LAB 2 ASSIGNMENT

PROBABILITY DISTRIBUTIONS

In this lab assignment, you will initially use numerical and graphical tools available in Excel to analyze a situation described by a Poisson distribution. You will use the process to further explore the properties of discrete and continuous distributions through computer experiment and simulation. We have used some functions available in Excel to develop three instructional templates (interactive worksheets) that allow you to utilize the properties of the Poisson, binomial, and negative binomial distributions. Some cells and charts in the interactive worksheets are protected and are, therefore, read-only.

The CAPTCHA Test

A CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a type of challenge-response system designed to differentiate humans from robotic computer programs. CAPTCHAs are used as security checks to deter spammers and hackers from using forms on web pages to insert malicious or frivolous code. Quite simply, CAPTCHA works by asking end users to perform some task that a software bot cannot do. If the user can do the task correctly, it provides authentication to the service that the user is a human being and not a spambot and allows the user to continue.

Tests often involve JPEG or GIF images because while bots can identify the existence of an image by reading source code, they cannot tell what the image depicts. Because some CAPTCHA images are difficult to interpret, human users are usually given the option to request a new CAPTCHA test. Requesting a new test is fairly rare, so these "request" events could be modelled by a Poisson distribution with a parameter of λ.

To evaluate the number of requests, suppose an internet website reviews their website data such that there exists a reasonably large amount of new test requests. Random samples of users who try to access the website only once a day are taken regularly and the data provided is analyzed against the target parameters. The internet website's employees review the data for any new test requests. The number of new test requests by these users in a single day is compared to a Poisson distribution with a parameter of λ, the average number of new test requests by these users in a single day. We will use Excel to simulate the outcomes of the new test requests for random users who try to access the website only once a day, or "once daily access" (ODA) users.

The necessary Excel file is available in the Data link located in the Lab 2 tab display in the Labs section on eClass. The data are not to be printed in your submission. The file consists of four worksheets: *Poisson Probabilities*, *Binomial Prob.*, *Neg. Bin. Prob.*, and *Simulation*. The first three are instructional templates; the fourth one contains data related to the above problem. **Unless otherwise stated, numerical answers should be to four decimal places.**

1. Open the *Poisson Probabilities* worksheet. The worksheet contains a graph of the Poisson probability distribution function. The top of the worksheet contains the value of λ. Any change to the distribution parameter is immediately reflected in the probability distribution function. In this part, you will see how various choices of λ. affect the shape of the Poisson distribution.

   Enter the value of λ as 8, then 4, 2, 1.5, 1, and finally 0.25. After each entry, carefully examine the shape of the corresponding density curve. You are not supposed to print the probability distribution functions.

   Describe briefly the change in the appearance of the probability distribution function as λ decreases from 4 to 0.25. What does the decrease mean to new test requests by ODA users in a single day?

2. In this question, you can use either the built-in Excel functions or the instructional templates to help answer the questions. Suppose that the number of new test requests by ODA users in a single day follows a Poisson distribution with parameter λ = 2.5. This implies that the expected number of new test requests by ODA users in a single day is 2.5. Answer the following questions. **In parts (c) to (e), also give the parameters of the distributions used.**

(a) What is the probability that there are no new test requests by random ODA users in a randomly selected day? In other words, what is the proportion of days with no new test requests?

(b) Suppose the CAPTCHA test is investigated if six or more new test requests are detected for random ODA users in a day. Use the template to determine the proportion of days that the CAPTCHA test is investigated.

(c) Refer to part (b). What is the probability that the CAPTCHA test is not investigated for a randomly selected period of four consecutive weeks? (Assume the first day of the entire period is selected randomly and each day is independent of the other days.)

(d) Refer to part (b). What is the probability that the CAPTCHA test is investigated for more than two days in a randomly selected period of four consecutive weeks? (Again, assume the first day of the entire period is selected randomly and independence between days.)

(e) Refer to part (b). What are the expected value and standard deviation for the number of days until the CAPTCHA test is investigated? (Again, assume independence.)

3. In this question, you can use either the built-in Excel functions or the instructional templates to help answer the questions. **In parts (a) and (b), also give the parameters of the distributions used.**

(a) Suppose again that days are independent but now the number of new test requests by ODA users in a single day follows a Poisson distribution with parameter $\lambda = 3.5$. What is the probability that the fifth day is the first one with no new test requests for random ODA users?

(b) If the website company has already identified that random ODA users have gone three days with at least one new test request in each day, what is the probability that the eighth day is the first one with no new test request? Compare your results with probability obtained in part (a). Give a justification for your results.

In Questions 4 and 5, you must use Excel in a Windows environment to obtain data (**different data may be produced by the *Random Number Generator* in Excel on MAC OS**). Excel 2016 or a newer version should be used in this part (older versions may produce different sequences of random numbers for the same seed).

Now suppose that the internet website's employees take a random sample of days to estimate the average number of new test requests. How likely will the estimate be close to the true value? Which estimate is more likely to be better: one based on a sample of 20 days or one based on a sample of 50 days? We will answer the above questions by simulation.

4. Open the worksheet *Simulation*. The worksheet allows you to simulate the outcomes of new test requests for 60 random samples of days. Use the *Random Number Generation* feature (Poisson, seed 10) to generate 60 samples of size $n = 20$ from the Poisson distribution with $\lambda = 2.5$. This corresponds to selecting 60 samples, each consisting of 20 days. The data will be entered in the form of 60 columns, each consisting of 20 rows into the range B10:BI29. In other words, the range contains the outcome of quality testing for 1200 days.

   Once the data are entered, the values of the variables AVERAGE and COUNT are automatically displayed in rows 61 and 63, respectively. They are further explained below, as needed.

   (a) Use the COUNTIF function to determine the number and proportion of days with no new test requests among the 1200 days. Compare the value with the probability obtained in Question 2, part (a). Should the values be identical? Explain briefly. The COUNTIF function was discussed in the *Lab 2 Instructions*.

   (b) The variable COUNT in row 63 counts the number of days with at least 6 new test requests in each sample. Use the values to determine the number and proportion of samples of 20 days containing less than 6 new test requests. Compare the value with the probability obtained in Question 2, part (b). Should the values be identical?

(c) The variable AVERAGE in row 61 shows the average number of new test requests for each sample. Obtain and print a histogram of the average number of new test requests using the following bins: 1.25, 1.50, ..., 3.75. The format of your histogram should be the same as the format from previous labs and the *Lab 1 Instructions*. Describe the shape of the histogram.

5. Now we repeat Question 4 with $n = 50$. First, clear the range B10:BI29 in the worksheet *Simulation*. Use the *Random Number Generation* feature (Poisson, seed 10) to generate 60 samples of size $n = 50$ from the Poisson distribution with the parameter $\lambda = 2.5$. The data will be entered into the range B10:BI59 in the form of 60 columns, each consisting of 50 rows.

   (a) Use the COUNTIF function to determine the number and proportion of days with no new test requests among the 3,000 days. Compare the value with the probability obtained in Question 2, part (a). Should the values be identical? Explain briefly. Also compare to the value in Question 4, part (a).

   (b) The variable COUNT in row 63 counts the number of days with at least 6 new test requests in each sample. Use the values to determine the number and proportion of samples of 50 days containing less than 6 new test requests. Compare the value with the probability obtained in Question 2, part (b). Should the values be identical? Also compare these results to those in Question 4, part (b).

   (c) Obtain a histogram of the average number of new test requests using the following bins: 1.25, 1.50, ..., 3.75. Describe the shape of the histogram, comparing to the histogram obtained in Question 4, part (c).

LAB 2 ASSIGNMENT MARKING SCHEMA

Question 1 (4)

Change in the appearance of the probability distribution function: 2 points
Implication to new test requests by ODA users in a single day: 2 points

Question 2 (16)

(a) Proportion of days with no new test requests: 3 points
(b) Proportion of days that the CAPTCHA test is investigated: 3 points
(c) Probability that the CAPTCHA test is not investigated for provided period: 3 points
(d) Probability that the CAPTCHA test is investigated for more than 2 days: 4 points
(e) Expected value and standard deviation: 1.5 points each (3 points total)

Question 3 (10)

(a) Parameters of the distribution(s): 2 points
Final probability: 3 points
(b) Parameters of the distribution(s): 2 points
Final probability: 3 points
Compare results in parts (a) and (b) and give justification of results: 2 points

Question 4 (16)

(a) Number and proportion of days with no new test requests: 2 points
Comparison with the probability: 2 points
(b) Number and proportion of samples of 20 days containing less than six new test requests: 2 points
Comparison with the probability: 2 points
(c) Correctly formatted histogram of the average number of new test requests: 6 points
Shape of the histogram (modality, skewness): 2 points

Question 5 (21)

(a) Number and proportion of days with no new test requests: 2 points
Comparison with the probability: 2 points
Comparison with Question 4: 2 points
(b) Number and proportion of samples of 50 days containing less than six new test requests: 2 points
Comparison with the probability: 2 points
Comparison with corresponding values in Question 4: 2 points
(c) Correctly formatted histogram of the average number of new test requests: 6 points
Shape of the histogram (modality, skewness) and comparison: 3 points

TOTAL = 67