Unit 2: Homework

John Andrus

September 8, 2020

1 Processing Pasta

A certain manufacturing process creates pieces of pasta that vary by length. Suppose that the length of a particular piece, L, is a continuous random variable with the following probability density function.

$$f(l) = \begin{cases} 0 & l \leq 0 \\ c \cdot l & 0 < l \leq 2 \\ 0 & 2 \leq l \end{cases}$$

1. Compute the constant C.
   For a continuous random variable L with PDF $F(l)$:

   $$\int_{-\infty}^{\infty} F(l)dl = 1$$

   Given $f(l)$

   $$\int_{-\infty}^{0} 0dl + \int_{0}^{2} c \cdot ldl + \int_{2}^{\infty} 0dl = 1$$

   $$\int_{0}^{2} c \cdot ldl = 1$$

   $$c = 0.5$$

2. Write down the complete expression for the cumulative probability function of L

   $$F(L \leq l) = \int_{-\infty}^{l} f(l)dl$$

   For $0 \leq l \leq 2$:

   $$F(L \leq l) = \int_{0}^{2} \frac{l}{2}dl$$

Which gives a CDF of :

$$F(L \leq l) = \begin{cases} 0 & l \leq 0 \\ \frac{l^2}{4} & 0 < l \leq 2 \\ 1 & 2 \leq l \end{cases}$$

3. Compute the median value of L.

$$L_{median} = \frac{l^2}{4} \text{ , where l} = 0.5$$

$$L_{median} = \frac{l^2}{4} = 0.5$$

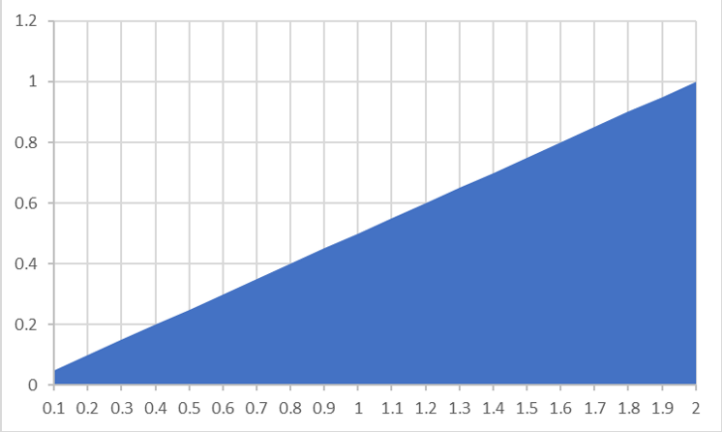
$$l = \sqrt{2}$$

2 Broken Rulers

You have a ruler of length 2 and you choose to break it using a uniform probability distribution. Let random variable X represent the length of the left piece of the ruler. X is distributed uniformly in [0,2]. You take the left piece of the ruler and once again choose a place to break it using a uniform probability distribution. Let random variable Y be the length of the left piece from the second break.

1. Draw a picture of the region in the X-Y plane for which the joint density of X and Y is nonzero.

   I had to rush through this question. Apologies if it doesn't make a lot of sense :(



2. Compute the joint density function for X and Y.

$$f_{XY} = \int_{-\infty}^{0} 0 + \int_{0}^{2}\int_{0}^{x} 0.5y + \int_{0}^{\infty} 0 = 0.5x$$

3. Compute the marginal probability density for Y, $f_Y(y)$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}dx = \int_{0}^{2} 0.5xdx = [\frac{x^2}{4}|_0^2 = 1$$

$$f_Y(y) = \begin{cases} 1 & 0 \leq x \leq 0 \\ 0 & otherwise \end{cases}$$

4. Compute the conditional probability density function of X, conditional on $Y = y$, $f_{X|Y}(x|y)$. Make sure you state the values of Y for which this exists.

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{0.5X}{1} = 0.5x$$

$$f_{X|Y}(x|y) = \begin{cases} 0.5x & 0 \leq x \leq 0 \\ 0 & otherwise \end{cases}$$

3 Concert Time

You are excited about a concert featuring your favorite a capella group: The Pitch Estimators. Tickets go on sale at noon, but before you can buy a ticket, you have to wait your turn in an online waiting room. Because tickets are in high demand, you enlist two of your friends to help you. All three of you enter the waiting room at noon, and as soon as one of you gets a ticket, you are done and you can all sign off.

Suppose your waiting time in minutes is a continuous random variable T. Your first friend's waiting time is a continuous random variable U. Your second friend's waiting time is a continuous random variable V.

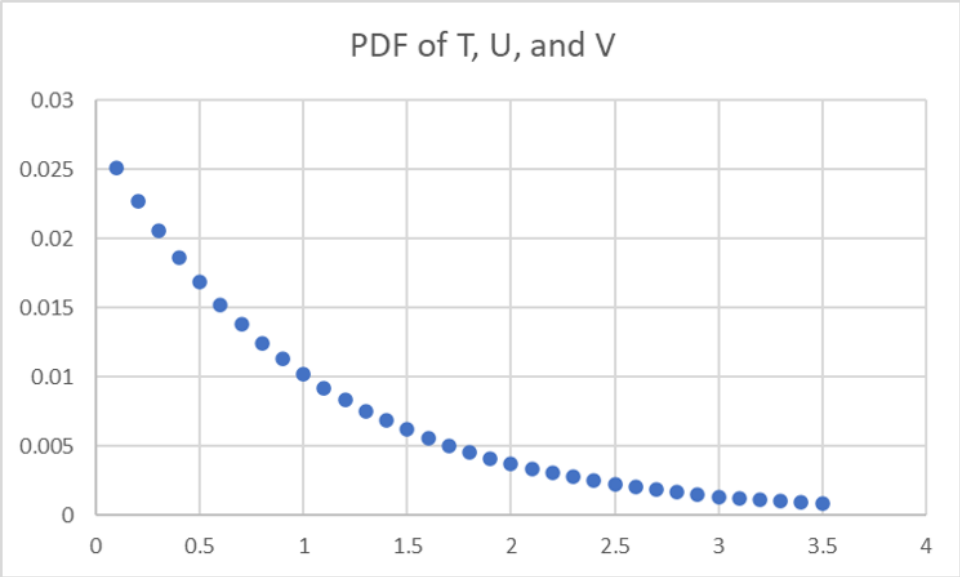
Suppose these random variables have probability density functions given by:

$$F_T(t) = \begin{cases} \frac{1}{3}e^{-\frac{1}{3}t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$U_T(t) = \begin{cases} \frac{1}{2}e^{-\frac{1}{2}t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$V_T(t) = \begin{cases} \frac{1}{6}e^{-\frac{1}{6}t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

1. Please sketch the probability density function for all three random variables on one graph.



2. For a particular time t, compute the probability that $T > t$, $U > t$, and $V > t$.

   For each random variable, the probability will be the complement of the cumulative density function.

$$P(T > t, U > t, V > t) = 1 - F_{T,U,V}$$

   Given that T, U, and V are independent:

$$f_{T,U,V} = f_T f_U f_V$$

   Plugging in f_T, f_U, f_V:

$$f_{T,U,V} = \frac{1}{36} e^{-t}$$

   Integrate the PDF to yield the CDF:

$$\int_0^t \frac{1}{36} e^{-t} dt = \frac{1}{36}(1 - e^{-t})$$

   Therefore:

$$P(T > t, U > t, V > t) = \begin{cases} 0 & t < 0 \\ 1 - \frac{1}{36}(1 - e^{-t}) & t \geq 0 \end{cases}$$

3. Let O represent your overall waiting time. O is the minimum of the waiting times for each of the three websites, $O = min(T, U, V)$. Compute the probability density function of O. (Hint: Use the previous answer to write down the cdf)

   The minimum of the waiting times, O, is equivalent to the expression solved in part two:

$$O(t) = P(T > t, U > t, V > t) = \begin{cases} 0 & t < 0 \\ 1 - \frac{1}{36}(1 - e^{-t}) & t \geq 0 \end{cases}$$

4 Lecture Fail

Suppose that the time of an event is a random variable T. One way to describe the distribution of T is with the hazard rate. If T has a pdf f_T and cdf F_T, then the hazard rate is given by:

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)}$$

The hazard rate at time t is the probability density conditional on the event $T \geq t$. To put it roughly, the hazard rate indicates how likely an event is to occur right away, given that it has not already occurred.

1. Say that the time a server breaks down is a random variable B, which is uniformly distributed on [0,2]. Compute the hazard rate of B.

   Given that B has uniform distribution over the interval [0,2]:

   $$f_B(b) = \begin{cases} 0.5 & 0 \leq b \leq 2 \\ 0 & otherwise \end{cases}$$

   Calculate the CDF as the integral of the PDF:

   $$F_B(b) = \begin{cases} 0 & b < 0 \\ 0.5b & 0 \leq b \leq 2 \\ 0 & b > 2 \end{cases}$$

   Plug these into the definition of hazard rate given in the problem description:

   $$F_B(b) = \begin{cases} 0 & b < 0 \\ \frac{0.5}{1-0.5b} & 0 \leq b \leq 2 \\ 0 & b > 2 \end{cases}$$

2. Prove that if X is a random variable with hazard rate h_X and Y is a random variable with hazard rate h_Y, then the hazard rate of $min(X, Y)$ is $h_X + h_Y$. (Hint: Write the hazard rate in terms of just the cdf, no pdf. Then remember what the cdf of a minimum of random variables looks like from the previous problem.)

5 Testing for Coronavirus

What we learn from a statistical test depends crucially on the population prevalence of the disease being tested for. Suppose that you are interested in a rapid assay for the coronavirus. Let T be the event that the Test comes back positives, C be the event that an individual has Coronavirus and P be the population prevalence of the disease.

The test is designed with a sensitivity of 0.94 and a specificity of 0.96. (Note: the instructors dislike the use of these terms because they are unnecessary jargon.

1. Define sensitivity and specificity in terms of the events above.

   Sensitivity is the probability that an individual tests positive for Coronavirus given that they have the disease.

   Specificity is the probability that in individual tests negative for Coronavirus given that they do not have the disease.

2. You are interested in the false discovery rate. Write a function that takes population prevalence as an argument and returns the false discovery rate.

   False Discovery Rate (FDR) is the expected proportion of Type I errors defined as:

$$FDR = \frac{\text{Number of False Positives}}{\text{Total Positive Tests}}$$

   False Discovery Rate function in R:

```
false_discovery_rate <- function(population_prevalence){

  sensitivity = 0.94
  specificity = 0.96

  FDR = 1 - (sensitivity*population_prevalence)/
            (sensitivity*population_prevalence +
            (1-specificity)*(1-population_prevalence))

  print(FDR)

}
```

3. Using the function that you have just written and the data supplied in the object d below, create a plot, using ggplot that has the following characteristics:

   - On the x-axis: The population prevalence rate
   - On the y-axis: The false discovery rate
   - Meaningful axis and plot titles

   False Discovery Rate plot in R:

```
d %>%
  ggplot() +
  aes(x=population_prev, y=FDR) +
  geom_line() +
  ggtitle("False Discovery Rate as a function
  of Population Prevalence") +
  xlab("Population Prevalence") +
  ylab("False Discovery Rate")
```