SOC 5050: Problem Set 04

Christopher Prener, Ph.D.

September 26th, 2016

Directions

Please complete all steps below. The final parts of this lab use the 2013 NHIS data. Your final do-files, log-file, plots, and markdown file with answers should be uploaded to your GitHub assignment repository by 4:20pm on Monday, October 3rd, 2016.

Part 1: Random Variables

For each question, provide both the probability as well as an interpretation that places the probability value into the context of the given scenario.

Scenario 1: A social services agency is helping its clients apply for a state housing program. One client's outcome in the application process does not predict another client's outcome, and each client has the same probability of being selected for the program because of the way that it is structured. The state reports back to the social services agency whether each client is 'accepted' or 'rejected'. The probability of being accepted is .35, and the agency is submitting 142 names to the program for consideration.

1. Evaluate the probability of 100 or more clients being selected?
2. Evaluate the probability of exactly 38 clients being selected?
3. Evaluate the probability of 25 or fewer clients being selected?
4. Evaluate the probability of 50 or more clients being selected?
5. Evaluate the probability of 80 or fewer clients being selected?
6. Evaluate the probability of exactly 75 clients being selected?

Scenario 2: A city health department is interested in evaluating the physical activity of residents who live in the city's elderly housing high rises. They use a test known as the SF-12v2, which evaluates each resident's mental and physical health. The result of the SF-12v2 is a variable named `pcs` - the "physical component score" - which is normed in such a way that the variable's distribution is

approximately normal with a mean of 50 and a standard deviation of 10.

7. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 42.
8. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 54.
9. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 84.
10. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 46.

Scenario 3: The FAA is interested in evaluating the risk of passenger fatalities due to airline accidents. In 2013, approximately 3.1 billion passengers flew on commercial aircraft. The five year average of fatalities in airline accidents covering 2009-2013 is 442.8 fatalities per year. The probability of dying in an airline accident for any one passenger is therefore estimated to be 0.000000143. In 2014, approximately 3.2 billion passengers flew on commercial aircraft. Assume that the probability of dying in an accident remained steady from 2013 to 2014.[1]

[1] These data were obtained from the 2014 and 2015 International Civil Aviation Organization annual reports.

11. Evaluate the probability of having exactly 457 airline accident fatalities in 2014.
12. Evaluate the probability of having 400 or fewer airline accident fatalities in 2014.
13. Evaluate the probability of having 500 or more airline accident fatalities in 2014.
14. The actual number of airline fatalities in 2014 was 904. It was large because the two Malaysia Airlines Flights - MH 17 and MH 370 - that crashed.[2] Evaluate the probability of having exactly 904 fatalities in 2014.

[2] MH 17 was the flight that crashed in Ukraine after being struck by a surface-to-air missle. MH 370 was the flight that disappeared from radar over the Indian Ocean.

15. The actual numbers of fatalities for the period covering 2009-2013 were 655, 626, 372, 388, and 173 respectively. Calculate the variance for this distribution. Does this impact your assessment of these probabilities?

Part 2: Skew and Kurtosis by Hand

16. The following are a distribution of scores on an evaluation of respondents' potential lifetime earnings: 25, 75, 28, 11, 66, 6, 48, 86, 94, 25. What is the skewness and kurtosis of this distribution of scores?

Part 3: Normality Testing in Stata

17. Use the variable `AGE_P` to conduct a full set of normality tests:
    (a) What is the variable's skew?
    (b) What is the variable's kurtosis?
    (c) Create and interpret a well laid-out p-p plot.
    (d) Create and interpret a well laid-out q-q plot.
    (e) Is the variable `AGE_P` appropriate for using either the Shapiro-Wilk or Shapiro-Francia normality tests?
    (f) Regardless of your answer to the above question, run and interpret both hypothesis tests.

Grading Rubric

Part 1 This section is worth ten points towards the total, with a third of a point point given for obtaining the correct probability and a third of a point point given for the evaluation of this probability on each question.

Part 2 This section is worth three points.

Part 4 This section is worth seven points - one point each for 17(a) through 17(e) and two points for 17(f).

Stata Do-File The overall quality of the Stata do-file stack is worth seven points. This grade will be based on the clarity, organization, and layout of your do-files.

Design An additional three points are based on the layout and design of each of your figures. This grade will be based on the use of schemes as well as customization of the plots (titles, subtitles, and notes).

Document Details

Document produced by Christopher Prener, Ph.D. for the Saint Louis University course SOC 5050 - QUANTITATIVE ANALYSIS: APPLIED INFERENTIAL STATISTICS. See the course wiki and the repository `README.md` file for additional details.