MTH 201: Probability and Statistics

Quiz 3

06/06/2023

Sanjit K. Kaul

No books, notes, or devices are allowed. Just a pen and eraser. Any exchange of information related to the quiz with a human or machine will be deemed as cheating. Institute rules will apply. Explain your answers. Show your steps. Approximate calculations are fine as long as the approximations are reasonable. You have 45 minutes.

Question 1. 30 **marks** You enter a casino to play $n > 0$ games. You can choose each game to be either a large bets game or a small bets game with equal probability. A large bets game results in a gain of 10 with probability 0.4 and a gain of -10 otherwise. A small bets game results in a gain of 1 with probability 0.4 and a gain of -1 otherwise. We are interested in the total gain G_n, which is the sum of gains obtained from the n games.

Answer the following questions.

(a) (20 marks) Derive the approximate CDF of G_n by assuming that G_n can be approximated as a Gaussian RV .

(b) (10 marks) Calculate the CDF of G_n in the limit as $n \to \infty$. Explain your answer.

Question 2. 30 **marks** A course has 39 lectures. For lecture $i \in \{1, 2, \ldots, 39\}$, let X_i be a random variable that takes the value 1 in case a student attends lecture i, an event that occurs with probability p. Otherwise, it takes the value 0. Let Y_i, $i \in \{1, 2, \ldots, 39\}$, be a random variable that takes a value 1 in case the lecturer records student attendance during lecture i, an event that takes place with probability q. Otherwise, Y_i takes a value 0. For a student to **record attendance** during any lecture, the student must attend the lecture and the lecturer must record attendance. Let Z be the total **recorded attendance** of the student at the end of 39 lectures.

Answer the following questions.

(a) (5 marks) Write down the RV Z in terms of the random variables X_i and Y_i, $i = 1, 2, \ldots, 39$.

(b) (5 marks) Derive the expected value of Z.

(c) (10 marks) Derive the variance of Z. Assume that all the RVs X_i, Y_i, $i = 1, 2, \ldots, 39$, are independent.

(d) (10 marks) Derive the moment generating function $E[e^{sZ}]$ of Z using the fact that Z is a sum of independent random variables. Use the fact to simplify the expected value you need to calculate.

Question 3. 40 **marks** Number of students that attend any lecture is given by the RV $Z = \sum_{i=1}^{m} S_i$, where m is the total number of enrolled students and the S_i are Bernoulli RVs with unknown parameter p. We would like to estimate the average number of students that attend any lecture. We want an unbiased estimator. We want a confidence interval estimate with interval length $2c = 0.1m$ and confidence 95%. Use the Chebyshev's inequality to calculate the minimum size of sample required.

Does the sample size increase or decrease with an increase in m? Provide a justification for why your observation is as expected.

Question 1. 30 **marks** You enter a casino to play $n > 0$ games. You can choose each game to be either a large bets game or a small bets game with equal probability. A large bets game results in a gain of 10 with probability 0.4 and a gain of -10 otherwise. A small bets game results in a gain of 1 with probability 0.4 and a gain of -1 otherwise. We are interested in the total gain G_n, which is the sum of gains obtained from the n games.

Answer the following questions.

(a) (20 marks) Derive the approximate CDF of G_n by assuming that G_n can be approximated as a Gaussian RV .

(b) (10 marks) Calculate the CDF of G_n in the limit as $n \to \infty$. Explain your answer.

Let X_i be the gain in the ith game.

$$G_n = \sum_{i=1}^{n} X_i$$

(a) To approximate using a Gaussian, we must calculate the expected value $E[G_n]$ & $Var[G_n]$.

$$E[G_n] = \sum_{i=1}^{n} E[X_i]$$

$$E[X_i] = (0.5)\left[(0.4)1 + (0.6)(-1)\right] + 0.5\left[(0.4)10 + (0.6)(-10)\right]$$

$$= \frac{-0.2}{2} + \frac{1}{2}(-2)$$

$$= -0.1 - 1 = -1.1.$$

$$E[G_n] = -1.1n.$$

Calculating the expectation of G_n

3

$Var[G_n] = Var\left[\sum_{i=1}^{n} X_i\right]$. Given that the X_i are independent RVs,

$$Var[G_n] = \sum_{i=1}^{n} Var[X_i]$$

$$Var[X_i] = E[X_i^2] - (E[X_i])^2$$

$$= E[X_i^2] - (1.1)^2$$

$$E[X_i^2] = (0.5)((0.4)(1)^2 + (0.6)(-1)^2) + (0.5)((0.4)(10)^2 + (0.6)(-10)^2)$$

$$= (0.5)(1) + (0.5)(100)$$

$$= \frac{101}{2} = 50.5$$

$$Var[X_i] = 50.5 - 1.21$$

$$= 49.29.$$

$$Var[G_n] = 49.29n.$$

Calculating the variance of G_n

5

The CDF $F_{G_n}(x)$

$$= P[G_n \le x]$$

$$= P\left[G_n - E[G_n] \le x - E[G_n]\right]$$

$$= P\left[\frac{G_n - E[G_n]}{\sqrt{Var[G_n]}} \le \frac{x - E[G_n]}{\sqrt{Var[G_n]}}\right]$$

$$\approx \Phi\left(\frac{x - E[G_n]}{\sqrt{Var[G_n]}}\right) = \Phi\left(\frac{x + 1.1n}{\sqrt{49.29n}}\right)$$

Correctly using the Gaussian approximation

10

(b) We were introduced to the CLT. However, the CLT doesn't say much about the CDF of $F_{G_n}(x)$ in the limit as $n \to \infty$.

All we know is that the CDF of the RV $\frac{G_n - E[G_n]}{\sqrt{Var[G_n]}}$ in the limit as $n \to \infty$ converges to the CDF of the standard normal.

Saying that $F_{G_n}(x) \to \Phi$ is wrong.

10 for stating what the CLT says about convergence.

Question 2. 30 **marks** A course has 39 lectures. For lecture $i \in \{1, 2, \ldots, 39\}$, let X_i be a random variable that takes the value 1 in case a student attends lecture i, an event that occurs with probability p. Otherwise, it takes the value 0. Let Y_i, $i \in \{1, 2, \ldots, 39\}$, be a random variable that takes a value 1 in case the lecturer records student attendance during lecture i, an event that takes place with probability q. Otherwise, Y_i takes a value 0. For a student to **record attendance** during any lecture, the student must attend the lecture and the lecturer must record attendance. Let Z be the total **recorded attendance** of the student at the end of 39 lectures.

Answer the following questions.

(a) (5 marks) Write down the RV Z in terms of the random variables X_i and Y_i, $i = 1, 2, \ldots, 39$.

(b) (5 marks) Derive the expected value of Z.

(c) (10 marks) Derive the variance of Z. Assume that all the RVs X_i, Y_i, $i = 1, 2, \ldots, 39$, are independent.

(d) (10 marks) Derive the moment generating function $E[e^{sZ}]$ of Z using the fact that Z is a sum of independent random variables. Use the fact to simplify the expected value you need to calculate.

(a) $$Z = \sum_{i=1}^{39} X_i Y_i$$ 5

(b) $$E[Z] = \sum_{i=1}^{39} E[X_i Y_i]$$

Okay to assume independence.

Given that the RVs are independent

$$E[X_i Y_i] = E[X_i]\,E[Y_i]$$ 5

$$\therefore E[Z] = \sum_{i=1}^{39} E[X_i]\,E[Y_i]$$

$$= \sum_{i=1}^{39} pq = 39pq.$$

Alternately, let $Z_i = X_i Y_i$.

$Z_i = 1$ w.p pq

$Z_i = 0$ otherwise.

$\therefore E[Z_i] = pq.$

(c) $$Var[Z] = \sum_{i=1}^{39} Var[X_i Y_i]$$ 10

$$= \sum_{i=1}^{39} Var[Z_i]$$

$$= 39\left(E[Z_i^2] - (E[Z_i])^2\right)$$

$$= 39\,pq(1-pq)$$

(d) $$E[e^{sZ}] = E\left[e^{s\sum_{i=1}^{39} Z_i}\right],$$ where $Z_i = X_i Y_i$.

$$E[e^{sZ}] = E\left[e^{s(Z_1 + Z_2 + \cdots + Z_{39})}\right]$$ 5

$$= E\left[e^{sZ_1} e^{sZ_2} \cdots e^{sZ_{39}}\right]$$

$$= E\left[e^{sZ_1}\right] E\left[e^{sZ_2}\right] \cdots E\left[e^{sZ_{39}}\right]$$

$\because$ the Z_i are independent, the expected value can be written as a product of expectations.

$$E[e^{sZ_i}] = pq\,e^s + (1-pq)(1)$$ 5

$$= 1 - pq + pq\,e^s.$$

$$\therefore E[e^{sZ}] = (1 - pq + pq\,e^s)^{39}.$$

Question 3. 40 **marks** Number of students that attend any lecture is given by the RV $Z = \sum_{i=1}^{m} S_i$, where m is the total number of enrolled students and the S_i are Bernoulli RVs with unknown parameter p. We would like to estimate the average number of students that attend any lecture. We want an unbiased estimator. We want a confidence interval estimate with interval length $2c = 0.1m$ and confidence 95%. Use the Chebyshev's inequality to calculate the minimum size of sample required.

Does the sample size increase or decrease with an increase in m? Provide a justification for why your observation is as expected.

$M_n(Z)$ is an unbiased estimator of $E[Z]$, where n is the sample size.

We want $P\left[|M_n(Z) - E[Z]| \geq \frac{0.1m}{2}\right] \leq 0.05$ 5

Since $M_n(Z)$ is an unbiased estimator

$$P\left[|M_n(Z) - E[Z]| \geq \frac{0.1m}{2}\right] = P\left[|M_n(Z) - E[M_n(Z)]| \geq \frac{0.1m}{2}\right]$$ 5

$$\leq \frac{Var[Z]}{n\left(\frac{0.1m}{2}\right)^2}$$

The bound is given by the Chebyshev's inequality.

Chebyshev's bound 10

$$\leq \frac{mp(1-p)}{n\left(\frac{0.1m}{2}\right)^2}$$

Correct $Var[Z]$ & RHS 10

We want

$$\frac{mp(1-p)}{n\left(\frac{0.1m}{2}\right)^2} \leq 0.05$$

$$n \geq \frac{mp(1-p)}{(0.05)\left(\frac{0.1}{2}\right)^2 m^2}$$

$$= \frac{0.25}{m(0.05)^3} = \frac{0.25}{m(5)^3 10^{-6}}$$

$$= \frac{25 \times 10^4}{125m}$$

$$= \frac{10^4}{5m}$$

$$= \frac{2000}{m}.$$

5 for simplifying correctly.

The sample size n is inversely proportional to m. This is because as m increases, the $STDDEV(Z_n)$ increases as $\sqrt{m}$. However the interval length $0.1m$ increases as m.

5