SOC 4015/5050: PS-02 - The Distribution of Random Variables

Christopher Prener, Ph.D.

Fall 2018

Directions

Please complete all steps below. The final parts of this assignment use the `gss14` data from the `testDriveR` package. This assignment should be uploaded to your Assignments Repository by 4:15pm on Monday, October 1st, 2018.

Analysis Development: Create a Project Folder System

This initial section follows the project workflow that is available in the `lecture-03` repo!

1. Using RStudio, add an R Project to the *existing* directory in your assignments repository named `PS-02`.
2. Add a new folder named `docs` to you project.
3. Create a new text file for your `README.md`. In the body of your `README.md` file, use Markdown formatting to write a sentance or two describing the purpose of this project. Then create an outline using bullets of the contents of the project itself.[1]
4. Create a new notebook with an expanded YAML heading.
5. Make sure your notebook has *completed* introductory, package loading, and data loading sections before proceeding with the parts below.
6. Be sure to "knit" your notebook at the end!

[1] See my write-up of the Markdown syntax in *Sociospatial Data Science* for details on creating lists.

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.

7. What is the appropriate distribution for these data? Make sure you provide a justification for your answer.
8. Evaluate the probability of more than 99 clients being selected?
9. Evaluate the probability of exactly 38 clients being selected?
10. Evaluate the probability of 25 or fewer clients being selected?
11. Evaluate the probability of 50 or more clients being selected?
12. Evaluate the probability of 80 or fewer clients being selected?
13. 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.

14. What is the appropriate distribution for these data? Make sure you provide a justification for your answer.
15. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 42.
16. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 54.
17. Evaluate the cumulative probability that a resident picked at random has a `pcs` value of 84.
18. 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.[2]

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

19. What is the appropriate distribution for these data? Make sure you provide a justification for your answer.
20. Evaluate the probability of having exactly 457 airline accident fatalities in 2014.
21. Evaluate the probability of having 400 or fewer airline accident fatalities in 2014.
22. Evaluate the probability of having 500 or more airline accident fatalities in 2014.
23. 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.[3] Evaluate the probability of having exactly 904 fatalities in 2014.
24. 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.[4] Does this impact your assessment of these probabilities?

[3] 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.

[4] See the course website to do this in `R`.

Part 2: Normality Testing

25. Using the `gss14` data, subset it so that you only have two columns left - `ID_` and `HRS1`. Rename both of those columns.
26. Use your *renamed* version of variable measuring hours worked 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 q-q plot for this variable.
    (d) Is the variable appropriate for using the Shapiro-Francia normality test?
    (e) Regardless of your answer to the above question, run and interpret this test.

Rubric

Individual Questions

Part 1		Part 2	
Question	Points	Question	Points
7 through 23	.5	25	4
24	1.5	26	6
Points Possible	10		10

Note: Partial credit possible

Project Organization

Category	Details	Points
Excellent	`PS-02` organized following workflow without error	5
Good	Minor concerns with organization	4.5
Improvement Needed	Significant concerns with organization	3
Unsatisfactory	No organizational strategy used	0
Points Possible		5

Notebook Formatting & RMarkdown

Category	Details	Points
Excellent	Syntax used appropriately & without error	5
Good	Minor concerns with syntax use	4.5
Improvement Needed	Significant concerns with syntax	3
Unsatisfactory	No RMarkdown used	0
Points Possible		5

Literate Programming

Category	Details	Points
Excellent	Narrative throughout with great detail	5
Good	Some narrative with inconsistent detail	4.5
Improvement Needed	Limited narrative with little detail	3
Unsatisfactory	No narrative included	0
Points Possible		5