Statistical Inference

HW Author: Mahshid Dashtianeh, Reyhane Vahedi
Instructor: Mohammadreza A. Dehaqani

Spring 2024

HW 2: Part 1

- Please note that this is half a homework. Given sooner with simpler questions, so you have time to learn as the course progresses.
- If you have any questions about the homework, don't hesitate to use the class group to ask questions or drop an email to the HW Authors.
- Please consult the course page for important information on submission guidelines and delay policies to ensure your homework is turned in correctly and on time.
- This course aims to equip you with the skills to tackle all problems in this domain and encourages you to engage in independent research. Utilize your learnings to extend beyond the classroom teachings where necessary.

Section 1: How Large is Large Enough?

Color blindness appears in 2% of the people in a certain population. How large must a random sample be in order to be 99% certain that a color-blind person is included in the sample?

Section 2: Bonus Problem: Other Limit Theorems

In addition to limit theorems that deal with sums, there are limit theorems that deal with extreme values such as maxima or minima. Here is an example: Let $U_1, \ldots, U_n$ be independent uniform random variables on $[0, 1]$, and let $U_{(n)}$ be the maximum. Find the cumulative distribution function or CDF of $U_{(n)}$ and a standardized $U_{(n)}$, and show that the CDF of the standardized variable tends to a limiting value.

Section 3: Sum of RVs and Limit Theorem

A factory builds X_n products on the nth day of production where all the X_n amounts are independent of each other and have similar distributions with a mean of 5 and variance of 9.

(a) Find the approximate probability of less than 240 products being made in the first 100 days.

(b) Find the largest n that fits in the following phrase:

$$P(X_1 + X_2 + \ldots + X_n \geq 200 + 5n) \leq 0.05$$

(c) If the Nth day is the first day where the number of total produced products passes 1000, find the approximate probability of $N \geq 220$.

Section 4: Convergence of Probabilites

Consider n rolls of a balanced dice. Let X_i be the outcome of the i-th roll, and let $S_n = \sum_{i=1}^{n} X_i$. Show that, for any $\epsilon > 0$:

$$P\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$.

Section 5: Error of Rounding Numbers

Numbers in decimal form are often approximated by the closest integers. Suppose n numbers $X_1, \ldots, X_n$ are approximated by their closest integers $J_1, J_2, \ldots, J_n$. Let $U_i = X_i - J_i$. Assume that U_i are uniform on $(-0.5, 0.5)$ and that U_i's are independent.

(a) Show that $\frac{\sum_{i=1}^{n} U_i}{\sqrt{\frac{n}{12}}} \sim N(0,1)$ as $n \to \infty$.

(b) Find $P\left(\frac{-5}{\sqrt{\frac{300}{12}}} \leq \frac{\sum_{i=1}^{n} U_i}{\sqrt{\frac{300}{12}}} \leq \frac{5}{\sqrt{\frac{300}{12}}}\right)$.

Section 6: Large Number of Customers

Suppose that 2500 customers subscribe to a telephone exchange. There are 80 trunk lines available. Any one customer has the probability of 0.03 of needing a trunk line on a given call. Consider the situation as 2500 trials with probability of "success" $p = 0.03$. What is the approximate probability that the 2500 customers will "tie up" the 80 trunk lines at any given time?

Section 7: Poisson Converges to Normal

Using the Central Limit Theorem, Answer the following parts.

(a) For large values of λ show the convergence of the Poisson distribution to the normal distribution. Also, using the moment generating function of the previous distribution, show that it is a standard normal distribution.

(b) The cylinders of a formula one car, oscillate about 900 times per hour. Supposing that this process follows a Poisson distribution, find the probability of such a cylinder oscillating 950 times per hour.

Section 8: Convergence Rate of Distributions

Use either R or Python to answer the following questions.

In this section, you will simulate, compare and discuss the convergence rate of multiple distributions.

(a) We will be discussing four different distributions in this question. Simulate a normal distribution with a mean of zero and a standard deviation of 100, an exponential distribution with a λ of 1, a Poisson distribution with a μ of 100, and a binomial distribution with n = 30 and p = 0.1.

(b) For each distribution try three different sampling rates and choose samples from them. Then calculate the mean of each distribution for each sampling rate by employing the estimation of central limit theorem.

(c) Compare the calculated mean in each set of distribution and sampling rate, with that distribution's actual mean and report the error of estimation by using the Mean Squared Error method.

(d) Finally, draw the graph of value of error per sampling rate for each distribution and discuss the results.

Section 9: Simulation of Normal vs. CLT

Use either R or Python to answer the following questions.

As you know, When the needed assumptions for the central limit theorem do not stand, we can use the normal estimation formula for calculating the accuracy of the estimated mean of samples.The normal estimation formula is as below:

$$P(|\bar{X}-\mu|<\epsilon) \approx 2\Phi(\sigma\epsilon\sqrt{n})-1$$

In this section you will need to simulate the above probability and compare it with its analytic alternative. For this purpose, generate a population of 1000 members which follow a normal distribution with a mean of zero and a standard deviation of 250.

(a) Considering a sample size of 30, execute 100 repetitions of sampling without permutations(or without putting the samples back into the population,) and draw the 99% confidence interval for each sampling repetition on a graph depicting the repetition number on the horizontal axis and the sample mean on the vertical axis.

(b) Using the above sampling repetitions, calculate the probability $P(|\bar{X}-\mu|<0.01)$ and compare it with the normal estimation probability given in the beginning of this section. (hint: since the given probability has a threshold of 0.01, you can use the concentration of the 99% confidence intervals from the previous part to calculate this probability.)

(c) Repeat the previous parts while executing the sampling repetitions with permutations. Compare and discuss the difference in results.

Section 10: Bonus Problem: Buffon's Needle

Use either R or Python to answer the following questions.

Buffon's Needle is one of the oldest problems in the field of geometrical probability. It was first stated by the French naturalist and mathematician, Comte de Buffon in 1777. The original version of it goes as follows. Suppose you have a floor made of long wooden planks each having a width equal to 1 unit of length. A large number, N, of needles, each also of length of 1 unit, are dropped randomly onto the floor.

(a) Firstly, find the probability of a needle landing on the crack between two adjacent wooden planks.

(b) If N_0 is the total number of the needles which either touch or cross the cracks between any two adjacent planks, show that the value of π can be estimated as $\frac{2N}{N_0}$.

(c) Buffon's Needle problem is essentially solved by Monte-Carlo integration. In general, Monte-Carlo methods use statistical sampling to approximate the solutions of problems that are difficult to solve analytically. Here, using either python or R, write a script that simulates the needle toss and uses the said method to calculate the probability of a needle intersecting a crack, and derives the value of π.