TRINITY COLLGE DUBLIN

School of Computer Science and Statistics

Extra Questions ST3009: Statistical Methods for Computer Science

NOTE: There are more example questions in Chapter 8 of the course textbook "A First Course in Probability" by Sheldon Ross, but ignore those questions involving continuous-valued random variables.

Question 1. Consider a six sided die and let X be the number that we observe when it is thrown. We know that $E[X] = 3.5$.

(a) What is $P(X \geq 5)$?

(b) Using Markov's inequality obtain a bound on $P(X \geq 5)$. How does it compare with the exact value in (a) ?

Solution

- $P(X \geq 5) = 2/6 = 0.333$
- By Markovs inequality $P(X \geq 5) \leq E[X]/5 = 3.5/5 = 0.7$

Question 2. Sometimes I forget a few items when I leave the house in the morning. For example, here are the probabilities that I forget various pieces of footwear: left sock 0.2, right sock 0.1, left shoe 0.1, right shoe 0.3. Let X be the number of these that I forget.

(a) What is $E[X]$? Hint. Let X_1 be 1 when I forget my left sock and 0 otherwise, similarly $X_2 = 1$ when I forget my right sock, $X_3 = 1$ when I forget my left shoe and $X_4 = 1$ when I forget my right shoe. Then $X = X_1 + X_2 + X_3 + X_4$.

(b) Use the Markov Inequality to upper bound the probability that I forget 3 or more items.

Now suppose that I forget each item independently.

(c) What is Var(X) ?

(d) Use Chebyshev's inequality to upper bound the probability that I forget 2 or more items.

Solution

- Let X_1 be 1 when I forget my left sock and 0 otherwise, similarly $X_2 = 1$ when I forget my right sock, $X_3 = 1$ when I forget my left shoe and $X_4 = 1$ when I forget my right shoe. Then $X = X_1 + X_2 + X_3 + X_4$ and $E[X] = E[X_1] + E[X_2] + E[X_3] + E[X_4] = 0.2 + 0.1 + 0.1 + 0.3 = 0.7$.
- $P(X \geq 3) \leq E[X]/3 = 0.2333$.
- Since the events are independent, $Var(X) = Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4)$. Now $E[X_1^2] = E[X_1]$ so $Var(X_1) = E[X_1^2] - E[X_1]^2 = 0.2 - 0.2^2 = 0.16$. Similarly, $Var(X_2) = 0.09$, $Var(X_3) = 0.09$, $Var(X_4) = 0.21$. Therefore $Var(X) = 0.55$.
- Chebyshev's inequality is that $P(|X - E[X]| \geq k) \leq Var(X)/k^2$. We have $E[X] = 0.7$ and $Var(X) = 0.55$ so this becomes $P(|X - 0.7| \geq k) \leq 0.55/k^2$. Selecting $k = 1.3$ then $P(|X - 0.7| \geq 1.3) = P(X \geq 2) \leq 0.55/1.3^2 = 0.3254$.

Question 3. A post office handles, on average, 10000 letters a day.

(a) Using Markov's inequality, what can be said about the probability that it will handle at least 15000 letters tomorrow ?

(b) Suppose now that the variance σ^2 in the number of letters per day is 2000. Using Chebyshev's inequality what can be said about the probability that this post office handles between 8000 and 12000 letters tomorrow?

(c) Using Chebyshev's inequality how can we bound the probability that it will handle at least 15000 letters tomorrow ? How does it compare with the bound in (a).

Solution

- Let X be the number of letters handled tomorrow. Then $E(X) = 10000$. By Markovs inequality we have $P(X \geq 15000) \leq E[X]/15000 = 2/3$
- We want $P(8000 < X < 12000) = P(-2000 < X - 1000 < 2000) = P(|X - 10000| < 2000$. By Chebyshev's inequality we have $P(|X - 10000| \geq 2000) \leq \sigma^2/2000^2 = 2000/2000^2 = 1/2000$. Now $P(|X - 10000| < 2000) = 1 - P(|X - 10000| \geq 2000)$ and so it follows that $P(|X - 10000| < 2000) \geq 1 - 1/2000 = 0.9995$.
- $P(X \geq 15000) = P(X - 10000 \geq 5000) \leq P(X - 10000 \geq 5000$ and $X - 10000 \leq -5000) = P(|X - 10000| \geq 5000)$. By Chebyshev, $P(|X - 10000| \geq 5000) \leq 2000/5000^2 = 1/12500$. This is much smaller than the bound of 2/3 using Markov's inequality.

Question 4. A biased coin, which lands heads with probability 1/10 independently each time it is flipped, is flipped 200 times consecutively. Using Markov's inequality give a bound on the probability that it lands heads 120 times or more.

Solution

- Let X_i equal 1 of a head is observed in the i'th toss and 0 otherwise. $E[X_i] = P(X_i = 1) = 1/10$. $X = \sum_{i=1}^{200} X_i$ be the number of heads observed. $E[X] = E[\sum_{i=1}^{200} X_i] = \sum_{i=1}^{200} E[X_i] = 200 \times 1/10 = 20$ by linearity of the expectation. Now by Markov's inequality $P(X \geq 120) \leq E[X]/120 = 20/120 = 1/6$.

Question 5. Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this weeks production will exceed 75?

(b) If the variance of a weeks production is known to equal 25, then what can be said about the probability that this weeks production will be between 40 and 60?

Hint: use Markov and Chebyshev inequalities.

Solution

- By Markovs inequality $P(X \geq 75) \leq E[X]/75 = 50/75 = 2/3$
- By Chebyshev's inequality $P(|X - 50| \geq 10) \leq \sigma^2/10^2 = 1/4$. Hence, $P(|X - 50| < 10) \geq 1 - 1/4 = 3/4$ and so the probability that this weeks production will be between 40 and 60 is at least 0.75.

Question 6. You would like to estimate the average number of hours p per day that a TCD student spends on youtube. To do this you plan to carry out a survey of the students by sampling n students independently and uniformly at random from the population. Letting X_i be the number of hours spent by student i in the sample, suppose the mean can be estimated as $X = \frac{1}{n} \sum_{i=1}^{n} X_i$.

(a) X is a Binomial random variable, with the X_i's all having mean p. Express $Var(X)$ in terms of p.

(b) Using the fact that $x(1-x) \leq 0.25$ for all $0 \leq x \leq 1$ (this can be verified using matlab), what is the maximum value of $Var(X)$?

(c) Using Chebyshev's inequality discuss how the value of n can be selected so as to ensure $P(|X-p| \geq 0.05) \leq 0.05$. Recall Chebyshev's inequality is $P(|X - E[X]| \geq k) \leq Var(X)/k^2$.

Solution

- $E[X_i] = p$ and $E[X_i^2] = p$ so $Var(X_i) = p - p^2 = p(1-p)$. Now $E[X] = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n E[X_i] = p$. And $Var(X) = Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = p(1-p)/n$ since the students are sampled independently.

- $Var(X) = p(1-p)/n$ and $p(1-p) \leq 0.25$. Therefore $Var(x) \leq 0.25/n$.

- Chebyshev's inequality gives us $P(|X - E[X]| \geq k) \leq Var(X)/k^2$. Since $E[X] = p$ and $Var(x) \leq 0.25/n$ it follows that $P(|X-p| \geq k) \leq 0.25/(nk^2)$. Selecting $k = 0.05$, then $P(|X-p| \geq 0.05) \leq 0.25/(0.05^2 n) = 100/n$. Selecting $n \geq 2000$ ensures that $100/n \leq 0.05$ and so $P(|X-p| \geq 0.05) \leq 0.05$ as required.

Question 7. In a study on cholestrol levels a sample of 12 men and women was chosen. The plasma cholestrol levels (mmol/L) of the subjects were as follows:

6.0,6.4,7.0,5.8,6.0,5.8,5.9,6.7,6.1,6.5,6.3,5.8

(a) Estimate the mean of the plasma cholestrol levels with a 95% confidence interval.

(b) What assumptions did you make about the sample in order to make your estimate ?

Solution We suppose that the cholestrol levels for the 12 people are random variable X_i, $i = 1, 2, \ldots, 12$. Let $X = \frac{1}{12}\sum_{i=1}^{12} X_i = 6.1917$ be our estimate of the mean. For the confidence interval we want to find b such that $P(|X - E[X]| \geq bE[X]) \leq 0.05$. There are lots of ways we could go about this.

- One is to use Chebyshev's inequality $P(|X - E[X]| \geq k) \leq \sigma^2/k^2$. Selecting $k = \sqrt{20}\sigma$ then $P(|X - E[X]| \geq \sqrt{20}\sigma) \leq 1/20 = 0.05$. We don't know the variance σ^2, but we can estimate it from the data as $(\frac{1}{12}\sum_{i=1}^{12} X_i^2) - X^2 = 38.4775 - 6.1917^2 = 0.1404$, giving an estimate $\sqrt{0.1404} = 0.3746$ for σ and so estimated confidence interval $P(|6.1917 - E[X]| \geq 1.67) \leq 0.05$ i.e. $P(4.52 \leq E[X] \leq 7.86) \leq 0.05$.

- We have assumed that the samples are independent random variables with the same mean and variance, which may well not be true since there is e.g. a mix of men and women plus some may be well and others unwell. We have also used an estimate for σ^2 to estimate the confidence interval. This estimate may not be the true variance value since its based on the measured data, but we have ignored this possible mismatch.