DATA ANALYSIS PROJECT

RAVURU CHIDAKSH(200010046)
200010046@iitdh.ac.in

Department of Computer Science, IIT Dharwad

September 24, 2021

Contents

1 Approach towards Problem 2

We are given a dataset with 50000 samples in it, of the random variable Z. We also know that $Z = X + 10Y$ where $Y = \sum_{i=0}^{k} W_i$ and we know W_i are independent and identically distributed random variables. And we know k belongs to {2,3,4} and W_i belong to one of either Exponential distribution or Rayleigh distribution or Half-normal distribution.

1.1 Calculation of mean and varience:

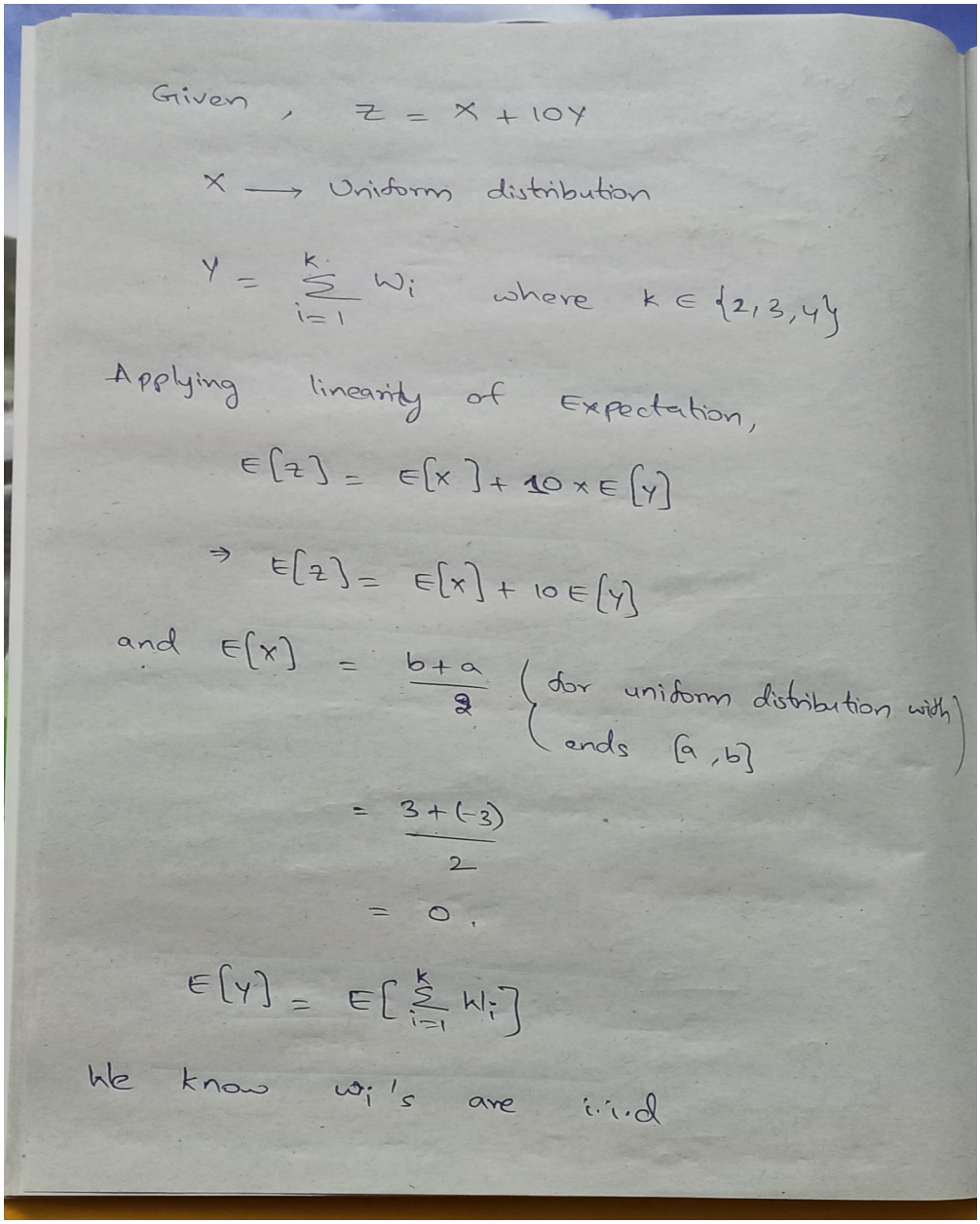
From the given data set we can calculate mean and varience of the entire data , i took the help of panda series objects and data frames to load the data from file and then took help of numpy inbuilt methods np.mean and np.var to calculate the mean and varience of the data.

1.2 Central Idea of the Approach:

since we have multiple values of k and multiple distributions we need to find the best possible k and distribution for the given data which means that we need to find that k and distribution such that the original mean and varience of data satisfies the relation between mean and varience of one of the distributions so closely for some k in {2,3,4}.

So we calculate mean for some k in {2,3,4} and the find varience (say var1) assuming some dirstribution and then we know actual varience from the data (say var2) and we will check for $min(var1 - var2)$ for every distribution for all k.

2 Images explaining the method



Given , $Z = X + 10Y$

$X \longrightarrow$ Uniform distribution

$Y = \sum_{i=1}^{k} W_i$ where $k \in \{2, 3, 4\}$

Applying linearity of Expectation,

$$E[Z] = E[X] + 10 \times E[Y]$$

$$\Rightarrow E[Z] = E[X] + 10\,E[Y]$$

and $E[X] = \frac{b+a}{2}$ (for uniform distribution with ends $[a, b]$)

$$= \frac{3+(-3)}{2}$$

$$= 0.$$

$$E[Y] = E\left[\sum_{i=1}^{k} W_i\right]$$

We know w_i's are i.i.d

⇒ let w_i's are random variables of W which follows one of raleigh, exponential or half normal distributions,

$$\Rightarrow E[Y] = E[w_1 + w_2 + \cdots + w_k]$$

$$= E[w_1] + E[w_2] + \cdots + E[w_k]$$

(all are $E[w]$)

$$= k \cdot E[w]$$

$$\therefore \boxed{E[Y] = k \cdot E[w]}$$

$$\Rightarrow E[z] = 0 + 10.k \cdot E[w]$$

$$\Rightarrow \boxed{E[w] = \frac{E[z]}{10k}} \rightarrow (1)$$

Similarly for varience,

$$\text{var}(z) = \text{var}(X + 10Y)$$

(X, Y are independent because if X, Y are dependent it would make ~~some~~ all of w_i's dependent which is a contradiction).

and hence X, Y are independent,

$$\text{var}(z) = \text{var}(x) + 100 \cdot \text{var}(y)$$

$$\downarrow$$

$$\frac{(b-a)^2}{12} \quad (b=3,\ a=-3)$$

$$\Rightarrow \text{var}(z) = \frac{(3-(-3))^2}{12} + 100 \cdot \text{var}(y)$$

$$\Rightarrow \text{var}(z) = 3 + 100 \cdot \text{var}(y)$$

$$\Rightarrow \boxed{\frac{\text{var}(z) - 3}{100} = \text{var}(y)} \quad \text{---} \textcircled{1}$$

but $\text{var}(y) = \text{var}(w_1 + w_2 + \cdots + w_k)$

($\because$ w_i's are i.i.d)

$$= \text{var}(w_1) + \text{var}(w_2) + \cdots + \text{var}(w_k)$$

$$= k\,\sigma^2_{(w)} \quad \text{(all are var(w))}$$

$$\therefore \boxed{\frac{\text{var}(z) - 3}{100 \cdot k} = \sigma^2(w)} \rightarrow \textcircled{2}$$

Now as we know $E[w]$, $\sigma^2[w]$ from data

⇒ For some k in [2, 3, 4] say (2) we can assume some distribution (say exponential) and calculate var[w] using the relation between mean and variance for that particular distribution. Say that variance be $\sigma^2[w_{new}]$.

by finding ratio between $\sigma^2[w_{new}]$ to $\sigma^2[w]$ for all 3 possible k's and 3 possible distributions, the case in which the ration is close to 1 the corresponding k and distribution is what we are looking for.

⇒ We can also take difference instead of ration and check for which combination it's zero.

3 Note:

- I wrote the code in python in Google Colab , which i'm submitting with extension "ipynb". So, to run the entire file , please upload the file on drive and open it with google colabaratory.
- I worked with .xls file in the entire process which i'm going to zip it with this pdf file . Please upload that .xls file on Google colab before running the cells.