Homework set 3 - Due 02/15/13

Math 3200 – Renato Feres

Preliminaries

We explore in this assignment a few of the most common discrete and continuous probability distributions. The basic theory is further described in sections 2.7, 2.8, and 2.9 of our textbook. R employs the following naming convention for functions associated to probability distributions, illustrated here with the normal distribution:

• `dnorm` density function (of the normal distribution)
• `pnorm` cumulative distribution function
• `qnorm` quantile function
• `rnorm` random variable

We’ll explain these functions shortly. They are available for many other probability distributions. Here is a short list of distributions, with their R names and associated parameters.

1. Discrete distributions:

<table>
<thead>
<tr>
<th>R name</th>
<th>distribution</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>binom</code></td>
<td>Binomial</td>
<td>( n = ) number of trials, ( p = ) probability of success for one trial</td>
</tr>
<tr>
<td><code>geom</code></td>
<td>Geometric</td>
<td>( p = ) probability of success for one trial</td>
</tr>
<tr>
<td><code>hyper</code></td>
<td>Hypergeometric</td>
<td>( m = ) # of white balls, ( n = ) # of black balls, ( k = ) # of balls drawn from urn</td>
</tr>
<tr>
<td><code>pois</code></td>
<td>Poisson</td>
<td>( \lambda = ) mean</td>
</tr>
</tbody>
</table>

2. Continuous distributions:

<table>
<thead>
<tr>
<th>R name</th>
<th>distribution</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>unif</code></td>
<td>Uniform</td>
<td>( \text{min} = ) lower limit, ( \text{max} = ) upper limit</td>
</tr>
<tr>
<td><code>norm</code></td>
<td>Normal</td>
<td>( \text{mean}, \text{sd} = ) standard deviation</td>
</tr>
<tr>
<td><code>exp</code></td>
<td>Exponential</td>
<td>( \lambda = ) rate</td>
</tr>
<tr>
<td><code>gamma</code></td>
<td>Gamma</td>
<td>shape; either rate or scale</td>
</tr>
<tr>
<td><code>beta</code></td>
<td>Beta</td>
<td>( \text{shape1}, \text{shape2} )</td>
</tr>
</tbody>
</table>

There are many others: cauchy, chisq, f, t, etc. For each you have the associated distribution (or density) function (d prefix), cumulative distribution function (p prefix), quantile function (q prefix), and the random variable function (r prefix) that you can use to generate random numbers with the given distribution. This naming convention was illustrated above for the normal distribution.

Example: the hypergeometric distribution. Consider a population of \( N = 100 \) people, of whom \( N_f = 52 \) are females and \( N_m = N - N_f = 48 \) are males. From this population we draw at random \( n = 10 \) individuals without replacement and count the number \( X \) of females. Then \( X \) is a random variable with the hypergeometric distribution:

\[
P(X = x) = \binom{N_f}{k} \binom{N_m}{n-x} \binom{N}{n} = \frac{52}{10-x} \binom{48}{10} \binom{100}{10}
\]
Note that \( P(X = x) = \text{dhyper}(x, 52, 48, 10) \).

- To draw a graph of \( f(x) = P(X = x) \):

```r
n = 10 # Number of individuals drawn from the population
Nf = 52 # Number of females in the population
Nm = 48 # Number of males. Therefore N=Nf+Nm=100
x = c(1:n) # Number of females in the sample of n individuals
y = dhyper(x, Nf, Nm, n) # Hypergeometric probabilities for each x
plot(x, y, type='b', xlab='number of females', ylab='probability')
grid()
```

![Graph of \( f(x) = P(X = x) \)](image)

- To draw a graph of the cumulative distribution function \( F(x) = P(X \leq x) \) for this same example, first note that \( F(x) = \text{phyper}(x, 52, 48, 10) \). Then

```r
z = phyper(x, 52, 48, 10)
plot(x, z, type='b', xlab='number of females', ylab='cumulative probability')
grid()
```

This produces the next graph.
• Suppose you'd like to obtain the 0.75 quartile of the distribution. This is simply \( qhyper(0.75, 52, 48, 10) \), which gives the number 6.

• Say you'd like to simulate the experiment of drawing 10 people at random without replacement and count the number of females. We wish to repeat this experiment 1000 times, then produce a histogram of the results. Finally, you'd like to compare the empirical data thus obtained with the exact probability distribution for the number of females.

```r
X=rhyper(1000,52,48,10) #This creates a vector of 1000 independent numbers from 0 to 10 drawn from the hypergeometric distribution.
hist(X,breaks=-0.5+c(0:11),freq=FALSE,ylim=range(c(0,0.3))) #This produces the histogram plot showing relative rather than absolute frequencies.
lines(x,y,type='b') #This adds the plot of dhyper to the histogram.
grid()
```

The resulting graph is shown below. (Note that I increased the range of the y axis for the histogram. Without doing this, the histogram itself would look fine but a small tip at the top of the overlaid graph would get chopped off.)
Density functions of some continuous distributions. The following graphs show the p.d.f.s of some useful random variables that are related to this assignment.

They were obtained in R with the following script:

```r
x=seq(from=0, to=6, length.out=100) #Defines the domain of the density functions
ylim=c(0,0.6)  #Sets the range of the y-axis
par(mfrow=c(2,2)) #Creates a 2x2 plotting area
#Plot of a uniform density:
plot(x,dunif(x,min=2,max=4),main='Uniform',type='l',ylim=ylim)
```
#Plot of a normal density:
plot(x, dnorm(x, mean=3, sd=1), main='Normal', type='l', ylim=ylim)

#Plot of an exponential density:
plot(x, dexp(x, rate=0.5), main='Exponential', type='l', ylim=ylim)

#Plot of a gamma density:
plot(x, dgamma(x, shape=2, rate=1), main='Gamma', type='l', ylim=ylim)

**Relationship between density and probability.** Areas under the density plot indicate probabilities. For example, the shaded area in the graph shown below represents the probability \( P(1 \leq Z \leq 2) \), where \( Z \) is a standard normal random variable. This interpretation of the p.d.f graph is, of course, general and doesn't only apply to normal random variables.

The particular type of graph just shown (with shading) won't be needed in this or any future assignments. But if you are curious about how I got it (I simply copied it from *R Cookbook* by Paul Teetor, O’Reilly 2011) here is the code:

```r
#The density curve of a normal distribution
x=seq(from=-3,to=3,length.out=100) #Points on the x-axis
y=dnorm(x) #Values of the normal density on those x values
plot(x,y,main="Standard Normal Distribution",type='l',ylab="Density",xlab='z')
abline(h=0) #Adds a horizontal straight line at y=0

#We want to shade the region under the graph over the interval [1,2].
region.x=x[1<=x & x<=2]
region.y=y[1<=x & x<=2]
region.x=c(region.x[1],region.x,tail(region.x,1))
region.y=c(0,region.y,0)
polygon(region.x,region.y,density=10)
```

**Exponential random variables.** We consider now in more detail exponential random variables. They are examples of continuous random variables. Recall from our class discussion that an exponential random variable \( T \) with rate \( \lambda \) has p.d.f \( f(t) = \lambda e^{-\lambda t} \). We often think of an exponentially distributed \( T \) as a *random waiting time*. Say that certain events happen in succession at random times \( 0 = T_0 < T_1 < T_2 < \ldots \), and that the time intervals (waiting times) \( T_{i+1} - T_i \) are independent exponential random variables with rate \( \lambda = 1 \). Suppose we wish to simulate these random times and count how many events have occurred over a fixed time interval, say \([0, 10]\). Since we mainly need to generate random numbers, the key command in R is `rexp`. The following program produces such \( T_i \):
tfinal=10 #Length of time interval
lambda=1 #exponential rate
t=0 #Current time; will be update during the course of simulation
T=c() #Empty vector of times of random events
while (t<tfinal){
    t=t+rexp(1,lambda) #This computes the time of the next event
    t=min(t,tfinal) #Makes sure final time not exceeded
    T=c(T,t) #Adds one more entry to vector T
}
Times=T[1:length(T)-1]

The above script creates a vector Times of (random) length N roughly equal to $\lambda t_{\text{final}}$, the first vector component being equal to $T_1 > 0$ and the last $T_N < 10$. The next plot gives the number of events that have occurred up to time $t$, where $0 \leq t \leq 10$. It was produced using stepfun. This R function creates a stepwise function out of two vectors, called Times and Number in this example. The first vector, Times, sets the points of discontinuity on the $x$-axis, and Number, which must have length one unit greater than Times, gives the constant values of the function on the intervals between those breakpoints.

Number=c(0:length(Times))
F=stepfun(Times,Number,f=0) #Creates a step function out of the vectors T and y
plot(F,xlab='Time',xlim=range(c(0,10)),ylab='Number of events',
     main='Total number of events as function of time')
grid()
Problems

1. **Playing darts, I.** In this experiment a person throws darts on a dartboard of radius 1. Each time the dart is thrown, it is assumed to hit at a random point with coordinates \((X, Y)\) centered at the origin (the bull's-eye), where \(X\) and \(Y\) are normal random variables with mean 0 and standard deviations 0.2 and 0.1, respectively. (These numbers indicate the level of skill of the player; big standard deviation is characteristic of a bad player.) The next figure depicts the dart board with a circle of radius 0.3 in dashed lines, and 50 hit marks.

Do the following experiment: simulate 10000 throws of the dart and determine empirically (i.e., simply by counting the number of good hits over total number of throws) the probability that the player can hit within a radius 0.3 from the center. (The code used to produce the above figure is shown next. Although no graph is asked for in this problem, you may find some of this script useful.)

```r
#Draw a circle of radius 1 centered at the origin, in solid line:
  a=seq(from=0,to=2*pi,length.out=100)
  C=cos(a)
  S=sin(a)
  plot(C,S,type='l',asp=1,axes=FALSE,xlab='',ylab='',
       xlim=range(c(-1,1)),ylim=range(c(-1,1)))

par(new=TRUE) #Superimpose the smaller circle to the one just drawn

#Draw another circle, now of radius 0.3, in dashed line
  C1=0.3*cos(a)
  S1=0.3*sin(a)
  plot(C1,S1,type='l',lty='dashed',asp=1,axes=FALSE,xlab='',ylab='',
       xlim=range(c(-1,1)),ylim=range(c(-1,1)))

#Now generate the random points and plot them over the dartboard
  X=rnorm(100,mean=0,sd=0.2)
  Y=rnorm(100,mean=0,sd=0.1)
  par(new=TRUE)
```
2. **Playing darts, II.** Imagine that a truly awful player is playing darts in his room. A dartboard of radius 1 hangs at the center of one wall. The rectangular wall has sides 6 by 8. Now the assumption is that all points on the wall are equally likely to be hit. In other words, the coordinates \(X\) and \(Y\) of a hit are uniformly distributed over the intervals \([0, 6]\) and \([0, 8]\), respectively.

(a) What is the exact probability that this player will hit the dartboard at all? (Do this by hand.)

(b) Obtain by simulation an approximation of the probability asked for in the first part of this problem. Suggestion: simulate 10000 throws and count the fraction that hit the dartboard. How well does your approximation compare with the exact value? (You may like to explore bigger numbers than 10000, if it does not take too long to run in your computer.)

(c) Draw a graphic similar to the above, now representing the rectangular wall in solid line, the dartboard at the center of the wall in dashed line, and 200 hit marks generated from the uniform distribution.

3. **Movement of a (mathematical) water bug, I.** Imagine a little water bug hopping about over the surface of a lake according to the following rule. It starts at time \(T_0 = 0\) at the middle of the lake (coordinates \((0, 0)\)); then at random times \(T_1 < T_2 < \ldots\) it jumps from its current position \((X, Y)\) to a new position \((X + \Delta X, Y + \Delta Y)\). The time intervals \(\Delta T_i = T_{i+1} - T_i\) are independent exponential random variables with rate \(\lambda = 1\), and the \(x\) and \(y\) components of the jumps are independent normal variables of mean 0 and standard deviation 0.1.

(a) Draw (in solid line) a graph of the \(X\)-coordinate of the bug's position as a function of time over the course of 1000 steps.

(b) Draw (without connecting the dots) a graph containing 10000 points, each representing the final point of an independent 100-steps path. This roughly shows the distribution of final positions of the bug's motion. (Here and below, use aspect ratio \(\text{asp}=1\) for a nice graph. As a reference, I suggest that you draw a circle of radius 1 in dashed line centered at the origin. When superimposing two graphs, make sure to set the range of the \(x\) and \(y\) coordinates some that a common scale is set for both. You'll need very small points to better see the result; see the command \(\text{cex}\) I used for the dartboard figure.)

4. **Movement of a water bug, II.** Suppose the same situation described in the previous problem. Assume that the exponential rate parameter is \(\lambda = 10\ \text{min}^{-1}\).

(a) Let \(N\) be the random number of times the bug jumps over the course of 1 minute. Obtain 10000 values of \(N\) by simulation and plot a histogram. Superimpose to your histogram the graph of the probability mass function of a Poisson distribution with mean \(\lambda = 10\). (For a nice graph, let your bins be centered at 0, 1, \ldots, 25.) How good an agreement do you observe?

(b) Obtain 10000 values of the time it takes for the bug to jump 5 times and plot a histogram. Superimpose to it the graph of the probability density function of the Gamma distribution with shape parameter 5 and rate parameter 10. How good an agreement do you observe? (For a nice graph, I suggest 100 bins and the density plot drawn over the interval from 0 to 2.)