Mathematics 375 – Probability Theory
Computer Lab Day 3 – Central Limit Theorem Demo in R
December 5, 2011

Goals

The goals of today's lab are:

- to gain some more practice with basic features of the R statistics package, including some very basic programming with loops, and
- to gain an intuitive understanding of the pattern expressed by the Central Limit Theorem, our final topic for the semester.

Basic R Mechanics – from Lab Day 1

Refer to the assignment sheet for the first lab day if you need to refresh your memory about running R in Haberlin 136, about how to save and print your results, etc. The Lab 1 assignment is posted on the course homepage, so you can open it and look at the instructions there as needed.

Background on Probability Distributions in R – from Lab Day 2

The R package contains built-in functions for computing probability functions for many of the discrete random variables we have discussed, plus functions for the probability densities and cumulative distributions of many of the standard types of continuous random variables that we will study this semester and next. Here's how it works. (Also look at table on 332 of Dalgaard *Introductory Statistics with R.*) Each type of random variable is covered by a *family* of 4 functions distinguished by a *prefix letter* `d, p, q`, or `r`:

- `d` – the probability function in the discrete case, or the density function continuous case,
- `p` – the cumulative distribution
- `q` – the quantile function (essentially the *inverse function* of the cumulative distribution, but of course the cumulative distribution can fail to be injective, so this needs to be taken with a grain of salt)
- `r` – random number generator

Following the prefix letter comes the rest of the function name and the inputs needed to compute the corresponding values. For instance the family of functions for exponential random variables (continuous) looks like this:

- `dexp(y,rate)` – the probability density function

$$f(y) = \begin{cases} rate \cdot e^{-rate \cdot y} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

for $Y \sim Exponential(1/rate)$ (note that the rate is $\frac{1}{\beta}$ in our usual formula for the exponential density).

- `pexp(y,rate)` – the cumulative distribution $P(Y \leq y)$
- `qexp(q,rate)` – computes the y such that $P(Y \leq y) = q$
- `rexp(N,rate)` – generates a list of N random numbers following a $Exponential(1/rate)$ distribution

Today's Lab – Distributions of Sample Means

Work through the following example before and then follow the pattern here in the other parts of question A below. below. You do not need to hand in any work for this preliminary discussion, just for the parts of the question itself.

1. Begin by generating a list of 100 random numbers from the exponential distribution with $\beta = 3$ (in `R` terms, $rate = 1/3$), and assign the result to the name `x`. We will think of this as a *sample* of values from that distribution.

2. Plot the density histogram for this data set `x` using the command

```
hist(x,freq=FALSE)
```

   This should look roughly like the part of the graph of the exponential density for positive y, reflecting the fact that the random numbers are generated according to probabilities given by that density function.

3. Our main goal is to understand what happens if we generate *lots* of such samples, find their sample means, then consider *the distributions of those sample means.*

4. Here is one way to do this in `R` using a simple *for loop.* First we create space to store the means of 1000 different random samples:

```
means <- array(1:1000)
```

   Next, we generate 1000 different random samples, find the means of each of them, and store them in the array created above:

```
for (i in 1:1000)
   {
   x <- rexp(100,1/3)
   means[i] <- mean(x)
   }
```

(Note: This could all be entered as one input line provided you leave spaces in the appropriate places. However, I think it is more readable (and it is easier to identify typos if you happen to make one) if you press ENTER at the end of each of the lines as you are typing this. If you do it that way, you should note that `R` generates a new input prompt `+` each time, indicating that you are still in the body of the `for` loop. The final } will close off the loop and take you back to the > prompt.)

5. What does the distribution of the sample means look like? We can see that by generating a density histogram for the data in the `means` array:

```
hist(means,breaks=20,freq=FALSE)
```

   This should look completely different from the histogram showing the distribution of the single sample above. In fact, what does this histogram remind you of?

6. If you said, "a normal density graph," good!! If you said something different, go back and look again.

7. Now, which normal density is it? Well, notice by our results about expected values and variances of linear combinations, if we have 100 sampled values $Y_1, \ldots, Y_{100}$ that are each exponentially distributed with $\beta = 3$, then the sample mean

$$\overline{Y} = \frac{1}{100}(Y_1 + \cdots + Y_{100})$$

   will have

$$E(\overline{Y}) = 100 \cdot \frac{\beta}{100} = \beta = 3$$

   and

$$V(\overline{Y}) = 100 \cdot \frac{\beta}{100^2} = \frac{\beta}{100} = .3.$$

8. Let's overlay that normal density and see how well things match:

```
curve(dnorm(x,3,.3),add=TRUE)
```

   This should match quite well (not perfectly, though!)

Lab Question

Repeat the parts of the worked example below for the means of 1000 random samples of size 100 from each of the following distributions. Note: In each case to find the appropriate normal density curve, you will need to recall facts about expected values and variances of the Y_i individually, then combine that with the patterns for linear combinations like the one computing $\overline{Y}$.

1. A uniform distribution with $min = 4$ and $max = 10$. (These are the endpoints of the interval where the uniform density different from zero.)

2. A Gamma distribution with $\alpha = 3$ and $\beta = 4$. (Note: the `rate` parameter in `R` is also $rate = 1/\beta$ here.)

3. A Poisson distribution with $\lambda = 2$.

What is happening in each case? Is the distribution of the sample means approximately normal every time?

Assignment

Lab reports containing input, output, and answers to the question posed above are due no later than Monday, December 12. If you do not finish during the hour today, you can return to HA 136 any time it is not in use by another class.