Bonus Statistics & Math Homework

Brainster Academy for Data Science

March 2021

You can solve the exercises on any platform/programming language that you prefer. A photo of solutions on a paper (with camscanner) is acceptable too whenever it is possible (for example in the first problem). You are NOT expected to solve all the questions (but the more, the better).

1 Basic probabilistic inequality: Markov's inequality

We will try to show the basic Markov inequality in the case of discrete and finite random variables.

1.1 Reminders

The expected value of a random variable X that takes values in $\{1, 2, ..., n\}$ is defined as

$$\mathbb{E}(X) = \sum_{1 \le i \le n} i \times P(X = i).$$

1.2 The Markov Inequality

The Markov inequality states that:

Theorem 1. *for every real number $r > 0$ we have*

$$P(X \ge r) \le \frac{\mathbb{E}(X)}{r}.$$

We will assume that r is a positive integer and $r \le n$, for the sake of simplicity.

First remark the decomposition

$$\mathbb{E}(X) = \sum_{1 \le i \le n} i \times P(X = i) = \sum_{1 \le i < r} i \times P(X = i) + \sum_{r \le j} j \times P(X = j)$$

and the equality

$$P(X \ge r) = P(\{X = r\} \text{ or } \cdots \text{ or } \{X = n\}) = \sum_{r \le j} P(X = j)$$

Question 1. *Justify the following two inequalities:*

$$\sum_{1 \le i < r} i \times P(X = i) \ge 0$$

and

$$\sum_{r \le j} j \times P(X = j) \ge \sum_{r \le j} r \times P(X = j).$$

Question 2. *Deduce that*

$$\mathbb{E}(X) \ge r \sum_{r \le j} P(X = j)$$

and with that, the Markov inequality.

REMARK: Using this inequality in a slightly different form, rigorous bounds can be given for the Three-sigma problem without assuming that the distribution is normal.

2 Three-sigma

2.1 Introduction

The goal of this problem is to convince ourselves that, for the sample of a random variable X with finite variance $\sigma^2 = \mathbb{V}(X)$, we have a weaker version of the three-sigma rule even if we do not assume that the random variable X is Gaussian click here for the Wikipedia article of the three-sigma, or 68–95–99.7 rule. In many cases, this rule can be easily used as a very simple way of finding outliers in our data.

2.2 Simulation

Let us start off by convincing ourselves that this is indeed the case.

Question 3. *First, either pick a numerical column from some dataset with* > 200 *samples, or generate some random data in the following way: Pick a few of the distributions from the numpy random module and then generate* > 200 *samples (for example, pick the Gaussian and the uniform distribution for some parameters, and generate 150 samples from each). Possible ways of doing this in excel can be found here and here*

Question 4. *Calculate the (empirical) mean* μ *and variance* σ^2*, as well as the standard deviation* $\sigma = \sqrt{\sigma^2}$ *of the datapoints.*

Question 5. *Check how many datapoints are within the intervals* $I_1 = [\mu - 0.5\sigma, \mu + 0.5\sigma]$*,* $I_2 = [\mu - \sigma, \mu + \sigma], I_3 = [\mu - 2\sigma, \mu + 2\sigma]$*. What percentage of all datapoints is within each interval?*

Question 6. *(good-to-know) Give an example where the "outliers" are actually close to the mean, and the "proper" datapoints are "further" from the mean.*

(Hint: Consider, for example, the following scenario: the samples are very close to two "centers", say for example (-2) and (2), and very few samples (the "outliers") are close to 0. This can be done for example by generating some samples from $N(2, 1)$ *and some other samples from* $N(-2, 1)$ *where* $N(.,.)$ *is the Gaussian normal distribution)*

For this question, you can just give a possible plot of a histogram of such an occurrence. By hand, or in python, excel,...