MATH 323 - Doing your own Research

Tyrel Stokes

Email tyrel.stokes@mail.mcgill.ca for any questions about these problems. (See mycourses → content → Tyrel's Tutorial → Zoom link for office hourse and tutorial for the link)

1. It has been shown that the length of internet comments (in bytes) can be well approximated by a log-normal distribution. The length of a comment under a Youtube video for example can be approximately be modelled by $X \sim lognormal(\mu, \sigma)$, where $\mu \approx 4$ and $\sigma \approx 1$. 2 bytes corresponds to approximately 1 character. For reference, the original length of a tweet was 140 characters and thus approximately 280 bytes. See this paper for more details - `https://epjdatascience.springeropen.com/articles/10.1140/epjds14`

Here it is given that for X distributed lognormal with parameters μ and σ that the mean and variance are given by:

$$E[X] = exp(\mu + \sigma^2/2)$$
$$Var(X) = [\exp(\sigma^2) - 1] * (\exp(2\mu + \sigma^2)$$

a) Find the mean and standard deviation for the length of a Youtube Comment

b) Say you find yourself, regretfully, deep in the comments section under a random Youtube Video. You have convinced yourself there is something suspicious about the comments, maybe the the video producer is using a paid service to leave comments. You use a program to count the bytes of each comment and find the average over all 100 of the comments is 113.6. Suppose we want to know if in fact this is unusual for a random Youtube comment. Does 113.6 give evidence to suggest this is unusually long. A testable probability would be under the assumption that the comments come from the general distribution for Youtube comments above, what is the probability of observing a mean length of 113.6 or greater.

i) What additional assumption would we need to use an approximation to the probability above from the course?

ii) Under the assumption in i), what is the approximate probability?

iii) How reasonable are the assumptions in b.i? How do you interpret your findings in b.ii)?

2. Prove under the assumptions in b.i), that the sample mean X_n converges in probability to $E[X]$,

3. Suppose that it is known that comments left under BBC articles varies by article type. The average length of a comment under a Political Article is 600 bytes, under music articles it is 250, financial articles 500, and for other articles it is 400.

a) If it is known that political articles make up 25%, music 5%, and financial 15% of all articles what is the average length of a comment under a BBC article?

b) Given that an article is political, the standard deviation in length is 150. For music the standard deviation is 50. For Financial and Other articles the standard deviation is 100. What is the standard deviation of a comment under a BBC article?