Exercises – R Course

Fall 2013

Day 2: Probability distributions

1. Generate 10 Bernoulli trials with probability of success equal to 0.35. Store them in a data object.

rbinom(10, 1, .35)

ber.35 <- rbinom(10, 1, .35)

1. Set the seed to today’s date (in MMDDYY format) and generate 10 random numbers from a uniform distribution over the interval [2, 3.5]. Generate 10 random numbers from an exponential distribution with rate 2. Store each of them in a data object.

set.seed(052213)

runi10 <- runif(10, 2, 3.5)

rexp10 <- rexp(10, 2)

1. Using the data set **cars** (already loaded in R), display the first 10 records of the data set. Display summaries of each of the variables in the data set, using the **summary()** and **fivenum()** functions. How do the results compare?

cars[1:10,]

summary(cars)

fivenum(cars$speed)

fivenum(cars$dist)

1. Construct a stem and leaf plot of both of the variables in the **cars** data set.

stem(cars$speed)

stem(cars$dist)

1. Construct a histogram of the frequency of speed variable in **cars**, using 4 different values for the numbers of bins. Give appropriate labels to both the x and y axes and give a title to the graph. How does the shape of the histogram change as the number of bins increase?

par(mfrow=c(1,4))

hist(cars$speed, breaks = 3, main = "Histogram of the Speed of Cars", xlab = "Speed")

hist(cars$speed, breaks = 5, main = "Histogram of the Speed of Cars", xlab = "Speed")

hist(cars$speed, breaks = 7, main = "Histogram of the Speed of Cars", xlab = "Speed")

hist(cars$speed, breaks = 9, main = "Histogram of the Speed of Cars", xlab = "Speed")

1. Using the dist variable in **cars**, construct a histogram of the probability with 6 bins, and overlay a dotted blue line for the density.

hist(cars$dist, breaks = 5, prob = T, main = "Histogram of the Distance of Cars", xlab = "Distance", ylab = "Probability")

lines(density(cars$dist), col="blue")

1. Construct boxplots for both variables of the **cars** data set, label the boxes, give a title to the graph, use different colors for each variable, and restrict the range of the plot to the minimum value of the two variables and the maximum value of the two variables in the data set.

boxplot(cars$dist, cars$speed, names = c("Distance", "Speed"), col = c("blue", "green3"), main = "Boxplots of Distance and Speed", ylim=c(min(cars$dist, cars$speed), max(cars$dist, cars$speed)))

1. Construct a qq-plot of the two variables you constructed in (2) above. Set the range of the x and y axes to [0, 3.5]. Give a title to the graph as well as label both the x and y axes. In this instance, how can you verify that your axes are labeled correctly by only looking at the plotted data?

qqplot(rexp10, runi10, main="Comparing Uniform and Exponential Random Variables", xlim=c(0, 3.5), ylim=c(0, 3.5), xlab="Exponential", ylab="Uniform")

1. Display a table for both variables in the **cars** data set. Is this an effective way to summarize these variables? If not, what would you use and why?

table(cars$dist)

table(cars$speed)