MATH-UA.0235 Probability and Statistics – Worksheet # 7

Problem 1 – Derivation of the Markov inequality for continuous random variables

Let X be a continuous random variable which only takes positive values, and let f_X be its probability density function. Let $a > 0$. We may write

$$\begin{aligned} E[X] = \int_0^{+\infty} x f_X(x)dx &= \int_0^a x f_X(x)dx + \int_a^{+\infty} x f_X(x)dx \\ &\geq \int_a^{+\infty} x f_X(x)dx \\ &= a \int_a^{+\infty} f_X(x)dx \\ &= aP(X \geq a) \end{aligned}$$

We thus also have the Markov inequality

$$E[X] \geq aP(X \geq a)$$

which we had derived for positive discrete random variables in Lecture 10.

Problem 2

Suppose that X is a random variable with mean 20 and variance 20. Can you try to provide an accurate lower bound for $P(0 < X < 40)$?

$$P(0 < X < 40) = P(|X - 20| < 20) = 1 - P(|X - 20| \geq 20)$$

We may now use Chebyshev's inequality to obtain a quite good upper bound for the probability on the right-hand side:

$$P(|X - 20| \geq 20) \leq \frac{\mathrm{Var}(X)}{400} = \frac{1}{20}$$

Hence, we have the lower bound

$$P(0 < X < 40) \geq 1 - \frac{1}{20} = \frac{19}{20}$$

Problem 3

Let N be a Poisson random variable with mean 20.

1. Use the Markov inequality to find an upper bound for $p = P(X \geq 26)$.

   According to the Markov inequality

   $$20 \geq 26p \quad \Leftrightarrow \quad p \leq \frac{10}{13}$$

2. Try to find a more accurate upper bound inspired by the Chebyshev inequality.

   Since we are interested in $P(X \geq 26)$, we cannot use Chebyshev's inequality as such. However, with some manipulations, we will be able to use that inequality.

   $$\begin{aligned} p = P(X \geq 26) = P(X - E[X] \geq 26 - E[X]) &= P(X - 20 \geq 6) \\ &\leq P((X - 20)^2 \geq 36) \end{aligned}$$

We are now ready to use Chebyshev's inequality:

$$P((X-20)^2 \geq 36) = P(|X-20| \geq 6) \leq \frac{\text{Var}(X)}{36} = \frac{20}{36} = \frac{5}{9}$$

We thus have

$$p \leq \frac{5}{9}$$

With see that Chebyshev's inequality, which relies on knowledge of the variance in addition to the expectation, gives us a tighter upper bound than Markov's inequality, which only relies on knowledge of the expectation.

Even more precise upper bound

With a bit of extra finessing, we can get an even tighter upper bound with similar reasoning to what we just showed. Here is how it goes. Let $b \in \mathbb{R}$.

$$p = P(X - 20 + b \geq 6 + b) \leq P((X-20+b)^2 \geq (6+b^2))$$

We may now use Markov's inequality:

$$P((X-20+b)^2 \geq (6+b^2)) \leq \frac{E[(X-20+b)^2]}{(6+b^2)} = \frac{\text{Var}(X-20+b) + (E[(X-20+b)])^2}{(6+b)^2} = \frac{\text{Var}(X)+b^2}{(6+b)^2}$$

We conclude that for all $b \in \mathbb{R}$ such that $6 + b \neq 0$,

$$p \leq \frac{20+b^2}{(6+b)^2}$$

Now, to obtain the tightest possible upper bound with this method, we look for the minimum of the function $h(b) = \frac{20+b^2}{(6+b)^2}$.

$$h'(b) = \frac{2b(6+b)^2 - 2(20+b^2)(6+b)}{(6+b)^4} = \frac{4(3b-10)}{(6+b)^3}$$

We see that the minimum of the function is reached for $b = \frac{10}{3}$, and h takes the value $\frac{5}{14}$. We conclude that the best estimate with this method is

$$p \leq \frac{5}{14}$$

This is indeed the tightest upper bound we have derived.

Note that for $b = 0$, $h(0) = \frac{5}{9}$, which is the result we had obtained with the more standard Chebyshev inequality.

Problem 4

An environmental engineer believes that there are two contaminants in a water supply: arsenic and lead. The actual concentrations of the two contaminants are independent random variables X and Y, measured in the same units. The engineer is interested in what proportion of the contamination is lead on average, i.e. she wants to know the expected value of $R = \frac{Y}{X+Y}$. She therefore decides to collect n pairs (X_1, Y_1), to compute $R_i = \frac{Y_i}{X_i+Y_i}$ for each pair, and to estimate $E[R]$ by the sample average $\overline{R}_n = \frac{1}{n}\sum_{i=1}^n R_i$. How many samples will she need if she wants to be 98% certain that she will have an error of less than 0.5%?

The engineer wants to find n such that

$$P(|\overline{R}_n - R| \geq 0.005) \leq 0.02$$

Let σ be the variance of any of the R_i. As we saw in class, we have

$$\text{Var}(\overline{R}_n) = \frac{\sigma^2}{n}$$

Now, by construction, the R_i can only take values between 0 and 1. Thus, $\sigma \leq 1$. Hence,

$$\text{Var}(\overline{R}_n) \leq \frac{1}{n}$$

Using Chebushev's inequality, we may write

$$P(|\overline{R}_n - R| \geq 0.005) \leq \frac{1}{0.005^2 n}$$

The engineer therefore needs n to be such that

$$\frac{1}{0.005^2 n} \leq 0.02 \quad \Leftrightarrow \quad n \geq \frac{1}{0.02 \cdot 0.005^2} = 2 \cdot 10^6$$

According to this estimate, she will need 2 million samples to have the desired confidence to estimate R with the desired accuracy. With more advanced methods from probability and statistics and more refined estimates, she may be able to obtain the desired performance with fewer samples.

Problem 5

The purpose of this problem is to illustrate the final remark we made in the previous problem: while Chebyshev's inequality is easy and convenient to apply to obtain upper or lower bounds, it may be fairly inaccurate.

Suppose that a fair coin is tossed n times in a row. For $i = 1, \ldots, n$, let $X_i = 1$ if a head is obtained on the ith toss, and $X_i = 0$ if a tail is obtained on that toss. We consider the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, which can naturally be interpreted as the proportion of heads obtained among the n tosses. Since the coin is fair, we expect $\overline{X}_n$ to converge to $\frac{1}{2}$ as $n \to +\infty$. Here, we want to find out how many times n one must toss the coin so that $P(0.4 \leq \overline{X}_n \leq 0.6) \geq 0.7$.

1. Estimate n using Chebyshev's inequality.

   We know that each X_i is a Bernoulli random variable with parameter $\frac{1}{2}$, so that $\text{Var}(X_i) = \frac{1}{4}$ for any i. We therefore have

   $$\text{Var}(\overline{X}_n) = \frac{1}{4n}$$

   Furthermore, by the linearity of expectation, $E[\overline{X}_n] = E[X_i] = \frac{1}{2}$ for any i. We have

   $$P(0.4 \leq \overline{X}_n \leq 0.6) = P(|\overline{X}_n - E[\overline{X}_n]| \leq 0.1) = 1 - P(|\overline{X}_n - E[\overline{X}_n]| > 0.1)$$

   We can now apply Chebyshev's inequality to the second term:

   $$P(|\overline{X}_n - E[\overline{X}_n]| > 0.1) \leq \frac{1}{0.04n}$$

   so that

   $$P(0.4 \leq \overline{X}_n \leq 0.6) \geq 1 - \frac{1}{0.04n}$$

   We need to find n such that

   $$1 - \frac{1}{0.04n} \geq 0.7 \quad \Leftrightarrow \quad n \geq \frac{1}{0.04 \cdot 0.3} = \frac{250}{3}$$

   We conclude that according to Chebyshev's inequality, we will need 84 tosses.

2. Let $n = 20$, and compute $P(0.4 \leq \overline{X}_n \leq 0.6)$ exactly. Show that for $n = 20$, the desired criterion is already satisfied.

   Let $S_{20} = 20 \cdot \overline{X}_{20} = \sum_{i=1}^{20} X_i$. By construction, S_{20} has a binomial distribution with parameters 20 and $\frac{1}{2}$.

   $$\begin{aligned} P(0.4 \leq \overline{X}_{20} \leq 0.6) &= P(0.4 \cdot 20 \leq S_{20} \leq 0.6 \cdot 20) \\ &= P(8 \leq S_{20} \leq 12) = P(S_{20} = 8) + P(S_{20} = 9) + P(S_{20} = 10) + P(S_{20} = 11) + P(S_{20} = 12) \\ &= \left(\frac{1}{2}\right)^{20} \left[\binom{20}{8} + \binom{20}{9} + \binom{20}{10} + \binom{20}{11} + \binom{20}{12}\right] \\ &\approx 0.737 \end{aligned}$$