Comparing SAT vs ACT

There are two major tests of readiness for college, the ACT and the SAT. ACT scores are reported on a scale from 1 to 36. The distribution of ACT scores is approximately Normal with mean = 21.5 and standard deviation = 5.4. SAT scores are reported on a scale from 600 to 2400. The SAT scores are approxi- mately Normal with mean = 1509 and standard deviation = 321.

- Tonya scores **1820** on the SAT.

- Jermaine scores **29** on the ACT.

3. If both tests measure the same thing, who has the higher score? Report both **z-scores**.

Problems 4–5: NSFG (2011–2013)

The following questions are based on information from the National Survey of Family Growth (NSFG) from 2011-2013. This is the nationally representative survey with information on fertility in the United States. From this dataset, you learn that the age of a woman at the time of her first birth is N(24,12).

4. What is the probability that a randomly drawn person is **older than 30** at first birth?

5. What percent of women are **teenagers (19)** when they have their first birth?

Problems 6–12: Madison County Jail Data (Stata)

For the following questions, use the dataset ”Madison county jail data.dta” located on blackboard. The unit of analysis is prisoner (each row in the dataset represents a different prisoner). The dataset currently has the following two variables:

- **Number of days**: This variable captures the number of days that a person has spent in jail.
- **Education level**: This variable captures the Highest education level achieved by the individual imprisoned. The numbers in the value represent categories in the following way:

  – 1 = Elementary School

  – 2 = Middle School

  – 3 = High School

  – 4 = College

  The director of the Madison County jail is planning to testify before the Wisconsin state legislature with the goal of obtaining more state funds for the county jail system. In preparing her remarks, the director plans to present data on the average number of days that prisoners remain in jail as well as data on the educational background on the individuals. The director asks you, as chief operations officer for the jail, to calculate descriptive statistics for a random sample of 250 prisoners processed in the last 9 months.

6. Obtain **one appropriate measure of central tendency** for `Number of days` and for `Education level`. Explain the choice and interpret it.

7. Create a **plot** for `Number of days` appropriate to its type. State whether the distribution is **symmetric, left-skewed, or right-skewed** and how you know.

8. Create a **plot** for `Education level` appropriate to a categorical variable.

9. Using a **linear transformation**, create `cost = 86 + 35 * (Number of days)`. Obtain a **summary** of `cost`.

10. **Regress** `cost` (y) on `Number of days` (x). Does the regression recover the linear transformation from (9)? Explain the intercept and slope.

11. **Regress** `Number of days` (y) on `Education level` (x). Report and interpret the **intercept** and **slope**.

12. Using results from (11), predict expected days for education levels **1–4** (Elementary, Middle, High School, College).

13. Briefly discuss **reservations/limitations** with the regression in (11) and predictions in (12) (e.g., treating education as numeric, omitted variables, linearity).