Homework 2

Dr. Purna Gamage

9/19/2022

Problem 1 (30 Points)

Consider the random variable defined by counting the number of failures until the first success, for independent trials with success probability p. Given p between 0 and 1, the **R** commands `myattempts(p)` and `rgeom(1,p)` both simulate this random variable. Find a way of demonstrating that the two commands indeed give the same results, for five different values of p. *You may find the table() function useful.*

Choose five p's of your choice, e.g. $p \in \{.1, .3, .5, .7, .9\}$. Run `myattempts` and `rgeom` each 10000 times.

a. For the first p, store the fraction of outcomes in the simulation in the columns of a suitable data frame and compare (Hint: use 3 columns to compare `rgeom()`, `myattempts()` and `dgeom()`) (10 points)

b. For the second p, Compare distribution by computing statistics such as mean and standard deviation. (5 points)

c. For the third p, plot both distributions as histograms in the same plot. (5 points)

d. For the fourth p, make side-by-side box plots. (5 points)

e. For the fifth p, plot the two empirical distribution functions in the same plot. (5 points)

Problem 2 (10 Points)

Consider the following random experiment: draw a uniformly distributed random number X_1 from the interval $(0,1)$. Next, draw a uniformly distributed random number X_2 from the interval $(0, 1+X_1)$, a uniformly distributed random number X_3 from the interval $(0, 1+X_2)$ and so on until X_{20}.

Use a monte carlo simulation to give an approximate answer to What is the mean value of X_{20}? and use a histogram to identify the distribution of X_{20}.

$X_1 \sim unif(0,1)$ $X_2 \sim unif(0, 1+X_1)$ $X_3 \sim unif(0, 1+X_2)$ . . . $X_{20} \sim unif(0, 1+X_{19})$

The **R** command for drawing a uniformly distributed random number from the interval $(0, b)$ is *runif(1,min = 0, max = b).*

Problem 3 (15 Points)

Suppose that the daily power consumption of a major city, X, has a Gamma distribution with shape parameter $r = 4$ and scale parameter $\rho = 2$. Use **R** to compute the following quantities: \

a. $Prob(X \leq 12)$ (3 Points)
b. $Prob(X > 5)$ (3 Points)
c. $Prob(|X - 8|) < 1$ (4.5 Points)

d. z such that $Prob(X < z) = .95$ (4.5 Points)

(Hint:$|X - 8| < 1$ is equivalent to $7 < X < 9$)

Problem 4 (15 Points)

Probability theory says that a binomial distribution, $B(n, p)$ is close to that of a normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$, if np and $n(1-p)$ are both sufficiently large, e.g. at least 10.

Check this by plotting both cumulative distribution functions in the same figure, using a staircase plot for the binomial distribution and a line plot for the normal distribution, for three different cases: a case where both np and $n(1-p)$ are large, a case where np is large and $n(1-p) < 10$, and a case where $np < 10$ and $n(1-p) < 10$.

Describe what happens in all three cases. In what sense are the cdf's not close in cases 2 and 3? (Hint: Compare with a cdf of the normal distribution)

(Hint: You can specify a staircase plot using the `lines()` function and passing the optional parameter `type=s`)

Problem 5 (15 Points)

A graphical technique for checking whether a sample has an approximate normal distribution is a “quantile-quantile” plot. The **R** command is `qqnorm(x)`, where x is the vector of sample values. If the plot is approximately a straight line, then this suggests that the sample comes from a normal distribution. Use the dataset from the `openintro` package to find out which of the the four distributions (three exams and course grade) is the closest to a normal distribution? You can load the dataset with the following commands `library(openintro)` and `data(exam_grades)`. Explore this by making `qqnorm()` and `qqline()` plots of the four distributions How close to straight lines are the plots in each case? How do the plots differ from straight lines? Hint: Make sure to remove the NA values.

Problem 6 (15 Points)

If X has a continuous distribution with cumulative distribution function F, then the new random variable $U = F(X)$ has a uniform $U(0, 1)$ distribution. Verify this with simulations for three different continuous distributions of your choice, by making a random sample of sufficient size, sorting it, plugging it into the cdf F, and plotting the result.

Bonus Problem (20 Points)

Suppose for $X = X_1 + X_2$ is the sum of two exponentially distributed random variables with the same parameter λ. Then X^α is very nearly normally distributed for a suitable choice of α. Determine an approximate value for α (within 0.05), using a simulation and `qqnorm()` plots for each of your choices of α.