Stochastics
Problem sheet 6 - Concentration theorems, some results
Fall 2021

3. A water cleaning facility cleans the waste water (sewage) from n factories. For each factory, the maximum daily output is 200 tons of waste water and the average daily output is 100 tons. The capacity of the water cleaning facility is C tons of water per day. We say that there is *overflow* on a given day if the total waste water produced by the factories exceeds the capacity.

   (a) $n = 100$ and $C = 14000$. Calculate the probability of overflow.

   (b) Assuming $n = 100$, determine an upper bound on C such that the probability of overflow is at most 10^{-6}.

   (c) Given $C = 12000$, calculate the maximum number of factories that can be allowed such that the probability of overflow is at most 10^{-6}.

   Solution.

   (a) We cannot use CLT since no deviation was given. We cannot use Cramer either, because to calculate the rate function, we would need to know the entire distribution of the output of a factory, which is not given. (Also, identical distribution for the daily output cannot be assumed either.) That leaves Hoeffding. Let $S = X_1 + \cdots + X_n$, where X_i is the daily output of factory i. We know $0 = a \leq X_i \leq b = 200$, and $\mathbf{E}X_i = 100$. In case $n = 100$ and $C = 14000$, this gives $\mathbf{E}S = n \cdot EX_i = 100 \cdot 100 = 10000$, and Hoeffding gives

$$\mathbf{P}(S > C) = \mathbf{P}(S > 14000) = \mathbf{P}(S > ES + 4000) \leq e^{-\frac{2 \cdot 4000^2}{100(200-0)^2}} = 3.35 \cdot 10^{-4}.$$

   (b) Now C is unknown; we set $C = \mathbf{E}S + t$ and we have

$$\mathbf{P}(S > C) = \mathbf{P}(S > ES + t) \leq e^{-\frac{2 \cdot t^2}{100(200-0)^2}} = 10^{-6},$$

   from which $t = 5256$ and $C = 15256$.

   (c) In this case, $\mathbf{E}S = n \cdot 100$ is unknown initially; still, we want to set $C = \mathbf{E}S + t$, which now gives $t = 12000 - n \cdot 100$. From Hoeffding, we have

$$\mathbf{P}(S > C) \leq e^{-\frac{2 \cdot (12000 - n \cdot 100)^2}{n(200-0)^2}} = 10^{-6},$$

   which will give a quadratic equation for n. The solutions are $n = 74$ and $n = 193$. Considering that n should be less than 100 (see part (a)), the answer is $n = 74$.

4. A student is participating in a test which has 100 simple choice questions (with 2 possible answers). For each question, she knows the answer with probability 1/2. If she doesn't know the answer, she picks one of the answers randomly. Estimate the probability that she gives a correct answer to at least 80 questions.

   Solution. For each question, she gives a correct answer with probability 0.75 (from total probability). $S = X_1 + \cdots + X_n$ is the number of correct answers, where X_i is 1 if the answer to question i was correct and 0 if it was incorrect and $n = 100$. Then $m = \mathbf{E}X_1 = 0.75$ and $\sigma = \mathbf{D}X_1 = 0.433$, and we apply the CLT to get

$$\mathbf{P}(S \geq 80) = 1 - \mathbf{P}(S < 80) = 1 - \mathbf{P}\left(\frac{S - nm}{\sqrt{n}\sigma} < \frac{80 - nm}{\sqrt{n}\sigma}\right) \approx$$
$$\approx 1 - \Phi\left(\frac{80 - nm}{\sqrt{n}\sigma}\right) = 1 - \Phi\left(\frac{80 - 100 \cdot 0.75}{\sqrt{100} \cdot 0.433}\right) = 1 - \Phi(1.15) = 1 - 0.875 = 0.125.$$

6. We toss a fair coin 1000 times. Give a large deviation estimate on the probability that it comes up heads at least 600 times.

   Solution. $S = X_1 + \dots X_n$ is the number of heads where X_i is 1 if flip i was heads and 0 if it was tails, and $n = 1000$. We know $0 \le X_i \le 1$, $\mathbf{E}X_i = 1/2$ and $\mathbf{E}S = n \cdot EX_i = 500$, so Hoeffding can be applied:

$$\mathbf{P}(S > 600) = \mathbf{P}(S > \mathbf{E}S + 100) \le e^{-\frac{2 \cdot 100^2}{1000 \cdot (1-0)^2}} = 2.06 \cdot 10^{-9}.$$

   Cramer can also be applied; for this, we use the rate function of the Bernoulli distribution (see problem 2, part (c)) with parameter $p = 1/2$, which is $I(x) = x \log\left(\frac{x}{1-x}\right) + \log(2(1-x))$. Cramer gives

$$\mathbf{P}(S > 600) = \mathbf{P}\left(\frac{S}{n} > \frac{600}{1000}\right) = \mathbf{P}\left(\frac{S}{n} \in [0.6, 1]\right) \le e^{-n \inf_{x \in [0.6,1]} I(x)}.$$

   The function $I(x)$ is 0 at the point $\mathbf{E}X_i = 1/2$ and increasing after that, so $\inf_{x \in [0.6,1]} I(x) = I(0.6) = 0.6 \log\left(\frac{0.6}{0.4}\right) + \log(2 \cdot 0.4) = 0.0201355$, so

$$\mathbf{P}(S > 600) \le e^{-1000 \cdot 0.0201355} = 1.8 \cdot 10^{-9}.$$

   Note also the numerical instability of Cramér: if we round the value of $I(0.6)$ to 0.02, we get

$$\mathbf{P}(S > 600) \le e^{-1000 \cdot 0.02} = 2.06 \cdot 10^{-9}.$$

15. A postal service delivery truck carries n packages. Each package has maximum weight 5 kg; the average package weight is 2 kg. The capacity of the truck is 1000 kg. Set the value of n so that the probability of n packages being overweight is under 10^{-4}.

    Result. $n = 393$. (Similar to 3. (c).)