Winter 2019 – ST 314 Data Analysis 02

Topics:
- Continuous Random Variables and their Probability Density Functions
- Lessons Covered: 8 - 13
  Textbook Chapter (Optional): 4

Grading:
- Points are listed next to each question and should total 25 points overall.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:
- Final Submission: **Monday, January 21st**. All submissions must be PDF files.

Instructions:
- Clearly label and **type answers** to the questions on the proceeding pages, ***without*** question prompts, in Word, Google Docs, or other word processing software.
- Insert **diagrams or plots as a picture** in an appropriate location.
- Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. *, /, sqrt() and ^
- Submit assignment to the Canvas link as a PDF. Verify the correct document has been uploaded. If not, resubmit. You can submit up to three times.

Allowances:
- You may use any resources listed or posted on the Canvas page for the course.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

Winter 2019 – ST 314 Data Analysis 02

Part 1. (6 points) Identify the distribution

For each random variable:

a. State the distribution that will best model random variable. Choose from the common distributions: Uniform, Exponential or Normal distribution. *Explain* your reasoning.
b. State the parameter values that describe the distribution.
c. Give the probability density function.

Random Variable 1.

A statistics student has a part time job as a coffee shop barista, she realizes the time between customer orders is a random variable. During an eight-hour shift, she measures time between successive customer order and finds that the time between customer orders is on average 30 seconds. Furthermore, she discovers times are more likely to be close to 0, and less likely as they get further away from 0.

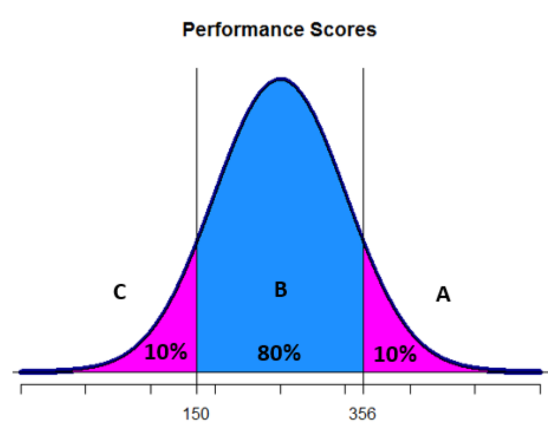
Random Variable 2.

In a board game, individuals must attempt to guess a phrase based on clues from their teammate. If they successfully guess the phrase before a buzzer sounds, their teammate may give clues for another phrase. Each correctly guessed phrase, before the buzzer sounds, gives them a point in the game. The buzzer is set to a random time increment anywhere between 30 and 90 seconds. Consider time until the buzzer sounds a random variable where any time between 30 and 90 has an equal likelihood.

Random Variable 3.

An industrial process yields a large number of steel cylinders. The length of the cylinder is a random variable with an average of 3.25 inches and a standard deviation of 0.003 inches. The distribution of cylinder lengths is symmetrical, where lengths are more likely to be close to the mean rather than further away from the mean.

Part 2. (6 points) Normal Distributions Some companies "grade on a bell curve" to compare the performance of their managers and professional workers. This forces the use of some low performance ratings so that not all workers are listed as "above average." Ford Motor Company's "performance management process" for this year assigned 10% A grades, 80% B grades, and 10% C grades to the company's managers. Suppose Ford's performance scores really are Normally distributed. This year, managers with scores less than 150 received C grades and those with scores of at least 356 received A grades.



a. (2 point) What are the z scores associated with the 10^{th} and 90^{th} percentiles from the standard normal distribution?
b. (2 point) What is the mean and standard deviation of the performance scores? Show work.

c. (2 point) Suppose the company adds grades D and F so there are 5 categories to grade performance. If they want to give A's only to those in the top 3%, what performance score must a manager exceed to get an A?

Part 3. (13 points) Simulation of Gamma Random Variables

Background: When we use the probability density function to find probabilities for a random variable, we are using the density function as a model. This is a smooth curve, based on the shape of observed outcomes for the random variable. The observed distribution will be rough and may not follow the model exactly. The probability density curve, or function, is still just a model for what is actually happening with the random variable. In other words, there can be some discrepancies between the actual proportion of values above x and the proportion of area under the curve above the same value x. Our expectation is as the number of observations increase, literally or theoretically, the observed distribution will align more with the density curve. Over the long run, the differences are negligible, the model is sufficient and more convenient to find desired information.

Simulation: Use R to simulate 1000 observations from a gamma distribution. To begin alpha = 2 and beta = 7. Highlight and run the parameters and observation values. Run the simulation code to plot the observations and fit the probability density function over the observations. You don't need to change anything. You may run the section all at once by highlighting all of the section and running it by clicking the run button at the top of the script window.

a. Given the values are from a gamma distribution with alpha= 2 and beta = 7,
   i. (1 points) What is the expression for the probability density function?
   ii. (1 point) What is the average and standard deviation of random variable? Show work.
   iii. (1 point) What is the probability x is less than 4? Show work.
b. (2 point) Run the simulation and paste your plot. Comment on the general shape of the distribution. How well does the density curve fit the observations?
c. (2 point) What is the exact proportion of values below 4? How does the actual proportion compare to the probability from the density curve in part 2-a-iii?
d. (1.5 point) Increase the number of observations to 10000, rerun the simulation. Paste your plot. How does increasing the number of observations affect the fit of the density curve?
e. (1.5 point) What is the exact proportion of values below 4? How does increasing the number of observations affect the accuracy of the model? Make a comparison between this proportion and 2-a-iii and 2c.
f. (1 point) Rerun the simulation with alpha = 1, beta = 7, and observations = 10000. Paste your plot. Comment on the general shape of the distribution.
g. (2 points) This model is a special case of the gamma distribution, what is it specifically? What is the expression for the probability density function?
h. Optional: Change the parameter values and take note of the effect of increasing or decreasing parameter values.