**Detection Problems in Assuming Gaussian Distribution for Sum of Independent Random variables**

Introduction – In estimation and detection theory many problems have closed form solution for gaussian distribution. We also know that distribution of the sum of independent and identically distributed random variables converges towards gaussian by the central limit theorem. The distribution of the random variables doesn’t matter. It is assumed that central limit theorem doesn’t take sum of too many IID random variables to converge towards a gaussian distribution. In nature, many times we only observe the sum of IID random variables. Hence, it is assumed that detection and estimation methods that work for normal distributions can be applicable to many problems involving other types of distributions. Due to this gaussian is used everywhere in industries like engineering, economics, statistics and sciences. Central limit theorem is considered central to the statistics and some people assume that name of the theorem comes from that.

But normal distribution is a very special type of distribution where probability of random variable goes down as a function of exponential squared. This means that probability of finding a random variable away from mean value is extremely low. 99.7% times random variable stays within 3 standard deviations (sigma) away from the mean and probability of 6sigma or 7sigma events is almost zero.

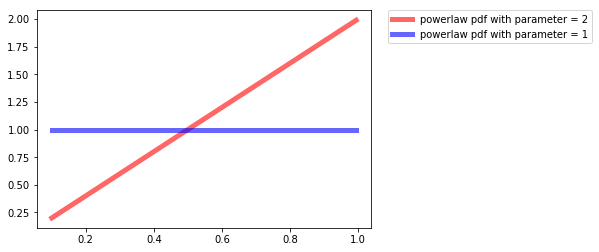
In this project, we will examine consequence of making the gaussian assumption in detection errors especially for rare events. We will look at the behavior of the following normalized sum random variable

According to the central limit theorem this random variable should converge towards the normal distribution with mean 0 and variance 1.

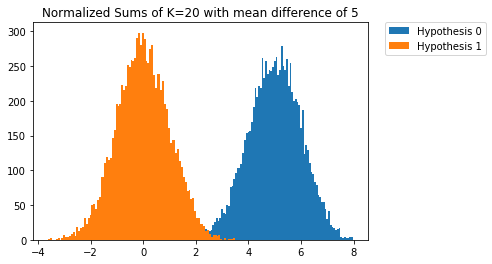
Experiments –

In all our experiments we assume that there are two separate hypotheses 0 and 1 which generate two different normalized sums. The normalized sum are separated from each other by adding a fixed mean difference to hypothesis 0. The detector is designed by assuming both hypothesis 0 and 1 behave like gaussian. We compare the empirical error and gaussian error of the detector.

Experiments with power law – In this set of experiments we assume that all Hypothesis 1 random variables are generated from a powerlaw distribution with parameter a = 1. This is same as uniform distribution. Hypothesis 0 random variables were generated from powerlaw distribution with parameter 0.9, 2, 4, 10 and 40. Figure 1 shows the powerlaw pdf with parameters 1 and 2. The higher parameter value in powerlaw makes the distribution heavy tail.

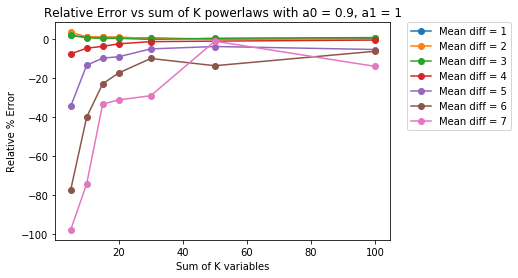


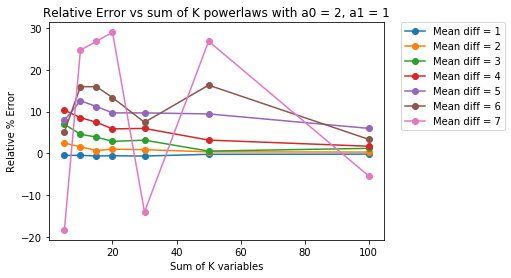
Our observations are the normalized sum of K random variables separated by a given mean difference. Figure 2 shows how the observations associated with different hypothesis look like. We use the threshold detector assuming gaussian distribution. After that we empirically try to determine the probability of error for this detector and compare it to the theoretical gaussian error. We do this comparison for different value of K (number of random variable that we sum) and mean difference.

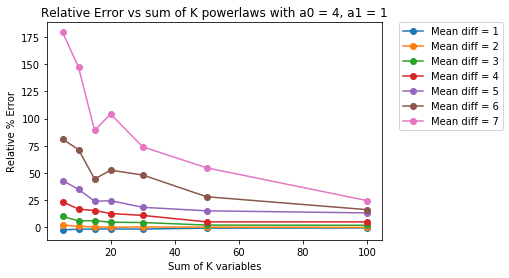


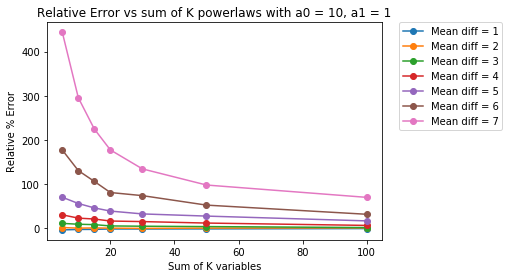
Results –

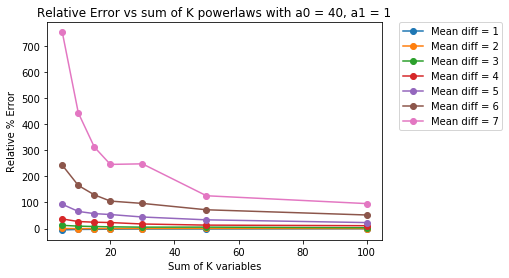
Figure 3-7 show the plot of relative error with respect to sum of K random variables in which hypothesis 0 has parameter a0 and hypothesis 1 has parameter a1 = 1.











As expected the relative error improves as we take the sum of larger number of random variables. What is unexpected is that Relative error is large when mean difference is large. Also, relative error is very high when powerlaw parameter is 40 as compared to when it is 2 or 4. Even with sum of 100 IID random variables the relative error stays large. The reasons for these are effect of tail probabilities. The relative error explodes as gaussian tail probability goes down much faster than actual tail probability. This behavior cannot be easily seen as low values of probabilities in both gaussian and non-gaussian distribution hide the actual difference.

Conclusion –

We have shown from our experiments that gaussian approximation for sum of independent and identically distributed variables only holds closer to the mean. These assumptions break down for tail probabilities. The central limit theorem only works closer to the mean, hence the name central. The use of gaussian distribution in detection of rare events or risk management can be highly flawed if underlying distribution has heavy tail. We think that gaussian distribution should not be used for failure analysis or any type of risk management even if the assumption of IID random variable holds.