University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2019)

Homework 1

Solutions

Problem 1: Show that for any $a \geq 1$ there exist distributions for which Markov's inequality is tight by showing the following:

- For each $a = 2, 3$, and 4 give a probability distribution $p(x)$ for a nonnegative random variable x for which
$$\mathbf{P}[\, x \geq a \,] = \frac{\mathbf{E}[\, x \,]}{a}.$$
- For arbitrary $a \geq 1$ give a probability distribution for a nonnegative random variable x where
$$\mathbf{P}[\, x \geq a \,] = \frac{\mathbf{E}[\, x \,]}{a}.$$

Solution: For an arbitrary $a \geq 1$ we consider the probability distribution

$$p(x) = \begin{cases} \frac{1}{a} & \text{if} \quad x = a, \\ 1 - \frac{1}{a} & \text{if} \quad x = 0. \end{cases}$$

Then, it holds that

$$\mathbf{E}[\, x \,] = a \cdot \mathbf{P}[\, x = a \,] + 0 \cdot \mathbf{P}[\, x = 0 \,] = 1,$$

and hence

$$\mathbf{P}[\, x \geq a \,] = \mathbf{P}[\, x = a \,] = \frac{1}{a} = \frac{\mathbf{E}[\, x \,]}{a}.$$

Problem 2: Show that for any $c \geq 1$ there exist distributions for which Chebyshev's inequality is tight, in other words,

$$\mathbf{P}[\, |x - \mathbf{E}[\, x \,]| \geq c \,] = \frac{\mathbf{Var}[\, x \,]}{c^2}.$$

Solution: For an arbitrary $c \geq 1$ we consider the following probability distribution

$$p(x) = \begin{cases} \frac{1}{2c} & \text{if} \quad x = c, \\ 1 - \frac{1}{c} & \text{if} \quad x = 0, \\ \frac{1}{2c} & \text{if} \quad x = -c. \end{cases}$$

Then, it holds that

$$\begin{aligned} \mathbf{E}[\, x \,] &= c \cdot p(c) + 0 \cdot p(0) + (-c) \cdot p(-c) = 0, \\ \mathbf{E}\left[\, x^2 \,\right] &= c^2 \cdot p(c) + 0 \cdot p(0) + (-c)^2 \cdot p(-c) = c, \\ \mathbf{Var}[\, x \,] &= \mathbf{E}\left[\, x^2 \,\right] - \mathbf{E}[\, x \,]^2 = c. \end{aligned}$$

Hence,

$$\mathbf{P}[\, |x - \mathbf{E}[\, x \,]| \geq c \,] = \mathbf{P}[\, |x| \geq c \,] = \frac{1}{c} = \frac{\mathbf{Var}[\, x \,]}{c^2}.$$

Problem 3: Consider the probability density function $p(x) = 0$ for $x < 1$ and $p(x) = c \cdot \frac{1}{x^4}$ for $x \geq 1$.

- What should c be to make p a legal probability density function?
- Generate 100 random samples from this distribution. How close is the average of the samples to the expected value of x?

***<u>Solution</u>*:** Recall that p is a valid probability density function if

1. $p(x) \geq 0 \qquad \forall x \in \mathbb{R}$;
2. $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$.

Working with the second condition we have that

$$\int_{-\infty}^{\infty} p(x)\, dx = \int_{1}^{\infty} c \cdot \frac{1}{x^4}\, dx = c \cdot \left.\frac{1}{-3x^3}\right|_1^{\infty} = \frac{c}{3},$$

and therefore the first condition holds when $c = 3$.

For the second part of the question we will use the Law of Large Numbers, i.e.,

$$\mathbf{P}\left[\left|\frac{x_1 + \cdots + x_{100}}{100} - \mathbf{E}[\,x\,]\right| \geq \epsilon\right] \leq \frac{\mathbf{Var}[\,x\,]}{100\epsilon^2}.$$

We know that

$$\mathbf{E}[\,x\,] = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{1}^{\infty} 3 \cdot \frac{1}{x^3}\, dx = \left.\frac{3}{-2x^2}\right|_1^{\infty} = \frac{3}{2},$$

$$\mathbf{E}\left[\,x^2\,\right] = \int_{-\infty}^{\infty} x^2\, p(x)\, dx = \int_{1}^{\infty} 3 \cdot \frac{1}{x^2}\, dx = \left.\frac{3}{-x}\right|_1^{\infty} = 3,$$

and

$$\mathbf{Var}[\,x\,] = \mathbf{E}\left[\,x^2\,\right] - \mathbf{E}[\,x\,]^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}.$$

Therefore, it holds that

$$\mathbf{P}\left[\left|\frac{x_1 + \cdots + x_{100}}{100} - \mathbf{E}[\,x\,]\right| \geq \epsilon\right] \leq \frac{3}{400\epsilon^2}.$$

For example, if $\epsilon = 0.2$, the probability that the mean of the 100 samples lies outside the interval $(1.3, 1.7)$ is not grater than 0.19.

Problem 4: Let G be a d-dimensional Gaussian with variance $1/2$ in each direction, centred at the origin. Derive the expected squared distance to the origin.

***<u>Solution</u>*:** Suppose $G = (g_1, g_2, \ldots, g_d)$ where each $g_i \sim \mathcal{N}(0, 1/2)$. Direct calculation gives us that

$$\mathbf{E}\left[\,\|G - \mathbf{0}\|^2\,\right] = \mathbf{E}\left[\sum_{i=1}^{d} g_i^2\right] = \sum_{i=1}^{d} \mathbf{E}\left[\,g_i^2\,\right] = \sum_{i=1}^{d}(\mathbf{E}\left[\,g_i^2\,\right] - \mathbf{E}[\,g_i\,]^2) = \sum_{i=1}^{d} \mathbf{Var}[\,g_i\,] = \frac{d}{2}.$$

Problem 5: Let $x_1, \dots, x_n$ be independent samples of a random variable x with mean μ and variance σ^2. Let

$$m_s = \frac{1}{n}\sum_{i=1}^{n} x_i$$

be the sample mean. Suppose one estimates the variance using the sample mean rather than the true mean, that is,

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_s)^2.$$

Prove that

$$\mathbf{E}\left[\,\sigma_s^2\,\right] = \frac{n-1}{n}\sigma^2$$

and thus one should have divided by $n-1$ rather than n.

Solution: First of all, we will rewrite σ_s^2 by

$$\begin{aligned}
\sigma_s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - m_s)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i m_s + m_s^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - 2m_s\frac{\sum_{i=1}^{n} x_i}{n} + \frac{1}{n}\sum_{i=1}^{n} m_s^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - m_s^2.
\end{aligned}$$

Now, using the linearity of expectation we have

$$\mathbf{E}\left[\,\sigma_s^2\,\right] = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{E}\left[\,x_i^2\,\right]\right) - \mathbf{E}\left[\,m_s^2\,\right],$$

where

$$\mathbf{E}\left[\,x_i^2\,\right] = \mathbf{Var}\left[\,x_i\,\right] + \mathbf{E}\left[\,x_i\,\right]^2 = \sigma^2 + \mu^2,$$

and

$$\begin{aligned}
\mathbf{E}\left[\,m_s^2\,\right] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] \\
&= \frac{1}{n^2}\mathbf{E}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] \\
&= \frac{1}{n^2}\left(\mathbf{Var}\left[\sum_{i=1}^{n} x_i\right] + \mathbf{E}\left[\sum_{i=1}^{n} x_i\right]^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbf{Var}\left[\,x_i\,\right] + \left(\sum_{i=1}^{n}\mathbf{E}\left[\,x_i\,\right]\right)^2\right) \qquad \text{using that } x_i \text{ are independent} \\
&= \frac{\sigma^2}{n} + \mu^2.
\end{aligned}$$

Therefore we get

$$\mathbf{E}\left[\,\sigma_s^2\,\right] = \frac{1}{n}\left(\sum_{i=1}^{n}\sigma^2 + \mu^2\right) - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$