Homework 3 Solutions

1. ***Identifying a pmf from data***

Suppose Ω is a sample space and $x : \Omega \to V$ is a random variable. In this question we will not specify Ω and x directly. Instead we will work with the *induced pmf* p^x on V. Here $V \subset \mathbb{R}$, given by

$$v = \{ v_1, v_2, \ldots, v_n \}$$

where

$$v = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 7 \\ 8 \\ 9 \end{bmatrix}$$

The random variable x has induced pmf $p^x : V \to [0, 1]$. We represent p^x by a vector $p^x \in \mathbb{R}^n$, so that

$$p^x(v_i) = p^x_i$$

Let

$$p^x = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.3 \\ 0.2 \\ 0.1 \end{bmatrix}$$

(a) Find the cdf of x, and use it to simulate x. Perform $m = 10,000$ trials, and collect data points $y(1), \ldots, y(m)$, where each $y(i) \in V$.

Let $s(j, n)$ denote the frequency of v_j in the first n data points, i.e.,

$$s(j, n) = \frac{1}{n} \sum_{i=1}^{n} I_j(y(i))$$

where I_j is the indicator function

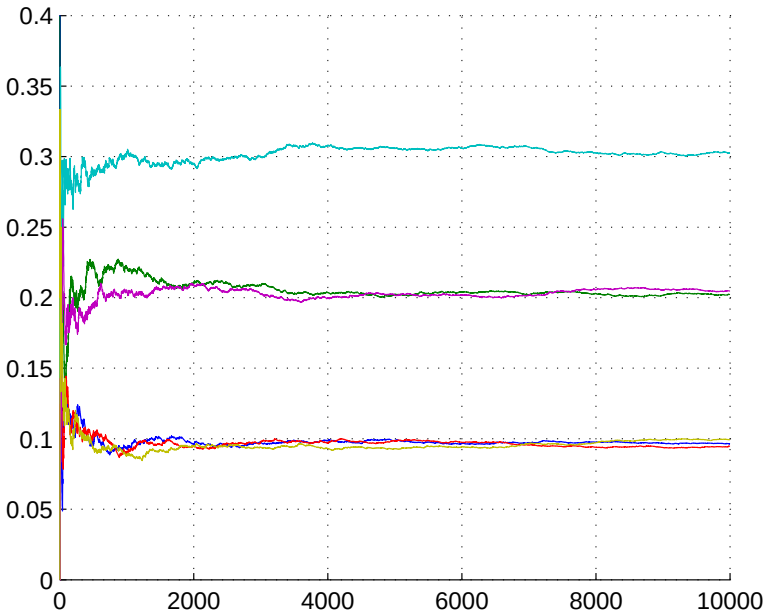
$$I_j(a) = \begin{cases} 1 & \text{if } a = v_j \\ 0 & \text{otherwise} \end{cases}$$

Plot $s(j, n)$ against n, for each j.

(b) Plot $s(4, n)$ against n, along with the bounds on its 90% confidence intervals given by the Chebyshev inequality.

Solution.

(a) A sample plot is shown below. As we can see, the empirical pmf converges to the actual pmf.



(b) Define the indicator random variable $I_4 : \Omega \to R$, by

$$I_4(x) = \begin{cases} 1 & \text{if } x = 7 \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$s(4,n) = \frac{1}{n}\sum_{i=1}^{n} I_4(x(n))$$

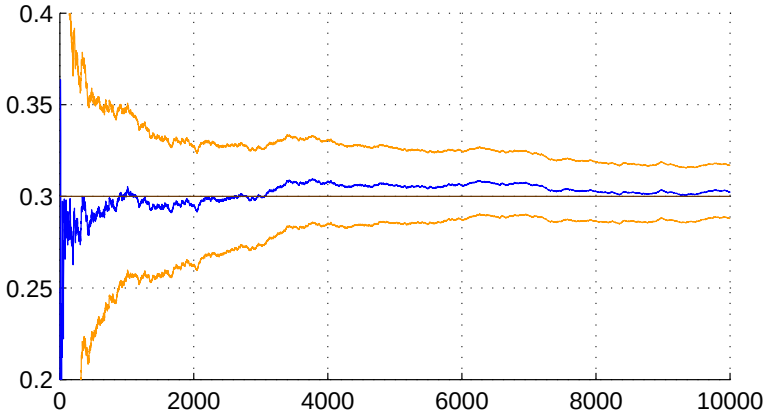
and $\mathrm{E}\, I_4 = 0.3$, hence

$$\Omega = \mathrm{cov}(I_4) = \mathrm{E}\, I_4^2 - (\mathrm{E}\, I_4)^2 = 0.21$$

For each n, half width of the 90% confidence interval is given by

$$\varepsilon = \sqrt{\frac{\Omega}{0.1n}}$$

The confidence intervals are shown below.



2. ***The geometric mean and products of IID random variables***

(a) Suppose $\Omega = \{1, 2, \ldots, n\}$ and $p : \Omega \to \mathbb{R}$ is the uniform pmf with

$$p(\omega) = \frac{1}{n} \qquad \text{for all } \omega \in \Omega$$

Suppose $x : \Omega \to \mathbb{R}$ is a random variable. Show that

$$\exp(\mathrm{E} \log x) = \prod_{i=1}^{n} x(\omega_i)^{1/n}$$

The right-hand side of this expression is the *geometric mean* of the values of x.

(b) Suppose $y_1, y_2, \ldots, y_n$ are IID Bernoulli random variables. Each y_i can take a value of either 6 or $\frac{1}{3}$, with

$$\mathrm{Prob}(y_i = 6) = \tfrac{1}{2} \qquad \mathrm{Prob}(y_i = \tfrac{1}{3}) = \tfrac{1}{2}$$

Let's define the product random variable $p_n : \Omega \to \mathbb{R}$ by

$$p_n = y_1 y_2 \ldots y_n$$

What is the expected value $\mathrm{E}\, p_n$ as a function of n?

(c) Now define the scalar random variable $q_n : \Omega \to \mathbb{R}$ by

$$q_n = \frac{1}{n} \log p_n$$

where log is the natural logarithm. Show that $\mathrm{E}\, q_n = \log \sqrt{2}$.

(d) As n becomes large, describe qualitatively the pmf of q_n.

Solution.

(a) We have

$$\begin{aligned} \mathrm{E} \log x &= \sum_{i=1}^{n} \frac{1}{n} \log x_i \qquad \text{by definition} \\ &= \log\left(\prod_{i=1}^{n} x_i^{\frac{1}{n}} \right) \end{aligned}$$

and the result follows.

(b) The expected value is

$$\begin{aligned} \mathrm{E}\, p_n &= \mathrm{E}(y_1 y_2 \ldots y_n) \\ &= \mathrm{E}\, y_1\, \mathrm{E}\, y_2 \ldots \mathrm{E}\, y_n \qquad \text{since the } y_i \text{ are IID} \\ &= \left(\frac{19}{6} \right)^n \end{aligned}$$

(c) We have

$$\begin{aligned} q_n &= \frac{1}{n} \log p_n \\ &= \frac{1}{n} \log(y_1 y_2 \ldots y_n) \\ &= \frac{1}{n} \sum_{i=1}^{n} \log y_i \end{aligned}$$

Hence the expected value is

$$\begin{aligned} \mathrm{E}\, q_n &= \mathrm{E}\, \frac{1}{n} \sum_{i=1}^{n} \log y_i \\ &= \mathrm{E} \log y_i \qquad \text{since the } y_i \text{ are IID} \\ &= \log \left(\frac{6^{\frac{1}{2}}}{3^{\frac{1}{2}}} \right) \\ &= \log \sqrt{2} \end{aligned}$$

(d) The random variable q_n is the sample mean of n IID random variables. Hence for large n, q_n is approximately Gaussian, and its covariance tends to zero as $n \to \infty$. That is, as $n \to \infty$, the pdf of q_n tends to a delta function at $\log \sqrt{2}$.

Interestingly p_n has mean $(19/6)^n$, even though $p_n = \exp(nq_n)$, and since q_n becomes a spike centered at $\log \sqrt{2}$, one would expect p_n to become a spike with a peak at $\sqrt{2}^n$.