Homework 1 Data Science II
Warm-up

Instructions

Please prepare a writeup following the general writeup instructions and submit on Brightspace. Please show all your work.

Problem 1. Included with this assignment is a time series dataset containing one year of measurements taken every 10 minutes. Besides the uninformative filename, what is wrong with these data?

Problem 2. Suppose an integer-valued $x \geq 0$ follows probability distribution $P(x)$. Then x has average value (expectation) of $\langle x \rangle = \mathbb{E}[x] = \sum_x xP(x)$ and likewise for $\langle x^2 \rangle = \mathbb{E}[x^2]$.

Now, imagine x has been "corrupted" into a new value $y \geq 0$, also integer-valued. The *distribution* of y, $Q(y)$, is related to $P(x)$ by

$$Q(y) = \sum_{x=y}^{\infty} P(x) \binom{x}{y} (1-p)^y p^{x-y}, \tag{1}$$

where $0 < p < 1$ is a "corruption probability". Assuming $\langle x \rangle$ (and therefore $\langle y \rangle$) is finite, write $\langle y \rangle$ and $\langle y^2 \rangle$ in terms of $\langle x \rangle$, $\langle x^2 \rangle$, and p.

Hint: x "choose" $y = 0$ if $y > x$. *Hint*: As you work, it's worth watching out for the first two moments of a binomial distribution.

Problem 3. *Torturing the data (analyst).*

We have at the ready an infinite number of true/false questions. Assume questions are unrelated and each question has an answer of T with probability $p = 1/20$, otherwise the answer is F. We continue to ask questions until the first T. How many questions should we *expect* to ask?