DA Project Question 2

Hrishikesh Pable

September 2021

Roll No: 200010037

1 Question

Disturbing distributions: You have been given the data (50000 samples) of a random variable Z. You know that $Z = X + 10Y$, where X is a uniform random variable between -3 and 3. You also know that

$$Y = \sum_{x=1}^{k} W_i$$

where k∈2, 3, 4

W_i's are independent and identically distributed (i.i.d.) and belong to one of the following:

- Exponential distribution characterized by its mean $1/\lambda$
- Rayleigh distribution characterized by σ
- Half-normal distribution characterized by σ

Come up with a mechanism to find k and the distribution of W_i along with the characterizing parameter rounded to the nearest integer (mean if it is an exponential distribution and σ otherwise). Justify your mechanism analytically.

2 Approach

Since X is a uniform random variable, $E[X] = 0$.

The variance of X is $Var(X) = (b-a)^2/12 = 3$

Since the W_i's are independent and identically distributed (lets assume that the value of $E[W_i] = E[W]$ and $Var(W_i) = Var(W)$ for all i, since they are identically distributed), we can simplify the expression for expectation as:

$$Z = X + 10Y$$
$$E[Z] = E[X] + 10 * E[Y]$$
$$E[Z] = E[X] + 10 \sum_{i=1}^{k} E[W_i]$$
$$E[Z] = E[X] + 10k * E[W]$$
$$E[Z] = 0 + 10k * E[W]$$

$$E[W] = \frac{E[Z]}{10k}$$

Also, we can simplify variance as:

$$Var(Z) = Var(X) + 100 * Var(\sum_{i=1}^{k} W_i)$$
$$Var(Z) = Var(X) + 100k * Var(W)$$
$$Var(Z) = 3 + 100k * Var(W)$$
$$Var(W) = \frac{Var(Z) - 3}{100k}$$

Now, we'll calculate some parameters (Error, deviation coefficient and t) for each of the distribution for all values of k.

Let's assume that the distibution is Exponential; and find the mean and variance of the 5 datasets, which is $E[Z]$ and $Var(Z)$ respectively.From this we can calculate $E[W]$ and $Var[W]$ using the above two formulae that we derived, for each of the 5 datasets ; then we'll find the 5 values of lambda from the five means and 5 values of lambda from the 5 variances, using the relation between them i.e.:

$$\lambda = \frac{10k}{E[Z]}$$
$$\lambda = \sqrt{\frac{100k}{Var(Z) - 3}}$$

We then find the variance of these 5 lambdas obtained from the 5 means and the variance of the 5 lambdas obtained from the 5 variances; for all values of k i.e. 2, 3 and 4. Then we define deviation coefficient as

deviation coefficient = (variance of 5 values of lambda obtained from mean + variance of the 5 values of l

And then calculate the deviation coefficient for all cases (i.e. k=2,3,4) Lesser is the deviation coefficient, more accurate is the distribution.

We also define the Error as:

$$Error = \frac{Mean\ of\ 5\ values\ of\ lambda\ obtained\ from\ mean}{Mean\ of\ the\ 5\ values\ of\ lambda\ obtained\ from\ variance}$$

Closer the Error is to 1, more accurate is the distribution.

Lastly, we define t, which is similar to Error, and is given by :

$$t = \frac{The\ value\ of\ lambda\ obtained\ from\ mean\ of\ entire\ 50000\ samples}{The\ value\ of\ lambda\ obtained\ from\ variance\ of\ entire\ 50000\ samples}$$

So, we'll compare the values of Error, deviation coefficient, t, to find the value of k and the type of distribution of the given data. But, we'll give highest preference to Error, then t, then deviation coefficient.

We do the same procedure for Rayleigh distribution and half normal distribution; The relations for them are:

For Rayleigh:

$$\sigma = \frac{E[Z]}{10k}\sqrt{\frac{2}{\pi}}$$

$$\sigma = \sqrt{\frac{2(Var(Z) - 3)}{100k(4 - \pi)}}$$

For Half-Normal:

$$\sigma = \frac{E[Z]}{10k}\sqrt{\frac{\pi}{2}}$$

$$\sigma = \sqrt{\frac{\pi(Var(Z) - 3)}{100k(\pi - 2)}}$$

So, first we calculate the mean and variance for the 5 parts of the dataset and the whole dataset, which are the $E[Z]$ and $Var(Z)$; then using the above mentioned formulae, we calculate the value of the parameter in each case (i.e. lambda for Exponential and sigma otherwise), then for finding the deviation coefficient, we find the variance in the set of 5 lamdas obtained from 5 means , and the set of 5 lambdas obtained from 5 variances, and take their average.Smaller is the deviation, more accurate is the distribution. For finding the Error, we find

the mean of the 5 lambdas obtained from 5 means and the 5 lambdas obtained from 5 variances of the 5 parts of the data. Then we take their ratio. Closer is this ratio to 1, more accurate is the distribution.

Finally, t is just the ratio of the lambda obtained from mean of the entire dataset, and the lambda obtained from variance of the entire dataset.

Priority order is

1. Error
2. T value
3. deviation coefficient

So, we get the values as:

The deviation coefficient for Exponential distribution with k=2 is: 2.318905264062391e-06
The Error for Exponential distribution with k=2 is: 0.6170014165209059
The value of t is: 0.6170241573563913

The deviation coefficient for Exponential distribution with k=3 is: 3.5603996305509e-06
The Error for Exponential distribution with k=3 is: 0.7556693205253251
The value of t is: 0.7556971722469571

The deviation coefficient for Exponential distribution with k=4 is: 4.856588486677428e-06
The Error for Exponential distribution with k=4 is: 0.872571771247276
The value of t is: 0.6170241573563913

The deviation coefficient for Rayleigh distribution with k=2 is: 0.0003046248072512952
The Error for Rayleigh distribution with k=2 is: 3.388733910895743
The value of t is: 3.3886725156211983

The deviation coefficient for Rayleigh distribution with k=3 is: 0.0004917835845951684
The Error for Rayleigh distribution with k=3 is: 6.225501716820484
The value of t is: 6.225388926499044

The deviation coefficient for Rayleigh distribution with k=4 is: 0.0007904921904657012
The Error for Rayleigh distribution with k=4 is: 9.584786912124757
The value of t is: 9.584613260064906

The deviation coefficient for Half-Normal distribution with k=2 is: 0.0001725578464265367
The Error for Half-Normal distribution with k=2 is: 1.2244647457353415
The value of t is: 1.2244425615358117

The deviation coefficient for Half-Normal distribution with k=3 is: 0.00010878997099776477
The Error for Half-Normal distribution with k=3 is: 0.9997712783594439
The value of t is: 0.9997531650363769
The value of sigma is: 2.9893098419987667

The deviation coefficient for Half-Normal distribution with k=4 is: 7.924925576585358e-05
The Error for Half-Normal distribution with k=4 is: 0.8658273250333217
The value of t is: 0.8658116384353989

Hence from the Error, the value of t, and the deviation coefficient, we can conclude that it is a Half Normal distribution with k = 3 and sigma = 3