**Lab Assignment 04**

I. Objectives: To explore probability distributions.

II. Datasets: Body Measurements:

http://www.openintro.org/stat/data/bdims.RData"

III. Packages: none

IV. Preparation

1) Open RStudio on a lab terminal by double-clicking the icon or selecting RStudio from the Start menu in Windows

2) Clear data from memory by typing: > rm(list=ls())

V. Instructions for Lab 04

A) Create a variable “x” with the values -2, -1, 0, 1, and 2.

>x <- c(-2,-1,0,1,2)

>x

B) Calculate the value of the cumulative distribution function at (or the probability to the left of) a given value of x. Record these probabilities.

>pnorm(x)

C) Now create a variable “y” with values: 0,1,2,5,8,10,15,20.

> y <- c(0,1,2,5,8,10,15,20)

D) Using a binomial distribution, calculate the value of the cumulative distribution function at a given value of y. Record these probabilities.

> pbinom(x,size=20,prob=.2)

E) Repeat step D, but use a Poisson distribution. Record these probabilities.

> ppois(x,6)

2)  Calculate the following probabilities :

A) Probability that a normal random variable with mean 22 and variance 25

(i)

lies between 16.2 and 27.5   
pnorm(27.5,22,sd=5)-pnorm(16.2,22,sd=5) 

(ii)

is greater than 29

1-pnorm(29,22,sd=5) 

(iii)

is less than 17

pnorm(17,22,sd=5) 

(iv)

is less than 15 or greater than 25

pnorm(15,22,sd=5)+1-pnorm(25,22,sd=5) 

B.

Probability that in 60 tosses of a fair coin the head comes up

(i)

20,25 or 30 times   
sum(dbinom(c(20,25,30),60,prob=0.5)) 

(ii)

less than 20 times   
pbinom(19,60,prob=0.5) 

(iii)

between 20 and 30 times pbinom(30,60,prob=0.5)-pbinom(20,60,prob=0.5) 

C.

A random variable X has Poisson distribution with mean 7. Find the probability that

(i)

X is less than 5    
  
> ppois(4,7)   
[1]0.1729916

(ii)

X is greater than 10 (strictly)   
> 1-ppois(10,7)

(iii)

X is between 4 and 16

> ppois(16,7)-ppois(3,7)

3. The following examples illustrate how to generate random samples from some of the well-known probability distributions.

A) The first sample is from N(0,1) distribution and the next one from N(5,1) distribution.

> z <- rnorm(10)

> z

> w <- rnorm(1000,mean=5,sd=1)

> hist(w)

B) Binomial Distribution:

> k <- rbinom(20,size=5,prob=.2)

> k

C) Poisson($\lambda$) Distribution :

> x <- rpois(20,6)

> x

4) This week we’ll be working with measurements of body dimensions. This data set contains measurements from 247 men and 260 women, most of whom were considered healthy young adults.

load(url("http://www.openintro.org/stat/data/bdims.RData"))

Let’s take a quick peek at the first few rows of the data.

head(bdims)

You’ll see that for every observation we have 25 measurements, many of which are either diameters or girths. A key to the variable names can be found [here](http://www.openintro.org/stat/data/bdims.php) (http://www.openintro.org/stat/data/bdims.php), but we’ll be focusing on just three columns to get started: weight in kg wgt, height in cm hgt, and sex (1 indicates male, 0 indicates female).

Since males and females tend to have different body dimensions, it will be useful to create two additional data sets: one with only men and another with only women.

mdims = subset(bdims, bdims$sex == 1)

fdims = subset(bdims, bdims$sex == 0)

Make a histogram of men’s heights and a histogram of women’s heights. How would you compare the various aspects of the two distributions?

hist(mdims$hgt)

hist(fdims$hgt)

p1 = hist(mdims$hgt)

p2 = hist(fdims$hgt)

xlim = c(min(fdims$hgt), max(mdims$hgt)) + c(-5, 5)

plot( p1, col=rgb(0,0,1,1/4), xlim=xlim)

plot( p2, col=rgb(1,0,0,1/4), add=T)

In your description of the distributions, did you use words like bell-shaped or normal? It’s tempting to say so when faced with a unimodal symmetric distribution.

To see how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data. We’ll be working with women’s heights, so let’s store them as a separate object and then calculate some statistics that will be referenced later.

fhgtmean = mean(fdims$hgt)

fhgtsd = sd(fdims$hgt)

Next we make a density histogram to use as the backdrop and use the lines function to overlay a normal probability curve. The difference between a frequency histogram and a density histogram is that while in a frequency histogram the heights of the bars add up to the total number of observations, in a density histogram the areas of the bars add up to 1. The area of each bar can be calculated as simply the height  times the width of the bar. Using a density histogram allows us to properly overlay a normal distribution curve over the histogram since the curve is a normal probability density function. Frequency and density histograms both display the same exact shape; they only differ in their y-axis. You can verify this by comparing the frequency histogram you constructed earlier and the density histogram created by the commands below.

hist(fdims$hgt, probability = TRUE, ylim=c(0, 0.07))

x = 140:190

y = dnorm(x = x, mean = fhgtmean, sd = fhgtsd)

lines(x = x, y = y, col = "blue")

Based on the this plot, does it appear that the data follow a nearly normal distribution?

Eyeballing the shape of the histogram is one way to determine if the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve. An alternative approach involves constructing a normal probability plot, also called a normal Q-Q plot for quantile-quantile.

qqnorm(fdims$hgt, col="orange", pch=19)

qqline(fdims$hgt, lwd=2)

A data set that is nearly normal will result in a probability plot where the points closely follow the line. Any deviations from normality leads to deviations of these points from the line. The plot for female heights shows points that tend to follow the line but with some errant points towards the tails. We’re left with the same problem that we encountered with the histogram above: how close is close enough?

A useful way to address this question is to rephrase it as: what do probability plots look like for data that I know came from a normal distribution? We can answer this by simulating data from a normal distribution using rnorm.

sim.norm = rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)

Make a normal probability plot of sim.norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

qqnorm(sim.norm, col="orange", pch=19)

qqline(sim.norm, lwd=2)

Even better than comparing the original plot to a single plot generated from a normal distribution is to compare it to many more plots using the following function. It may be helpful to click the “zoom” button in the plot window.

qqnormsim(fdims$hgt)

Does the normal probability plot for fdims$hlstd look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

Using the same technique, determine whether or not female weights appear to come from a normal distribution.

wgtm = mean(fdims$wgt)

wgts = sd(fdims$wgt)

xlim = c(min(fdims$wgt), max(fdims$wgt)) + c(-5, +5)

hist(fdims$wgt, probability=T, xlim=xlim)

x = seq(xlim[1], xlim[2], 0.5)

y = dnorm(x=x, mean=wgtm, sd=wgts)

lines(x=x, y=y, col="blue")

qqnorm(fdims$wgt, col="orange", pch=19)

qqline(fdims$wgt, lwd=2)

sim.norm = rnorm(n=length(fdims$wgt), mean=wgtm, sd=wgts)

qqnorm(sim.norm, col="orange", pch=19)

qqline(sim.norm, lwd=2)

qqnormsim(fdims$wgt)