Bazi Olasilik Dagilimlari 2

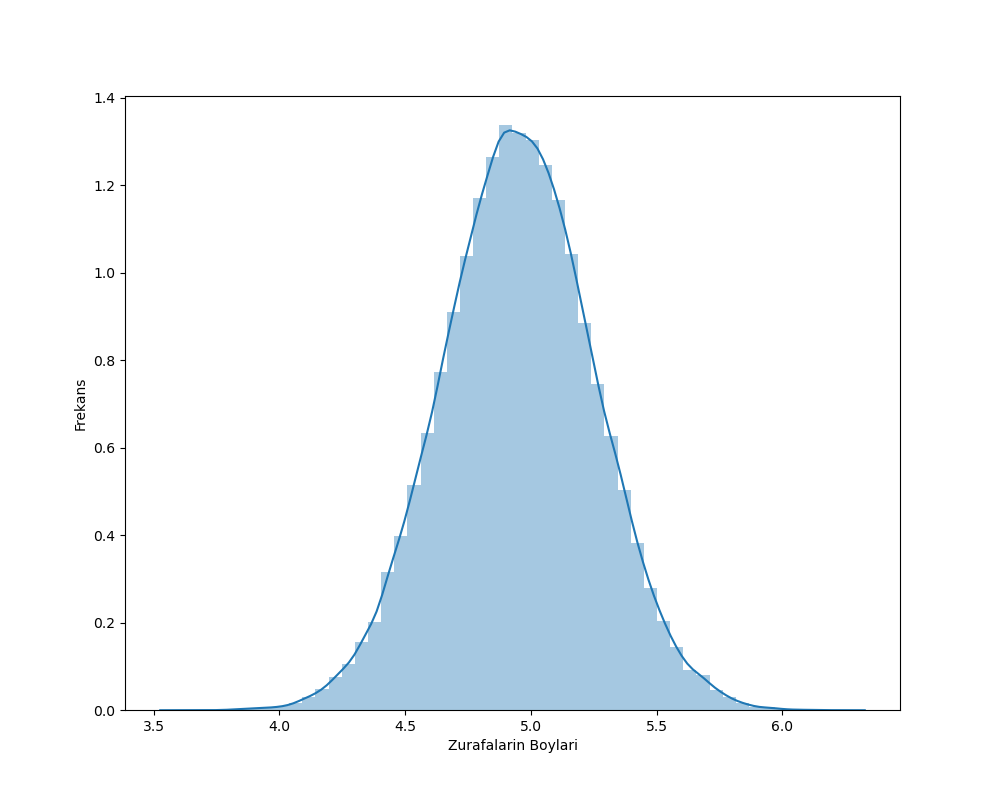
Bu bolumde surekli daglimlarla devam edecegiz.

1. Surekli Dagilimlar(Continuous Distribution): Sureksiz dagilimlarda verinin belirli bir ornek dagilimi vardi. Mesela bir zar atiyorduk ve 6 tane olasi sonucumuz vardi ve her sonucun arasindaki fark birbirine esit ve iki sonuc arasinda daha kucuk bir sonuc yoktu, Bu durum surekli dagilimlarda farklidir. Elimizdeki iki sonuc arasinda sonsuz tane daha sonuc mevcuttur. Bu yuzdende elimizdeki sonuclari tablo ile gostermemizin bir olanagi yoktur. Bunun yerine olasilik yogunluk fonksiyonu(Probability Density Function (PDF)) kullanilir. Cunku olasiligi hesaplarken o kadar cok olasi sonuc vardir ki, herhangi bir sonucun butun bu sonuclar icinde olmasi cok kucuk bir ihtimaldir ve yahut direkt 0 diyebiliriz. Bu bolumde ekstra olarak birikimli dagilim fonksiyonundan ve niye boyle bir seye ihtiyac duydugumuzdan da bahsedecegim. Hadi baslayalim.

Olasilik Yogunluk Dagilimi(PDF) : Ustte bahsettigim gibi, surekli dagilimlarda ornek uzayi sonsuz oldugu icin her olasi sonucun olasi sonucu 0 olacaktir. Bu yuzden herhangi bir sonucun olasiligini hesaplamak yerine bir aralik belirleyip o araligin olasiligini hesaplamamiz gerekecektir. Verilen araliklardaki olasiliklari artik toplamak icin integral kullanmamiz gerekecek. Bu ayrinti integralin tanimini hatirlamaniz icin iyi bir firsat 😊. Eger butun dagilimi integre edersek, elimizdeki sonuc 1 olacaktir.

Birimli Olasilik Dagilimi(Cumulative Distribution Function(CDF)) : Aslinda tanimindan da anlasilacagi uzere fonksiyonu cizerken onceki araliktaki olasiklari toplayarak cizmeye devam edilen bir fonksiyondur. Yani CDF aslinda PDF’in integralinden baska bir sey degildir. Eksi sonsuzdan herhangi bir sonuca kadar olan bir aralagin olasiligini bulmak istedigimizde CDF’i kullaniriz. CDF icin ozel ornekler verecegim.

1. Normal Dagilim(Normal Distribution) : Gaussian dagilimi, Gauss dagilimi veya Gauss-Laplace dagilimi olarak da literaturde isimlendirilmistir. Normal dagilim istatistikte en cok kullanilan dagilimdir diyebiliriz. Cunku aslinda normal dagilimin kendisi doganin bir parcasidir ve temeli rastgelilige dayanir. Mesela yetiskin bir zurafanin boyu 4.4 metre ve 5.5 metre arasinda degisiklik gosterir. Eger dunyadaki butun zurafalarin boylarini bir frekans grafigine koysaydik eger soyle bir sey olacakti  :



Rastgele sayilardan olusan her dagilim Gauss dagilimini olusturacaktir. Deney yaparken aldigimiz verileri eger histograma koyarsak yine normal dagilimi goruruz. Bunun sebebi ise rastgele gurultulerden kaynaklarin. Gurultu doganin bir parcasidir ve bundan kacamayiz demistik. O yuzden herhangi bir deney yapiyorsaniz bu dagilim sizin en yakin arkadasiniz olacaktir. Eger elinizdeki veri seti Normal dagilima sahip degil ise o zaman sistematik bir hataya sahipsiniz demektir ve bunu olabildigince duzeltmeye calismaniz gerekmektedir. Bu konuya ayrintili olarak ilerki yazilarimizda deginecegiz. Gauss dagiliminin olasilik dagilim fonksiyonu su sekildedir :

Burada ortalama veya beklenen degerdir. ise standart sapmadir. Simdi kod ile yukarida yaptigim zurafalarin boy dagilimini nasil yaptigimi gorelim.

import random

import matplotlib.pyplot as plt

import seaborn as sns

veri = []

for i in range(100000):

    veri.append(random.gauss(4.95, 0.3))

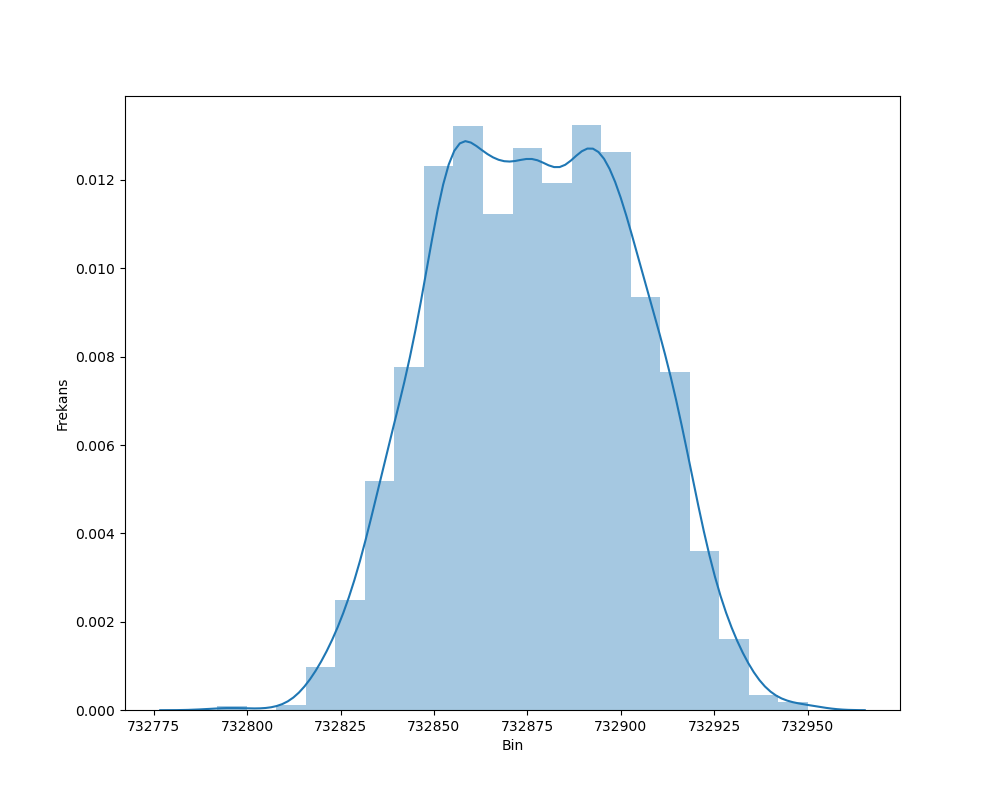
plt.figure(figsize = (10, 8))

ax = sns.distplot(veri, kde = True, bins = 50)

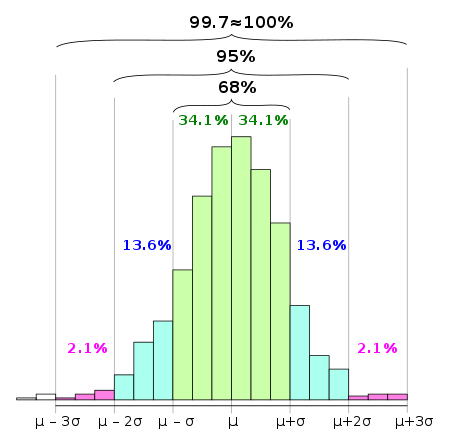
ax.set(xlabel = 'Zurafalarin Boylari', ylabel = 'Frekans')

plt.show()

Random modulunun kendine ait gauss rastegele sayisi olusturmak icin fonksiyonu var ve bu fonksiyonun ilk parametresi beklenen deger(ortalama deger), ikinci parametresi ise standart sapmadir. Bu olayi gercek bir veri setiyle de gostermek istiyorum. Elimde LHC DCCT’nin ne kadar stabil oldugunu gormek icin sabit bir akim gondererek aldigim veri seti var. Alinan veriyi grafik uzerinde gosterelim.



Neredeyse Gauss dagilimi gibi. Fakat tam olarak Gauss dagilimi degil. Burada ilerde uzerine deginecegimiz etki miktari(influence quantity) ayrik bir etkenin etkisi var. Kisaca olculen nicelik olmayan ama olcumun sonuclarina etki eden nicelik. Buradaki nicelik ise sicakliktir. Sabah saatlerinde sicaklik dusuk oldugu icin, olcum yapilan cihazlarin elektronigi aksam saatlerine dogru daha dusuktur. Olcum aldigim laboratuvarin gun boyu gunes gormesininde bu nicelige buyuk katkitisi oluyor ve gorundugu uzre olcumlere yansiyor. Ozetlemek gerekirse Gauss dagilimi, rastgele gelisen her sey icinde olan bir dagilimdir. Yani buna biz rastgele hata diyoruz. Doganin karakteristigi ise rastgeleliktir. Normal dagilimin standart sapmasi bize olctugumuz verinin ortalama degerden ne kadar saptiginin olcusunu verir. Bu yuzden normal dagilimin standart sapmasi cok onemlidir ve aslinda rastgele hatanin ve belirsizliklerin hesaplamasinda kullanilir. Alinan veriye gore; verinin %68‘lik kismini, verinin %95‘lik kismini ve verinin %99.7’lik kismini kapsar.



[1] Standart sapmanin veri seti uzerindeki dagilimi

Belirsizligi hesaplarken bir guvenlik araligi belirledigimiz zaman standart sapmayi kullancagiz. Burada derinlemesine incelemeyecegim bir dagilimdan daha bahsetmek istiyorum. Students’T dagilimi. Eger az sayida olcum yapilan bir veriyi bu dagilim ile inceliyoruz. Bu dagilim tipki normal dagilim gibidir. Istatistikte sikilikla kullanilan bu dagilim genelde sinirli veri ile hipotez testi yaparken kullanilir.

1. Ki-Kare Dagilimi(Chi-Squared Distribution) : Ki-Kara dagilimini anlatmadan once bir tane tanim yapmak istiyorum.
   * + Serbestlik Derecesi(Degrees of Freedom) : Degerlendirilmesi yapilacak niceligin kac tane birbirinden bagimsiz nicelige bagli oldugunu gosterir. n tane degiskenimiz var ise serbestlik derecesi = n – 1 olacaktir.Mesela elimizde bir diyet programimiz var. Bu diyet programina katilan kisilerin bir ayda ortalama kac kilogram kaybettiklerini ogrenmek istiyoruz. Eger bu diyete katilan kisilerden 4 kisiyi secersek, n = 4 olacaktir. Bu durumda serbestlik derecesi N = n – 1 = 3 olacaktir.

Ki-Kare dagilimini hipotez testlerinde, guven araligi hesaplamada ve en onemlisi fit yapilan veri setinin ne kadar iyi fit yapildigini gostermek icin kullanilir. Biz fit’in ne kadar iyi yapildigini ve guven araligini hesaplayacagiz. Baslamadan once dagilima bir de matematiksel olarak bakalim.

Burada gozlem sayisi, ise i’ninci terimin beklenen degeri. Hemen gra