**Probability distributions for continuous random variables**

Examples of continuous random variables:

* Miles per gallon (MPG) a car gets for its gas consumption
* Amount of precipitation for a city during a year
* Height for people
* GPA
* Price of diamonds
* Salary
* Distance for a field goal in football
* Lifetime before failure of an object
* How long it takes you to get to class everyday

Continuous random variables have an infinite number of values that they can take on within an interval. However, all of these examples are limited by our ability to measure them. For example, the amount of precipitation may only be measured to two decimal places. The distance of a field goal is typically rounded to the nearest yard. While these variables may be measured as discrete random variables, they often can be approximated as continuous. This is especially true if there are a lot of values the random variable can take on.

The possible values that a continuous random variable can take on can be quantified using a probability distribution. These probability distributions will be very similar to what we saw for discrete random variables, but there are some VERY IMPORTANT DIFFERENCES! These similarities and differences are explained in the next example.

Example: GPA probability distribution (gpa.R)

Let Y be a random variable representing GPAs of students on campus. Suppose the probability distribution function for the random variable Y is



for 0 ≤ y ≤ 4 and f(y) = 0 for all other possible values of y. Notice the use here of Y and y again.

Below is a plot of the probability distribution

> curve(expr = 15/2048 \* x^4 \* (4 - x), xlim = c(0,

 4), col = "red", main = "Probability

 distribution for GPAs", ylab = "f(y)", xlab =

 "y", lwd = 2)

> abline(h = 0)



The “area” underneath the curve represents probability. For example, P(2.5 < Y < 3.5) is represented by the area underneath the plotted curve between 2.5 and 3.5 and above.

How do you find this area? Use calculus!!!

 

Questions:

* Suppose the interval was smaller, say 2.9 to 3.1. What will happen to the area underneath the curve?
* What happens if this interval becomes smaller and smaller and … ? What is P(Y = 3) then?

Unlike probability distributions for discrete random variables, we cannot simply substitute a value for y into the mathematical function to find a probability. For example, f(3) =  P(Y = 3).

Characteristics of a PDF for a continuous random variable

Suppose a continuous random variable Y has a PDF f(y). Then

* f(y) ≥ 0
* 
* P(a < Y < b) =  for constants a and b

Relate these bullets to what we saw before with discrete random variables.

Probability distribution functions for continuous random variables are often called probability density functions.

Why are we integrating to find probabilities? Please see the next example.

Example: MPG (MPG-uniform.R)

Suppose we are interested in the miles per gallon (MPG) of gas a car gets. Let X be a continuous random variable denoting the MPG for a particular car for one tank of gas. In a very simplistic setting, let



The function quantifies probabilities of getting particular MPG. This is actually a special type of probability distribution called a uniform probability distribution. Characteristics about it will be discussed later in the course.

Below is a plot of the function.

> # Create an empty plotting area

> plot(x = c(20,20,40,40), y = c(0,0.12,0,0.12), type =

 "n", xlab = "x", ylab = "f(x)")

> abline(h = 0, col = "black")

> segments(x0 = 19, y0 = 0, x1 = 25, y = 0 , col =

 "red", lwd = 2)

> segments(x0 = 25, y0 = 0, x1 = 25, y = 0.1, col =

 "red", lwd = 2)

> segments(x0 = 25, y0 = 0.1, x1 = 35, y = 0.1, col =

 "red", lwd = 2)

> segments(x0 = 35, y0 = 0, x1 = 35, y = 0.1, col =

 "red", lwd = 2)

> segments(x0 = 35, y0 = 0, x1 = 41, y = 0, col =

 "red", lwd = 2)



Area underneath the “curve” down to f(x)=0 represents probability (remember the field goal kicking). For example, the probability that the car gets MPG between 25 and 30 is P(25 ≤ X ≤ 30) is the area underneath the curve between 25 and 30. Because this is a rectangle, we can see use base×height = 5×0.1 = 0.5 as the probability.

In calculus, you learned how to find this area by integrating:



What is the entire area under the curve?

 

Thus, probability is equated to the area underneath a curve. The area underneath all curves that we will study in the class will be 1 because no probabilities can be greater than 1!

What is P(X = 30)? 0

Why?



In general, the probability a continuous random variable is equal to one value is 0. Also, notice this means

P(25 < X < 30) = P(25 ≤ X < 30)

 = P(25 < X ≤ 30)

 = P(25 ≤ X ≤ 30).

In a more realistic setting, MPG may be represented with the following PDF:

 for -∞ < x < ∞

This is another special type of probability distribution called a normal probability distribution. Characteristics about it will be discussed later in the course.

Below is a plot of the PDF (ignore the red part for now).



The R code is in the corresponding program, but you are not responsible for it until later in the course.

Possible values of X could be < 20 and > 29 here, but these are not displayed in the plot since the area underneath the curve on the ends is very small. According to this PDF, MPG could be negative; however, notice the probability will be EXTREMELY small for it to happen. Thus, this probability distribution can still be thought of as a realistic portrayal of MPG.

In this setting, we may be interested in the probability that the MPG is between 25 and 26. This probability is represented by the area in red in the plot above. It can be found by:

 

This integral is a lot more difficult to do in comparison to the previous example!

Example: Transaction time (Transaction.R)

The number of seconds between transactions (e.g., purchases) on a website can be represented by a PDF. Let X be a random variable representing the seconds. Suppose the PDF for X is



This is another special type of probability distribution known as an exponential probability distribution. Characteristics about it will be discussed later in the course.

Below is a plot of the PDF.

> # x and y-axes start/end at the specified x and y

 limits in curve()

> # Change back to the default with par(xaxs = "r",

 yaxs = "r")

> par(xaxs = "i", yaxs = "i")

> curve(expr = 1/30\*exp(-x/30), xlim = c(0,140), col =

 "red", ylab = "f(x)", xlab = "x")



Notes:

* Use the plot above to provide a general idea of how long between transactions.
* Why would a PDF for this application be important to know?
* Are there other examples where this PDF would be useful (outside of web transactions)?

Find the probability that the time between a set of randomly selected consecutive transactions will be longer than 30 seconds.



Let u = -x/30 then du = -1/30 dx ⇒ -30du = dx

Note that when x = 30, u = -1; when x = ∞, u = -∞.

 







= e-1 ≈ 0.367879

One could be more formal by working with limits for , but this is not necessary for our course.

In R, e-1 is found by using exp(-1).

Find the probability the time between a set of randomly selected consecutive transactions will be between 25 and 100 seconds.



= 

= 

= e-25/30 – e-100/30 ≈ 0.3989

Suppose the web service provider is concerned with meeting demand whenever there is less 0.5 seconds between transactions. Based on this PDF, should the web services provider be concerned?

VERY IMPORTANT POINT:

In addition to thinking of the total area underneath the curve as probability, one can think of this as the percentage of items in the population which satisfy a particular range of X. For example, the area underneath the curve for X > 30 could represent that 36.79% of all times are more than 30 seconds. Next are examples to help illustrate why this is important.

Example: A sample of times between transactions (Transaction.R, ExampleSample.csv)

Below is a sample that comes from a population characterized by the PDF of



> # Specify the folder AND the data file

> # I already specified the folder using setwd()

> # so I did not need to specify it in read.csv()

> set1 <- read.csv(file = "ExampleSample.csv")

> head(set1)

 Observation x

1 1 8.782565

2 2 2.810164

3 3 71.201537

4 4 16.549732

5 5 23.581126

6 6 2.165690

> tail(set1)

 Observation x

995 995 38.96618

996 996 44.53581

997 997 58.59082

998 998 17.57703

999 999 34.00815

1000 1000 52.04435

We will discuss how to take this sample in R later in the course.

Questions:

* What proportion of these times would you expect to be greater than 30?
* What proportion of these times would you expect to be between 25 and 100?

The proportion greater than 30 is found in R using the following code:

> # Estimate P(X > 30) using sample

> n <- length(set1$x)

> n

[1] 1000

> head(set1$x > 30)

[1] FALSE FALSE TRUE FALSE FALSE FALSE

> tail(set1$x > 30)

[1] TRUE TRUE TRUE FALSE TRUE TRUE

> sum(set1$x > 30)

[1] 377

> sum(set1$x > 30)/n

[1] 0.377

> mean(set1$x > 30)

[1] 0.377

The proportion between 25 and 100 is found in R using the following code:

> mean(set1$x > 25 & set1$x < 100)

[1] 0.412

> mean(set1$x < 100) - mean(set1$x < 25)

[1] 0.412

Notes:

* Both proportions are relatively close to the probabilities from the PDF.
* Suppose another sample is taken of size 1000. Would you expect the same proportions?
* Suppose a larger sample was taken. What would you expect to occur with these proportions relative to the probabilities?

A histogram of the observed values:

> hist(x = set1$x, xlab = "x", main = "")



The histogram bar heights have the same general shape as the PDF!

Find the frequencies for each category:

> save.hist <- hist(x = set1$x, xlab = "x")

> names(save.hist)

[1] "breaks" "counts" "density" "mids" "xname"

 "equidist"

> # Information on the classes and the frequencies per

 class

> save.hist$breaks #Notice there is one more "break" than

 there are counts

 [1] 0 20 40 60 80 100 120 140 160 180 200 220

> save.hist$counts

 [1] 480 256 139 77 26 9 7 4 1 0 1

> sum(set1$x > 0 & set1$x <= 20)

[1] 480

> sum(set1$x > 20 & set1$x <= 40)

[1] 256

> #Frequency distribution

> Rel.Frequency <- round(save.hist$counts /

 sum(save.hist$counts), digits = 2)

> Cumul.Rel.Freq = round(cumsum(save.hist$counts /

 sum(save.hist$counts)), digits = 2)

> data.frame(class = save.hist$breaks[-1], Frequency =

 save.hist$counts, Rel.Frequency = Rel.Frequency,

 Cumul.Rel.Freq = Cumul.Rel.Freq)

 class Frequency Rel.Frequency Cumul.Rel.Freq

1 20 480 0.48 0.48

2 40 256 0.26 0.74

3 60 139 0.14 0.88

4 80 77 0.08 0.95

5 100 26 0.03 0.98

6 120 9 0.01 0.99

7 140 7 0.01 0.99

8 160 4 0.00 1.00

9 180 1 0.00 1.00

10 200 0 0.00 1.00

11 220 1 0.00 1.00

The PDF can be overlaid onto the histogram to show how similar they are. To do this, I need to re-scale what appears on the histogram’s y-axis in order to match the scale for the PDF plot. This is done by dividing each frequency by the category width (20) and sample size, which is accomplished automatically in R using the freq = FALSE argument value in hist() (freq = TRUE is the default).

> # Histogram with PDF overlay

> # Changed y-axis scale to get PDF in plot better

> hist(x = set1$x, xlab = "x", main = "", freq = FALSE,

 ylim = c(0, 0.03))

> curve(expr = 1/30\*exp(-x/30), col = "red", add = TRUE,

 n = 1000)



Notice how similar they are! Now you have an example of what it means for a population to be characterized by a PDF!

Notes:

* R uses a default y-axis label of “density” to represent that the probability density function scale is being used.
* Suppose a smaller sample size was taken. What would you expect to happen for the similarity of the histogram shape and the PDF?
* For a general data problem outside of this example, how could one check if a PDF is appropriate for a particular data setting?

Example: Transaction time (Transaction.R, ExampleSample.csv)

Find the time between transactions such that 95% are less than this value.

Find c in 

0.95 = 

 = -(e-c/30 – e0)

 = 1 – e-c/30

⇒ e-c/30 = 1-0.95

⇒ c/30 = -log(0.05) where “log” means natural log

⇒ c = 30log(20) ≈ 89.87

The natural log is calculated as 30\*log(20) in R.

Thus, the time is 89.87 seconds and P(X < 89.87) = 0.95.

Find the time between transactions such that 5% are less than this value.

 ⇒ c = -30ln(0.95) ≈ 1.54

From the sample:

> quantile(x = set1$x, probs = c(0.05, 0.95), type = 5)

 5% 95%

 1.573308 78.490649

Question: Suppose a time is 1000 seconds. What could this indicate potentially about the website?

The past examples show the importance of finding probabilities such as P(Y < \_\_). Similar to how the cumulative distribution function was defined for discrete random variables, we can define it here for continuous random variables! The cumulative distribution function (CDF) for a continuous random variable Y, denoted F(y), is

F(y) = P(Y < y) = 

Notes:

* The t is used here to help avoid confusion between the limits of integration and the random variable.
* If the lower limit for the PDF is not -∞, use the this other lower limit and define F(y) = 0 for values less than this lower limit. See the upcoming example.
* F(y) “cumulates” probabilities as Y increases and how the function ranges from 0 to 1. Compare to the CDF for a discrete random variable Y:

F(y) = P(Y ≤ y) = 

Example: Transaction time (Transaction.R)

The CDF for X is

 = 1 – e-x/30

for 0 < X < ∞ and F(x) = 0 otherwise

Plot in R:

> curve(expr = 1 - exp(-x/30), xlim = c(0,140), col =

 "red", ylab = "f(x)", xlab = "x", panel.first =

 grid(), ylim = c(0,1.05))



Typically, CDFs are not shown where F(x) = 0 or too far past where it visually appears F(x) = 1.

The CDF for X provides a convenient way to find probabilities, such as P(X < 75) or P(X > 75).

P(X < 75) = F(75) = 1 – e-75/30 = 0.9179

P(X > 75) = 1 – P(X < 75) = 1 – F(75) = e-75/30 = 0.0821

**Very often**, we will want to find probabilities such as P(25 < X < 30). Here are two ways to do it:



and





Notice the 2nd way is MUCH easier if one has already obtained the CDF! Also, this a result from the “Fundamental Theorem of Calculus”.

We saw earlier that a histogram gives a graphical estimate of the PDF. Similarly, the empirical CDF (ECDF) gives a graphical estimate of the CDF! In other words, a plot of the cumulative relative frequency for the sample that can be calculated for any value of the random variable. The ECDF for a random variable Y is



where

* n is the sample size
*  is the indicator function;  = 1 when the item within parentheses is true and  = 0 otherwise
* y is a particular value of Y of interest; e.g.,  would be the proportion of observations in the sample that are less than or equal to 2.
* Sometimes, this sum is simply represented as



Example: A sample of times between transactions (Transaction.R, ExampleSample.csv)

We found the following earlier

> #Frequency distribution

> Rel.Frequency <- round(save.hist$counts /

 sum(save.hist$counts), digits = 2)

> Cumul.Rel.Freq = round(cumsum(save.hist$counts /

 sum(save.hist$counts)), digits = 2)

> data.frame(class = save.hist$breaks[-1], Frequency =

 save.hist$counts, Rel.Frequency = Rel.Frequency,

 Cumul.Rel.Freq = Cumul.Rel.Freq)

 class Frequency Rel.Frequency Cumul.Rel.Freq

1 20 480 0.48 0.48

2 40 256 0.26 0.74

3 60 139 0.14 0.88

4 80 77 0.08 0.95

5 100 26 0.03 0.98

6 120 9 0.01 0.99

7 140 7 0.01 0.99

8 160 4 0.00 1.00

9 180 1 0.00 1.00

10 200 0 0.00 1.00

11 220 1 0.00 1.00

> # ECDF with CDF overlay

> F.hat <- ecdf(x = set1$x) # Creates a function to find

 ECDF

> F.hat(seq(from = 20, to = 220, by = 20)) # Compare to

 previous calculations

[1] 0.480 0.736 0.875 0.952 0.978 0.987 0.994 0.998 0.999

 0.999 1.000

> sum(set1$x <= 20)/n

[1] 0.48

> sum(set1$x <= 40)/n

[1] 0.736

We can also use F.hat() with any other possible value of y:

> F.hat(0)

[1] 0

> F.hat(10)

[1] 0.262

> F.hat(300)

[1] 1

Thus,  when 26.2% of the observations were less than or equal to 10 in the sample.

Compare CDF to ECDF:

> plot.ecdf(x = set1$x, lwd = 2, panel.first = grid(),

 ylab = "Probability", xlab = "x", col = "blue", main

 = "", verticals = TRUE, do.p = FALSE)

> curve(expr = 1 - exp(-x/30), col = "red", add = TRUE, n

 = 1000, xlim = c(0,250))

> legend(x = 100, y = 0.6, legend = c("ECDF", "CDF"), lty

 = 1, col = c("blue", "red"), lwd = 2, bty = "n")



Notice how similar they are!

Note: Suppose you are given the CDF, how can you find the PDF? Take the derivative of the CDF with respect to the random variable! For a random variable Y, we have

****

Example: Transaction time (Transaction.R)

F(x) = 1 – e-x/30 for 0 < X < ∞ and F(x) = 0 otherwise

 **** 

Example: Wind speed in Lincoln (Lincoln\_Feb\_wind.csv)

Wind speeds are often characterized by PDFs! Suppose Y is the wind speed for a particular day. A PDF that often does a good job of accounting for Lincoln’s February wind speeds is



where δ, γ, and β are population parameters. This PDF is known as a three-parameter Weibull distribution. Below is the PDF with δ = 2.1327, γ = 1.8832, and β = 6.2926 plotted on a histogram of sampled wind speeds.



Notice how well f(y) serves as a “model” for reality!

How could this PDF be used?

Question: Can f(y) be greater than 1 for a continuous random variable Y?