ST 314 Data Analysis 02

Part 1. (6 points) Identify the distribution

Random Variable 1. (2 points)

A board game buzzer is set to a random time increment anywhere between 45 and 120 seconds. Consider time until the buzzer sounds a random variable where any time between 45 and 120 has an equal likelihood. Players of the game must guess a phrase from clues given by their teammates before the buzzer sounds.

a. State the distribution that will best model random variable. Choose from the common continuous distributions: Uniform, Exponential or Normal distribution. *Explain* your reasoning.

   The best model for this random variable would be Uniform distribution since each increment between 45 and 120 seconds is equally likely.

b. State the parameter values that describe the distribution.

   The parameter values we need are the minimum and maximum values for a uniform distribution, so they would be a = 45 and b = 120.

c. Give the probability density function.
   In this case, the probability density function would be f(x) = $\frac{1}{75}$ for 45 ≤ x ≥ 120.

Random Variable 2. (2 points)

The time between patients arriving at an emergency room is a random variable. During a one-hour period, a staff member measures the time between successive patients and finds that, on average, the time between patients is 7 minutes. Furthermore, they observe that times are more likely to be close to 0 and less likely as they get further from 0.

a. State the distribution that will best model random variable. Choose from the common continuous distributions: Uniform, Exponential or Normal distribution. *Explain* your reasoning.

   The best model for this random variable would be Exponential distribution since the events occur independently and at a constant average rate.

b. State the parameter values that describe the distribution.

   The parameter values in this case would be the rate (lambda), which is the reciprocal of the mean. So in this case, the lambda would be $\frac{1}{7}$.

c. Give the probability density function.
   The probability density function would be $f(x) = \frac{1}{7}e^{-\frac{1}{7}x}$ for x ≥ 0

Random Variable 3. (2 points)
Steel cylinders are produced as a part of a manufacturing process. The length of the cylinder is a random variable with an average of 3.25 inches and a standard deviation of 0.009 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.

a. State the distribution that will best model random variable. Choose from the common distributions: Uniform, Exponential or Normal distribution. *Explain* your reasoning.

   The best model for this random variable would be normal distribution because it is symmetric and the values near the mean are more likely. It is a bell curve.

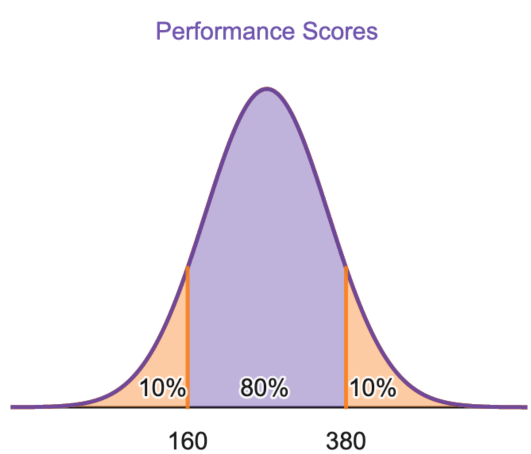
b. State the parameter values that describe the distribution.

   The parameters we need are the mean and the standard deviation.
   μ = 3.25 inches
   σ = 0.009 inches

c. Give the probability density function.

   In this case, the probability density function would be $f(x) = \frac{1}{0.009\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-3.25}{0.009})^2}$

Part 2. (6 points) Normal Distributions The Normal distribution curve to the right displays the distribution of grades given to managers based on management performance at Ford. Of the large population of Ford managers, 10% were given A grades, 80% were given B grades, and 10% were given C grades. A's were given to those who scored 380 or higher and C's were given to those who scored 160 or lower.



a. (2 point) What are the z scores associated with the 10^{th} and 90^{th} percentiles from the standard normal distribution? Recall that a z-score is value from the Standard Normal distribution and represents the number of standard deviations a value is away from its mean.

   The z-score for the 10^{th} percentile is -1.28 and it is 1.28 for the 90^{th} percentile. This means that these percentiles are 1.28 standards deviations away from the mean.

b. (2 point) From part a, you should have two values - the z-scores associated with the 10th and 90th percentiles. Using these two values and the mathematical definitions of a z-score, calculate the mean and standard deviation of the performance scores? Show work.

To calculate this, we can use the equation $z = \frac{(X-\mu)}{\sigma}$
So for the 10th percentile, we have $-1.28 = \frac{(160-\mu)}{\sigma}$ and for the 90th percentile we have $1.28 = \frac{(380-\mu)}{\sigma}$. Using these 2 equations we can solve for the mean and the standard deviation. In the end, the mean is 270 and the standard deviation is 86.

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?

If we were to limit A's to the top 3%, a manager must exceed a score of 431.6, which corresponds to the 97th percentile.

Part 3. (11 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, set alpha = 2 and beta = 4. 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 = 4,
   i. (1 points) What is the expression for the probability density function?

      The PDF function for this problem is $f(x) = 4^2 x e^{-4x}$ for x > 0

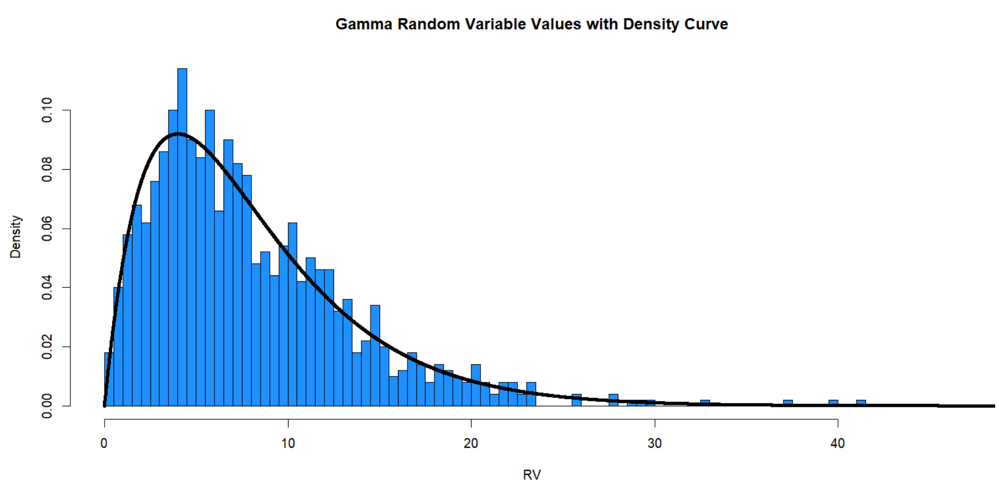
ii. (1 point) What is the average and standard deviation of the random variable? Show work in regards to how you derived these quantities.

The average of a gamma distribution can be calculated by $\mu = \frac{\alpha}{\beta}$ and the standard deviation can be calculated by $\sigma = \frac{\sqrt{\alpha}}{\beta}$. If we replace the placeholders with 2 and 4, we get a mean of 0.5 and a standard deviation of $\frac{\sqrt{2}}{4}$.

iii. (1 point) What is the probability x is less than 6? Show work.

To calculate the probability that x is less than 6, we would need to integrate the PDF from 0 to 6: $\int_0^6 4^2 x e^{-4x}$. If you were to integrate this out, you would get a probability of 0.44.

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?

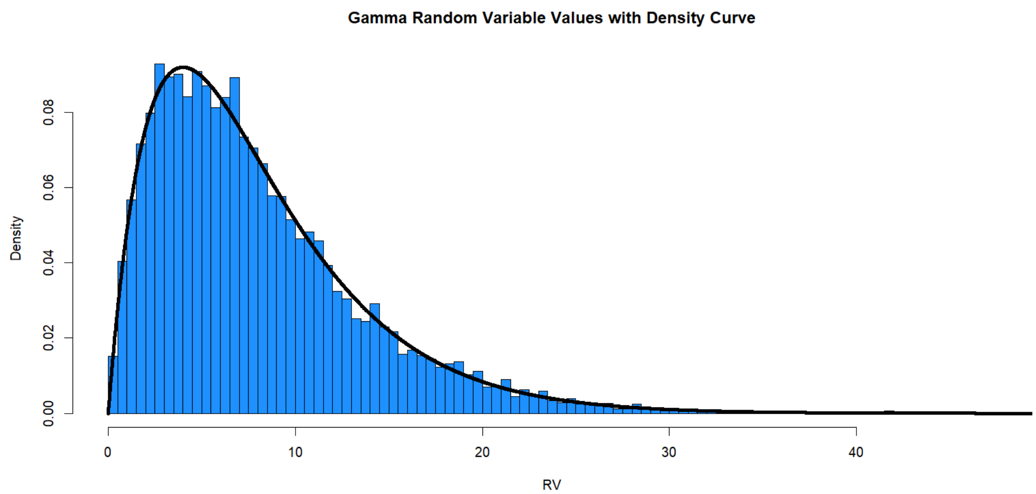


The density curve fits the observations fairly well, but definitely not perfectly. There are some outliers and some observations on the lower RV end are too dense.

c. (2 point) What is the exact proportion of values below 6? How does the actual proportion compare to the probability from the density curve in part 2-a-iii?

The exact proportion of values below 6 is 0.48, which is close to the answer of 0.44 I got in part 2.

d. (1 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?

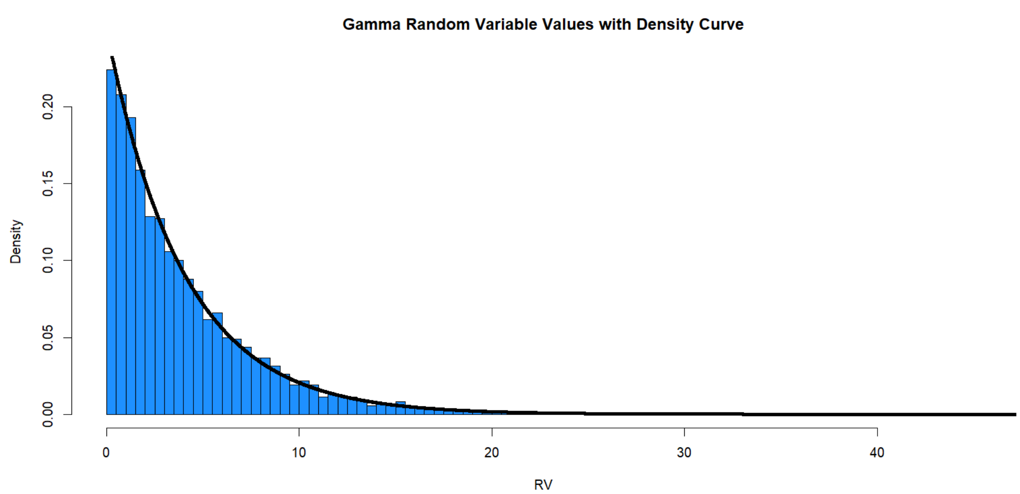


Increasing the number of observations allows them to fit tighter within the density curve.

e. (1 point) What is the exact proportion of values below 6? 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.

The proportion value is now 0.439 which is incredibly close to our answer of 0.44 from earlier. Increasing the number of observations makes our proportion value more accurate.

f. (1 point) Rerun the simulation with alpha = 1, beta = 4, and observations = 10000. Paste your plot. Comment on the general shape of the distribution.



Changing the alpha value to one makes the curve start dense and decline quickly.

g. (1 point) The model in part (f) is a special case of the gamma distribution, what is it specifically? What is the expression for the probability density function?

   This special case of the gamma distribution is known as exponential distribution. The PDF expression for this model would be $f(x) = 4e^{-4x}$ for x ≥ 0

h. Optional: Change the parameter values and take note of the effect of increasing or decreasing parameter values.

Gradescope Page Matching (2 points)
When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.