SOLUTIONS TO THE LAB 2 ASSIGNMENT

Question 1

As the parameter λ changes, the probability distribution function changes as well. The probability distribution function is slightly to moderately skewed to the right. The smaller the value of λ, the more extreme the right-skewness.

As λ decreases from 8 to 0.25, however, the shape of the probability distribution function becomes less and less symmetric. Note that even larger values of λ (even further beyond 8) improves the symmetry even further; unfortunately, as the range on the horizontal axis is fixed, the corresponding histogram cannot be displayed fully as lambda continues to increase.

Since the parameter λ shows the expected number of new test requests by ODA users in a single day, a large λ indicates low CAPTCHA test success with several new tests requests in a day. However, as the number λ decreases, there are less and less new test requests, which is less stress on the website's servers.

Question 2

(a) To obtain the required information, we enter the value of the parameter λ as 2.5 and the value of x as 0. The probability that there are no new test requests by random ODA users in a randomly selected day is 0.0821. In other words, the proportion of days with no new test requests is 0.0821.

(b) The CAPTCHA test is investigated if six or more new test requests are detected for random ODA users in a day. Note that $P(X \geq 6) = 1 - P(X \leq 5)$.

The cumulative probability $P(X \leq 5)$ returned by the template is 0.9580. Thus, $P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.9580 = 0.0420$.

Thus, the proportion of days that the CAPTCHA test is investigated is 0.0420.

(c) Let Y be the number of days the CAPTCHA test is investigated. Since a week consists of 7 days, there are $7 \times 4 = 28$ days in any four consecutive weeks. As the days are independent, then Y follows a binomial distribution with $n = 28$ and $p = 0.0420$.

Thus, using the Lab 2 Excel file or Excel functions, $P(Y = 0) = 0.9580^{28} = 0.3008$.

(Note that using the unrounded value from (b) also gives 0.3006. It is also possible to do this question such that $N \sim B(n = 28, p = 0.9580)$ and finding $P(N = 28)$, where N be the number of days the CAPTCHA test is not investigated.)

(d) Let W be the number of days the CAPTCHA test is investigated. Since the days are independent, then W follows a binomial distribution with $n = 28$ and $p = 0.0420$.

Thus, using the Lab 2 Excel file or Excel functions, $P(W \geq 3) = P(W > 2) = 0.1115$.

(Note that using the unrounded value from (b) gives 0.1116.)

(e) Let D be the number of days until the CAPTCHA test is investigated. Since the days are independent, then D follows a geometric distribution with $p = 0.0420$. Thus, the expected value and standard deviation are the following values.

$$E(D) = \frac{1}{p} = \frac{1}{0.0420} = 23.8095,\ SD(D) = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.0420}{0.0420^2}} = 23.3042$$

(Note that using the unrounded value from (b) respectively gives 23.7976 and 23.2922.)

Question 3

(a) First use the *Poisson* worksheet. If X is the number of new test requests by ODA users in a single day, then $X \sim Poisson(\lambda = 3.5)$. Then, $P(X = 0) = 0.0302$. If Y is the number of days until the first one with no new test requests, then $Y \sim Geometric(p = 0.0302)$ or Y follows a negative binomial distribution with $r = 1$ and $p = 0.0302$. Using the *Neg. Bin.* worksheet, then the answer would be $P(Y = 5) = 0.0267$.

(b) As with part (a), X is the number of new test requests by ODA users in a single day and follows a Poisson distribution with $\lambda = 3.5$. Also, Y is the number of days until the first one with no new test requests such that Y is either geometric with $p = 0.0302$ or negative binomial with $r = 1$ and $p = 0.0302$. Since the three days are "given" information, the probability can be expressed and answered as follows.

$$P(Y = 8 \mid Y > 3) = P(Y = 3 + 5 \mid Y > 3) = P(Y = 5) = (1 - 0.0302)^4(0.0302) = 0.0267$$

The answer is identical to the answer in part (a) due to the lack of memory property of the geometric distribution.

Question 4

The worksheet *Simulation* is not protected. The students should be careful not to remove the formulas entered in the rows AVERAGE, COUNT, and the summary statistics for AVERAGE.

(a) The number of days with no new test requests is equal to the number of cells containing zero in the range B10:BI29. Thus, to determine the proportion of days with no new test requests among the 1200 days, use the following formula.

=COUNTIF(B10:BI29,"0")/1200

The value returned by Excel is 107/1200 = 0.0892. The observed proportion (0.0892) is very close to the proportion (0.0821) calculated in Question 2, part (a). The two values, although extremely close, do not have to be identical. Note the probability of 0.0821 refers to the relative frequency obtained in an infinite sequence of days from the production run.

(b) The variable COUNT in row 63 counts the number of days with at least six new test requests in each sample. Samples containing less than six new test requests produce a value of COUNT equal to zero. Thus, to obtain the percentage of samples of 20 days containing less than six new test requests, use the following formula

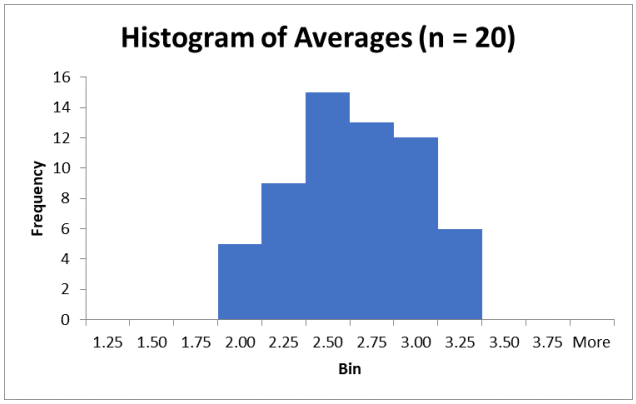
=COUNTIF(B63:BI63,"0")/60

The value returned by Excel is 25/60 = 0.4167.

Since each of the 20 days is independent of the others, the value should be compared to $0.9580^{20} = 0.4239$ (or 0.4238 if using the unrounded value from Question 2, part (b)).

The relative frequency of 0.4167 is fairly close to the value of 0.4239, given that this is a simulation of 60 samples.

(c) The histogram of the average number of new test requests is displayed below.



The histogram appears unimodal and approximately symmetric (like the normal distribution). It is worth noting that the parent distribution (Poisson with $\lambda = 2.5$) is moderately right-skewed, so the above histogram appears suitable.

Question 5

The worksheet *Simulation* is not protected. The students should be careful not to remove the formulas entered in the rows AVERAGE, COUNT, and the summary statistics for AVERAGE.

(a) The number of days with no new test requests is equal to the number of cells containing zero in the range B10:BI59. Thus, to determine the proportion of days with no new test requests among the 3,000 days, use the following formula.

=COUNTIF(B10:BI59,"0")/3000

The value returned by Excel is 240/3000 = 0.0800. The observed proportion (0.0800) is extremely close to the proportion (0.0821) calculated in Question 2, part (a). The two values, although close, do not have to be identical. Note the probability of 0.0821 refers to the relative frequency obtained in an infinite sequence of days from the production run.

The value of 0.0800 is a closer/better estimate than the 0.0892 from Question 4. This is appropriate considering the larger number of observations.

(b) The variable COUNT in row 63 counts the number of days with at least six new test requests in each sample. Samples containing less than six new test requests produce a value of COUNT equal to zero. Thus, to obtain the proportion of samples of 50 days containing less than six new test requests, use the following formula

=COUNTIF(B63:BI63,"0")/60

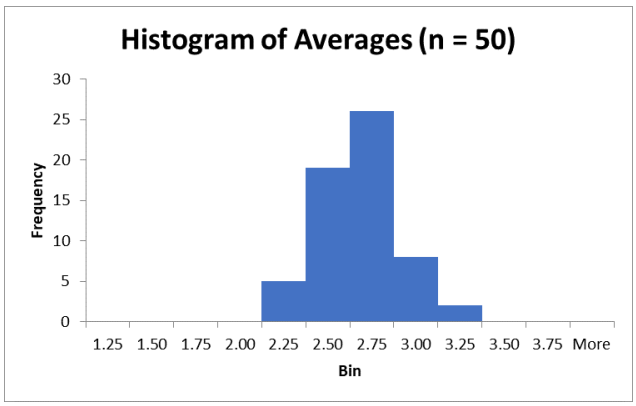
The value returned by Excel is 6/60 = 0.1000.

Since each of the 50 days is independent of the others, the value should be compared to $0.9580^{50} = 0.1170$ (or 0.1169 if using the unrounded value from Question 2, part (b)).

The relative frequency of 0.1000 is legitimately close to the value of 0.1170, given that this is a simulation of 60 samples.

The corresponding comparison in Question 4 shows closer values than the pair seen here. Theoretically, the larger sample size should provide the closer values, but the further distance is possible with random simulations.

(c) The histogram of the average number of new test requests is displayed below.



The histogram of averages for $n = 50$ appears unimodal and approximately symmetric or slightly right-skewed. Compared to the histogram in Question 4, there is more right-skewness present (the previous one was more symmetric), but this is likely because the original population with a Poisson distribution with $\lambda = 2.5$ was moderately right-skewed. The range of the histogram is reduced, though.

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: 2 points
(b) Parameters of the distribution(s): 2 points
Final probability: 2 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