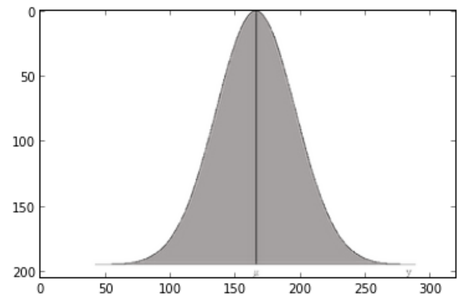
Bu notlar makine ogrenimi, veri madenciligi gibi konularda gerekli olasilik ve istatistik bilgisini paylasmak icin hazirlaniyor. Notlarda olasilik ve istatistik ayni anda anlatilacak, ve uygulamalara agirlik verilecek.

Dağılımlar Hakkında

Doğadan yapılan çoğu ölçümlerin, sıklık grafiğini alınca sonucun aşağıda gibi çıkmasi ilginctir.



Mesela, Türkiye'deki 2000 yetişkinin kilosunu ölçün. Grafiğini alın, kesinlikle yukarıdaki tepe şekli çıkacak. Ya da, 1000 kişinin boyunu ölçün, aynı tepe şekli. Keskin nişancının hedefe attığı kurşunların hedefe gelişini en iyi 12 en kötü 1 olmak üzere ölçün, sıklık grafiğini alın. Gene aynı tepe şekli!

Nasıl oluyor bu iş?

Açıklama için, normal dağılım eğrisinden bahsetmemiz gerekecek.

Not olarak düşelim: Sıklık grafiği, X sayısının ne kadar çıktığını sayıp, Y ekseni üzerinde bu sayıyı X'e tekabül ederek kolon olarak göstermeye denir. Mesela, 60 kilo değeri 13 kere çıktı ise, X=60, Y=13 gibi bir kolon çizilecektir.

Normal Dağılım Eğrisi

Normal dağılımın olasılık kavramı ile yakın bağları var. Bu konuda ünlü bir deney zar atma deneyidir. Mesela, elimizde tek bir zar olsun, ve bu zarı arka arkaya atalım. Sabrımız yeterse 1000 kere atalım. Sonuçta, sıklık grafiği eşit bir dağılım gösterecektir. (Zar tutmuyorsanız :))

Bunun sebeplerini anlamak zor değil. Her zar atış olayı birbirinden bağımsız, ve her sayının üstte gelme ihtimali birbirine eşit olduğu için (1/6), her sayıdan eşit miktarda gelecektir. Tabii bunun için deneyin birçok kere tekrarlanması gerekiyor.

Fakat, bir yerine 2 zar atalım. Hatta hatta, 4 zar atalım, ve bu sefer sıklık grafik hanesine yazmadan çıkan sayıları önce toplayalım. Bu çıkan toplamın sıklık grafiğini alalım.

İşte bu sıklık grafiği göreceğiz ki, ustte gorulen tepe grafiğine yaklaşıyor. Ne kadar çok zar atarsanız, bu benzerlik o kadar daha fazla olacaktır.

Bunun sebebi sezgisel olarak tahmin edilebilir, 1..6 arası sayıların tek bir zardan gelme olasılığı aynı, evet. Fakat toplamlara gelince, mesela iki zarlı örnekte, 10 sayısının olasılığı 2 sayısından daha yüksek. Çünkü, 10 sayısını 5-5, 4-6 ya da 6-4 ile alabiliyoruz. 2 sayısı sadece 1-1 ile geliyor.

Buradan şu sonuç cıkabilir: Eğer doğada ölçtüğümüz bir kavramın oluşmasında birden fazla etken var ise, o ölçümlerin sıklığı her zaman çan şekli ile olacaktır. Bir kisinin boyunu, kilosunu etkileyen pek cok diger faktor oldugu icin bu tek olcutleri dagilimlarinin normal ciktigi iddia edilebilir.

Toplamlarin dağılımının çan eğrisine yaklaşması durumu Istatistikte Merkezi Limit Teorisi ile ispatlanmistir.

Simulasyon

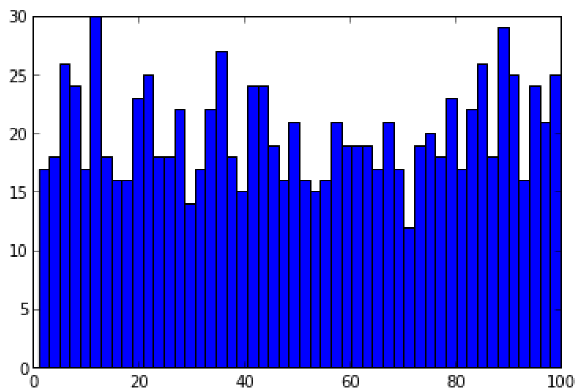
Eğer bu kavramları simulasyon ortamında göstermek istersek, Python ile bunu yapabiliriz.

İlk önce, Random.org sitesinden rasgele sayı üretip bilgiyarımıza kopyalacağız. Bahsettiğimiz site, kimsenin kullanmadığı radyo kanallarından atmosfer gürültüsü dinleyip, bu gürültüleri sayısal değere çevirerek rasgele sayı üretiyor.

Gerçek rasgele sayı üretmek pek kolay bir iş değil. Her ne kadar bilgisayarımızda rasgele sayı üreten birçok algoritma olsa bile, bu algoritmalar belli bir sayı üretiminden sonra kendini tekrar etmeye başlıyorlar. Gerçek rasgele sayılar için dış bir kaynağa bağlanmak bir secenek olabilir. Ama sunu da soylemek lazim, simulasyon tekniklerinin tamami icin yari-rasgele (pseudorandom) sayilar yeterlidir.

Siteden rasgele sayıları üretip, bir veri dosyasına koyuyoruz. Python ile bu sayıları okuyup, ilk önce teker teker sayıların sıklık grafiğini, ondan sonra sayıları üçer üçer toplayıp, onların grafiğini alıp göstereceğiz.

```
A = loadtxt('rasgele.dat')
plt.hist(A, 50)
plt.savefig('dagilim_1.png')
```

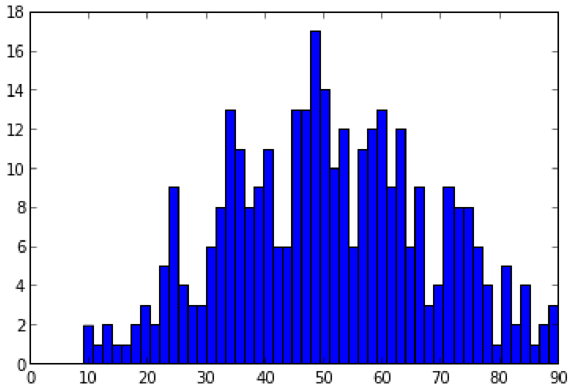



```
A = loadtxt('rasgele.dat');
B = []
```

```
i = 1;

while (i < 998):
  toplam = 0
  s = A[i]
  toplam = toplam + s
  s = A[i+1]
  toplam = toplam + s
  s = A[i+2]
  toplam = toplam + s
  B.append(toplam/3)
  i = i + 3

plt.hist(B, 50);
plt.savefig('dagilim_2.png')
```



Olasilik

Orneklem Uzayi (Sample Space)

Orneklem uzayi Ω bir deneyin mumkun tum olasiliksal sonuclarin (outcome) kumesidir. Eger deneyimiz ardi ardina iki kere yazi (T) tura (H) atip sonucu kaydetmek ise, bu deneyin mumkun tum sonuclari soyledir

$$\Omega = \{HH, HT, TH, TT\}$$

Sonuclar ve Olaylar (Outcomes and Events)

Ω icindeki her nokta bir sonuctur (outcome). Olaylar Ω'nin herhangi bir alt kumesidir ve sonuclardan olusurlar. Mesela ustteki yazi-tura deneyinde "iki atisin icinden ilk atisin her zaman H gelmesi olayi" boyle bir alt kumedir, bu olaya A diyelim, $A = \{HH, HT\}$.

Ya da bir deneyin sonucu ω fiziksel bir olcum , diyelin ki sicaklik olcumu. Sicaklik $\pm$, reel bir sayi olduguna gore, $\Omega = (-\infty, +\infty)$, ve sicaklik olcumunun 10'dan buyuk ama 23'ten kucuk ya da esit olma "olayi" $A = (10, 23]$. Koseli parantez kullanildi cunku sinir degerini dahil ediyoruz.

Ornek

10 kere yazi-tura at. A = "en az bir tura gelme" olayi olsun. T_j ise j'inci yazi-tura atisinda yazi gelme olayi olsun. $P(A)$ nedir?

Bunun hesabi icin en kolayi, hic tura gelmeme, yani tamamen yazi gelme olasiligini, A^c'yi hesaplamak, ve onu 1'den cikartmaktir. ^{c} sembolu "tamamlayici (complement)" kelimesinden geliyor.

$$P(A) = 1 - P(A^c)$$

$$= 1 - P(\textit{hepsi yazi})$$

$$= 1 - P(T_1)P(T_2)...P(T_{10})$$

$$= 1 - \left(\frac{1}{2}\right)^{10} \approx .999$$

Rasgele Degiskenler (Random Variables)

Bir rasgele degisken X bir eslemedir, ki bu esleme $X : \Omega \to \mathbb{R}$ her sonuc ile bir reel sayi arasindaki eslemedir.

Olasilik derslerinde bir noktadan sonra artik ornekleme uzayindan bahsedilmez, ama bu kavramin arkalarda bir yerde her zaman devrede oldugunu hic aklimizdan cikartmayalim.

Ornek

10 kere yazi-tura attik diyelim. VE yine diyelim ki $X(\omega)$ rasgele degiskeni her ω siralamasinda (sequence) olan tura sayisi. Iste bir esleme. Mesela eger ω = HHTHHTHHTT ise $X(\omega) = 6$. Tura sayisi eslemesi ω sonucunu 6 sayisina esledi.

Ornek

$\Omega = \{(x, y); x^2 + y^2 \leqslant 1\}$, yani kume birim cember ve icindeki reel sayilar (unit disc). Diyelim ki bu kumeden rasgele secim yapiyoruz. Tipik bir sonuc $\omega = (x, y)$'dir. Tipik rasgele degiskenler ise $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$ olabilir. Goruldugu gibi bir sonuc ile reel sayi arasinda esleme var. X rasgele degiskeni bir sonucu x'e eslemis, yani (x, y) icinden sadece x'i cekip cikartmis. Benzer sekilde Y, Z degiskenleri var.

Toplamsal Dagilim Fonksiyonu (Cumulative Distribution Function -CDF-)

Tanim

X rasgele degiskeninin CDF'i $F_X : \mathbb{R} \to [0, 1]$ tanimi

$$F_X(x) = P(X \geqslant x)$$

Eger X ayriksal ise, yani sayilabilir bir kume $\{x_1, x_2, ...\}$ icinden degerler aliyorsa olasilik fonksiyonu (probability function), ya da olasilik kutle fonksiyonu (probability mass function -PMF-)

$$f_X(x) = P(X = x)$$

Bazen f_X, ve F_X yerine sadece f ve F yazariz.

Tanim

Eger X surekli (continuous) ise, yani tum x'ler icin $f_X(x) > 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$ olacak sekilde bir f_X mevcut ise, o zaman her $a \leqslant b$ icin

$$P(a < X < b) = \int_a^b f_X(x)dx$$

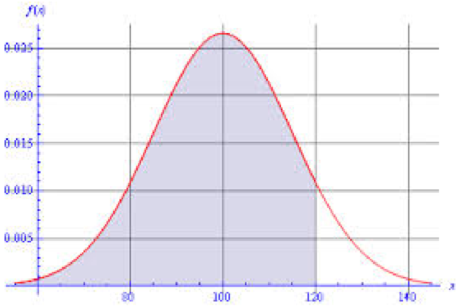
Bu durumda f_X olasilik yogunluk fonksiyonudur (probability density function -PDF-).

$$F_X = \int_{\infty}^{x} f_X(t)dt$$

Ayrica $F_X(x)$'in turevi alinabildigi her x noktasinda $f_X(x) = F'_X(x)$ demektir.

Dikkat! Eger X surekli ise o zaman $P(X = x) = 0$ degerindedir. $f(x)$ fonksiyonunu $P(X = x)$ olarak gormek hatalidir. Bu sadece ayriksal rasgele degiskenler icin isler. Surekli durumda olasilik hesabi icin belli iki nokta arasinda entegral hesabi yapmamiz gereklidir. Ek olarak PDF 1'den buyuk olabilir, ama PMF olamaz. PDF'in 1'den buyuk olabilmesi entegrali bozmaz mi? Unutmayalim, entegral hesabi yapiyoruz, noktasal degerlerin 1 olmasi tum 1'lerin toplandigi anlamina gelmez. Bakiniz *Entegralleri Nasil Dusunelim* yazimiz.

Olasilik degerleri, $P(a < X < b)$ ifadesi, alan hesabi ve rasgele degiskenler arasindaki baglantiyi biraz daha detaylandirmak gerekirse; X bir rasgele degisken, nokta (kesin) degeri olmasa da denklemde kullanilabiliyor, toplanip cikartilabiliyor, vs. Bu degiskene "degeri soruldugunda" bu deger o X'in bagli oldugu dagilimin zar atmasi sonucunda gelecektir. Bu zar atisi ise olasilik fonksiyonunun yuksek deger verdigi x degerlerini daha fazla uretecektir dogal olarak. Bunu kavramsal olarak soyluyoruz tabii, istatistiki problemlerde illa bu zar atisini yapmamiz gerekmeyebilir.



Mesela ustteki dagilim icin 100 ve cevresindeki degerlerinin olasiligi cok yuksek, mesela grafige bakarsak, kabaca, $f_X(100) = 0.027$, ya da $f_X(120) = 0.015$. Demek ki bu dagilima bagli bir X, o cevreden daha fazla deger uretir.

Rasgele degiskene bagli olasilik hesabi icin ise, mesela $P(X < 120)$ diyelim, bu ifade ile ne diyoruz? Sordugumuz sudur, zar atislarinin belli deger altinda gelmesi olasiligi... Bu hesap tabii ki bir alan hesabidir, x eksenindeki belli araliklar, bolgelerin toplam olasiliginin ne olacagi o bolgenin tam uzerindeki yogunlugun toplami olacaktir, aynen tekil degerlerin olasiliginin o degerin tekil yogunluk degeri olmasi gibi. Yani bu tur olasilik hesaplari direk $f_X(x)$ uzerinden yapilacaktir. Zar atildiginda 100'den kucuk degerlerin gelme olasiligi nedir? Alana bakarsak 0.5, yani 1/2, tum alanin yarisi. Bu normal, cunku 100'den kucuk degerler dagilimin yarisini temsil ediyor. 200'den kucuk degerler gelme olasiligi nedir, yani $P(X < 200)$? Olasilik 1. f_X alaninin tamami. Yani kesin. Cunku yogunluk fonksiyonunun tamami zaten 200'den kucuk degerler icin tanimli. "Yogunluk orada".

Tanim

X rasgele degiskeninin CDF'i F olsun. Ters CDF (inverse cdf), ya da ceyrek fonksiyonu (quantile function)

$$F^{-1}(q) = \inf \left\{ x : F(x) \leqslant q \right\}$$

ki $q \in [0, 1]$. Eger F kesinlikle artan ve surekli bir fonksiyon ise $F^{-1}(q)$ tekil bir x sayisi ortaya cikarir, ki $F(x) = q$.

Eger inf kavramini bilmiyorsak simdilik onu minimum olarak dusunebiliriz.

$F^{-1}(1/4)$ birinci ceyrek

$F^{-1}(1/2)$ medyan (median, ya da ikinci ceyrek),

$F^{-1}(3/4)$ ucuncu ceyrek

olarak bilinir.

Iki rasgele degisken X ve Y dagilimsal olarak birbirine esitligi, yani $X \overset{d}{=} Y$ eger $F_X(x) = F_Y(x)$, $\forall x$. Bu X, Y birbirine esit, birbirinin aynisi demek degildir. Bu degiskenler hakkindaki tum olasiliksal islemler, sonuclar ayni olacak demektir.

Uyari! "X'in dagilimi F'tir" beyanini $X \sim F$ seklinde yazmak bir gelenek. Bu biraz kotu bir gelenek aslinda cunku $\sim$ sembolu ayni zamanda yaklasiksallik kavramini belirtmek icin de kullaniliyor.

Dagilimlar

Bernoulli Dagilimi

X'in bir yazi-tura atisini temsil ettigini dusunelim. O zaman $P(X = 1) = p$, ve $P(X = 0) = 1 - p$ olacaktir, ki $p \in [0, 1]$ olmak uzere. O zaman X'in dagilimi Bernoulli deriz, $X \sim Bernoulli(p)$ diye gosteririz. Olasilik fonksiyonu, $x \in \{0, 1\}$.

$$f(x; p) = p^x (1 - p)^{(1-x)}$$

Yani x ya 0, ya da 1. Parametre p, 0 ile 1 arasindaki herhangi bir reel sayi.

Beklenti ve varyans

$$E(X) = p$$

$$Var(X) = p(1 - p)$$

Uyari!

X bir rasgele degisken; x bu degiskenin alabilecegi spesifik bir deger; p degeri ise bir **parametre**, yani sabit, onceden belirlenmis reel sayi. Tabii istatistiki problemlerde (olasilik problemlerinin tersi olarak dusunursek) cogunlukla o sabit parametre bilinmez, onun veriden hesaplanmasi, kestirilmesi gerekir. Her halukarda, cogu istatistiki modelde rasgele degiskenler vardir, ve onlardan ayri olarak parametreler vardir. Bu iki kavrami birbiriyle karistirmayalim.

Binom Dagilimi (Binomial Distribution)

Her biri birbirinden bagimsiz ve birbiriyle ayni Bernoulli Dagilimina sahip deneylerden n tane yapildigini farzedelim, ki bu deneylerin sadece iki sonucu olacak (1/0. basari/basarisizlik, vs). Bu deneylerin p'si ayni olacak. O zaman n deney icinden toplam kac tanesinin basarili oldugunu gosteren X rasgele degiskeni Binom Dagilimina sahiptir denir.

Bu dagilimin yogunlugu

$$f(x; p, n) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \frac{n!}{x!(n - x)!} p^x (1 - p)^{n-x}$$

Bu fonksiyonun parametreleri p, n degerleridir. Beklenti ve varyans

$$\mu = E(X) = np$$

$$\sigma^2 = Var(X) = np(1-p)$$

Duz (Uniform) Dagilim

X duz, $Uniform(a, b)$ olarak dagilmis deriz, ve bu $X \sim Uniform(a, b)$ olarak yazilir eger

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \ icin \\ 0 & digerleri \end{cases}$$

ise ve $a < b$ olacak sekilde. CDF hesabi olasilik egrisinin entegralini temel alir, duz dagilim bir a, b arasinda $1/b - a$ yuksekliginde bir dikdortgen seklinde olacagi icin, bu dikdortgendeki herhangi bir x noktasinda CDF dagilimi, yani o x'in baslayip sol tarafin alaninin hesabi basit bir dikdortgensel alan hesabidir, yani $x - a$ ile $1/b - a$'nin carpimidir, o zaman

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

Normal (Gaussian) Dagilim

$X \sim N(\mu, \sigma^2)$ ve PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \ x \in \mathbb{R}$$

ki $\mu \in \mathbb{R}$ ve $\sigma > 0$ olacak sekilde. Bazilari bu dagilimi

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-\mu)\sigma^{-2}(x-\mu) \right\}$$

olarak gosterebiliyor, cunku bu sekilde (birazdan gorecegimiz) cok boyutlu Gaussian formulu ile alaka daha rahat gozukuyor.

Ileride gorecegiz ki μ bu dagilimin "ortasi", ve σ onun etrafa ne kadar "yayildigi" (spread). Normal dagilim olasilik ve istatistikte cok onemli bir rol oynar. Dogadaki pek cok olay yaklasiksal olarak Normal dagilima sahiptir. Sonra gorecegimiz uzere, mesela bir rasgele degiskenin degerlerinin toplami her zaman Normal dagilima yaklasir (Merkezi Limit Teorisi -Central Limit Theorem-).

Eger $\mu = 0$ ve $\sigma = 1$ ise X'in standart Normal dagilim oldugunu soyleriz. Gelenege gore standart Normal dagilim rasgele degiskeni Z ile gosterilmelidir, PDF ve CDF $\phi(z)$ ve $\Phi(z)$ olarak gosterilir.

$\Phi(z)$'nin kapali form (closed-form) tanimi yoktur. Bu, matematikte "analitik bir forma sahip degil" demektir, formulu bulunamamaktadir, bunun sebebi ise Normal PDF'in entegralinin analitik olarak alinamiyor olusudur.

Bazi faydali puf noktalari

1. Eger $X \sim N(\mu, \sigma^2)$ ise, o zaman $Z = (X - \mu)/\sigma \sim N(0, 1)$.

2. Eger $Z \sim N(0, 1)$ ise, o zaman $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

3. Eger $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, ...$ ve her X_i digerlerinden bagimsiz ise, o zaman

$$\sum_{i=1}^{n} X_i = N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma^2\right)$$

Tekrar $X \sim N(\mu, \sigma^2)$ alirsak ve 1. kuraldan devam edersek / temel alirsak su da dogru olacaktir.

$$P(a < X < b) = ?$$

$$= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Ilk gecisi nasil elde ettik? Bir olasilik ifadesi $P(\cdot)$ icinde esitligin iki tarafina ayni anda ayni toplama, cikarma operasyonlarini yapabiliriz.

Son ifadenin anlami sudur. Eger standart Normal'in CDF'ini hesaplayabiliyorsak, istedigimiz Normal olasilik hesabini yapabiliriz demektir, cunku artik X iceren bir hesabin Z'ye nasil tercume edildigini goruyoruz.

Tum istatistik yazilimlari $\Phi(z)$ ve $\Phi(z)^{-1}$ hesabi icin gerekli rutinlere sahiptir. Tum istatistik kitaplarinda $\Phi(z)$'nin belli degerlerini tasiyan bir tablo vardir. Ders notlarimizin sonunda da benzer bir tabloyu bulabilirsiniz.

Ornek

$X \sim N(3, 5)$ ise $P(X > 1)$ nedir? Cevap:

$$P(X > 1) = 1 - P(X < 1) = 1 - P(Z < \frac{1-3}{\sqrt{5}})$$

$$= 1 - \Phi(-0.8944) = 1 - 0.19 = .81$$

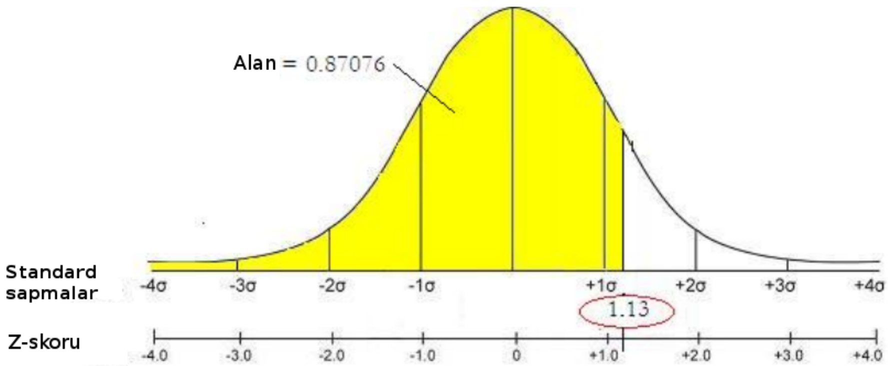
Soru $P(a < X < b)$ formunda a kullanmadi, sadece b oldugu icin yukaridaki form ortaya cikti. Python ile

```
from scipy.stats.distributions import norm
print norm.cdf(-0.8944)
print 1-norm.cdf(-0.8944)
```

```
0.18555395624
0.81444604376
```

Soru

$\Phi(1.13)$ nedir?



z-skoruna bakmanin bir diger yolu o degerin "kac standart sapma uzakta" oldugunu gosterilmesidir. Yani olcumuz standart sapma, ve bu deger sola ya da saga cekildikce ona tekabul eden alan (ustte sari renkle gosterilen kisim), yani olasilik azalip cogaliyor. Grafikte mesela "1.13 standart sapma" yani z-skor nereyi gosteriyor deyince, gorulen sekil / olasilik ortaya cikiyor. Tabii dagilim standart dagilim ve standart sapma 1 oldugu icin "kac standart sapma" ile z-skor arasinda direk bir baglanti var.

Ornek

Simdi oyle bir q bul ki $P(X < q) = .2$ olsun. Yani $\Phi^{-1}(.2)$'yi bul. Yine $X \sim N(3, 5)$.

Cevap

Demek ki tablodan .2 degerine tekabul eden esik degerini bulup, ustteki formul uzerinden geriye tercume etmemiz gerekiyor. Normal tablosunda $\Phi(-0.8416) = .2$,

$$.2 = P(X < q) = P(Z < \frac{q - \mu}{\sigma}) = \Phi(\frac{q - \mu}{\sigma})$$

O zaman

$$-0.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}}$$

$$q = 3 - 0.8416\sqrt{5} = 1.1181$$

Gamma Dagilimi

Y rasgele degiskeninin, verilmis $r > 0$ ve $\lambda > 0$ uzerinden Gamma yogunluk fonksiyonuna sahip oldugu soylenir, eger bu fonksiyon

$$f_Y = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{\lambda y}$$

$$y > 0$$

Peki Γ sembolu nerede geliyor? Bu bir fonksiyondur; Herhangi bir $r > 0$ icin Gamma fonksiyonu $\Gamma(r)$ su sekilde gosterilir,

$$\Gamma(r) = \int_0^\infty y^{r-1} e^{-y} dy$$

olarak tanimli ise.

Eger Y Gamma olarak dagilmis ise, beklenti $E(Y) = r/\lambda$, ve $Var(Y) = r/\lambda^2$.

Iki Degiskenli Dagilimlar

Tanim

Surekli ortamda (X, Y) rasgele degiskenleri icin yogunluk fonksiyonu $f(x, y)$ tanimlanabilir eger i) $f(x, y) > 0$, $\forall (x, y)$ ise, ve ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ ise ve her kume $A \subset \mathbb{R} \times \mathbb{R}$ icin $P((X, Y) \in A) = \int\int_A f(x, y) dx dy$. Hem ayriksal hem surekli durumda birlesik (joint) CDF $F_{X,Y}(x, y) = P(X \leqslant x, Y \leqslant y)$ diye gosterilir.

Bu tanimda A kumesi olarak tanimlanan kavram uygulamalarda bir olaya (event) tekabul eder. Mesela

Ornek

(X, Y)'in birim kare uzerinde duz (uniform) olsun. O zaman

$$f(x, y) = \begin{cases} 1 & \textit{eger } 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1 \text{ ise} \\ 0 & \textit{diger durumlarda} \end{cases}$$

$P(X < 1/2, Y < 1/2)$'yi bul.

Cevap

Burada verilen $A = \{X < 1/2, Y < 1/2\}$ bir altkumedir ve bir olaydir. Olaylari boyle tanimlamamis miydik? Orneklem uzayinin bir altkumesi olay degil midir? O zaman f'i verilen altkume uzerinden entegre edersek, sonuca ulasmis oluruz.

Ornek

Eger dagilim kare olmayan bir bolge uzerinden tanimliysa hesaplar biraz daha zorlasabilir. (X, Y) yogunlugu

$$f(x,y) = \begin{cases} cx^2y & eger\ x^2 \leqslant y \leqslant 1 \\ 0 & digerleri \end{cases}$$

Niye c bilinmiyor? Belki problemin modellemesi sirasinda bu bilinmez olarak ortaya cikmistir. Olabilir. Bu degeri hesaplayabiliriz, cunku $f(x, y)$ yogunluk olmali, ve yogunluk olmanin sarti $f(x, y)$ entegre edilince sonucun 1 olmasi.

Once bir ek bilgi uretelim, eger $x^2 \leqslant 1$ ise, o zaman $-1 \leqslant x \leqslant 1$ demektir. Bu lazim cunku entegrale sinir degeri olarak verilecek.

$$1 = \int\int f(x,y)dydx = c\int_{-1}^{1}\int_{x^2}^{1} x^2y$$

$$= c\int_{-1}^{1} x^2 \int_{x^2}^{1} ydydx = \int_{-1}^{1} x^2(\frac{1}{2} - \frac{x^4}{2})dx = 1$$

$$= c\int_{-1}^{1} x^2(\frac{1-x^4}{2})dx = 1$$

$$= \frac{c}{2}\int_{-1}^{1} x^2 - x^6 dx = 1$$

Devam edersek $c = 21/4$ buluruz.

Simdi, diyelim ki bizden $P(X \geqslant Y)$'yi hesaplamamiz isteniyor. Bu hangi A bolgesine tekabul eder? Elimizdekiler

$$-1 \leqslant x \leqslant 1,\ x^2 \leqslant y,\ y \leqslant 1$$

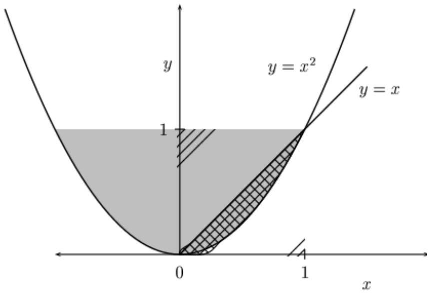
Simdi bunlara bir de $y \leqslant x$ eklememiz lazim. Yani ortadaki esitsizlige bir oge daha eklenir.

$$-1 \leqslant x \leqslant 1$$

$$x^2 \leqslant y \leqslant x$$

$$y \leqslant 1$$

$x^2 \leqslant y$'yi hayal etmek icin $x^2 = y$'yi dusunelim, bu bir parabol olarak cizilebilir, ve parabolun ustunde kalanlar otomatik olarak $x^2 \leqslant y$ olur, bu temel irdelemelerden biri.



Ayni sekilde $y \leqslant x$ icin $y = x$'i dusunelim, ki bu 45 derece aciyla cizilmis duz bir cizgi. Cizginin alti $y \leqslant x$ olur. Bu iki bolgenin kesisimi yukaridaki resimdeki golgeli kisim.

Ek bir bolge sarti $0 \leqslant x \leqslant 1$. Bu sart resimde bariz goruluyor, ama cebirsel olarak bakarsak $y \geqslant x^2$ oldugunu biliyoruz, o zaman $y \geqslant 0$ cunku x^2 muhakkak bir pozitif sayi olmali. Diger yandan $x \geqslant y$ verilmis, tum bunlari yanyana koyarsak $x \geqslant 0$ sarti ortaya cikar.

Artik $P(X \geqslant Y)$ hesabi icin haziriz,

$$P(X \geqslant Y) = \frac{21}{4}\int_0^1 \int_{x^2}^{x} x^2 y \, dy \, dx = \frac{21}{4}\int_0^1 x^2 \left[\int_{x^2}^{x} y \, dy \right] dx$$

$$= \frac{21}{4}\int_0^1 x^2 \frac{x^2 - x^4}{2} dx = \frac{3}{20}$$

"Hafizasiz" Dagilim, Ustel (Exponential) Dagilim

Ustel dagilimin hafizasiz oldugu soylenir. Bunun ne anlama geldigini anlatmaya ugrasalim. Diyelim ki rasgele degisken X bir aletin omrunu temsil ediyor, yani bir $p(x)$ fonksiyonuna bir zaman "sordugumuz" zaman bize dondurulen olasilik, o aletin x zamani kadar daha islemesinin olasiligi. Eger $p(2) = 0.2$ ise, aletin 2 yil daha yasamasinin olasiligi 0.2.

Bu hafizasizligi, olasilik matematigi ile nasil temsil ederiz?

$$P(X > s + t \,|X > t) = P(X > s), \; \forall s, \; t \geqslant 0$$

Yani oyle bir dagilim var ki elimizde, $X > t$ bilgisi veriliyor, ama (kalan) zamani hala $P(X > s)$ olasiligi veriyor. Yani t kadar zaman gectigi bilgisi hicbir seyi degistirmiyor. Ne kadar zaman gecmis olursa olsun, direk s ile gidip ayni olasilik hesabini yapiyoruz.

Sartsal (conditional) formulunu uygularsak ustteki soyle olur

$$\frac{P(X > s + t, X > t)}{P(X > t)} = P(X > s)$$

ya da

$$P(X > s + t, X > t) = P(X > s)P(X > t)$$

Bu son denklemin tatmin olmasi icin X ne sekilde dagilmis olmalidir? Ustteki denklem sadece X dagilim fonksiyonu ustel (exponential) olursa mumkundur, cunku sadece o zaman

$$e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t}$$

gibi bir iliski kurulabilir.

Ornek

Diyelim ki bir bankadaki bekleme zamani ortalam 10 dakika ve ustel olarak dagilmis. Bir musterinin i) bu bankada 15 dakika beklemesinin ihtimali nedir? ii) Bu musterinin 10 dakika bekledikten sonra toplam olarak 15 dakika (ya da daha fazla) beklemesinin olasiligi nedir?

Cevap

i) Eger X musterinin bankada bekledigi zamani temsil ediyorsa

$$P(X > 15) = e^{-15 \cdot 1/10} = e^{-3/2} \approx 0.223$$

ii) Sorunun bu kismi musteri 10 dakika gecirdikten sonra 5 dakika daha gecirmesinin olasiligini soruyor. Fakat ustel dagilim "hafizasiz" oldugu icin kalan zamani alip yine direk ayni fonksiyona geciyoruz,

$$P(X > 5 >= e^{-5 \cdot 1/10} = e^{-1/2} \approx 0.60$$

Bilesen (Marginal) Dagilimlar

Surekli rasgele degiskenler icin bilesen yogunluk

$$f_X(x) = \int f(x, y) dx$$

ve

$$f_Y(y) = \int f(x,y)dy$$

Ustteki integraller gercek bir dagilim fonksiyonu $f(x,y)$ verilince alt ve ust limit te tanimlamak zorundadir. Cunku bilesen yogunluk icin bir veya daha fazla degiskeni "integralle disari atmak (integrate out)" ettigimiz soylenir, eger ayriksal (discrete) ortamda olsaydik bu atilan degiskenin tum degerlerini goze alarak toplama yapan bir formul yazardik. Surekli ortamda integral kullaniyoruz, ama tum degerlerin uzerinden yine bir sekilde gecmemiz gerekiyor. Iste alt ve ust limitler bunu gerceklestiriyor. Bu alt ve ust limitler, atilan degiskenin "tum degerlerine" bakmasi gerektigi icin $-\infty, +\infty$ olmalidir. Eger problem icinde degiskenin belli degerler arasinda oldugu belirtilmis ise (mesela alttaki ornekte $x > 0$) o zaman entegral limitleri alt ve ust sinirini buna gore degistirebilir.

Ornek

$f_{X,Y}(x,y) = e^{-(x+y)}$, olsun ki $x, y \geqslant 0$. O zaman $f_X(x)$

$$f_X(x) = e^{-x}\int_0^\infty e^{-y}dy = e^{-x}\cdot 1 = e^{-x}$$

Ornek

$$f(x,y) = \begin{cases} x+y & eger\ 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1 \\ 0 & diger \end{cases}$$

$$f_Y(y) = \int_0^1 (x+y)dx = \int_0^1 x dx + \int_0^1 y dx = \frac{1}{2} + y \tag{1}$$

Tanim

Iki rasgele degisken A, B bagimsizdir eger tum A, B degerleri icin

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

esitligi dogru ise. Bu durumda X II Y yazilir.

Teori

X, Y'nin birlesik PDF'i $f_{X,Y}$ olsun. O zaman ve sadece $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ ise X II Y dogrudur.

Ornek

Diyelim ki X, Y bagimsiz, ve ikisinin de ayni yogunlugu var.

$$f(x) = \begin{cases} 2x & eger\ 0 \leqslant x \leqslant 1 \\ 0 & digerleri \end{cases}$$

$P(X+Y<1)$'i hesaplayin.

Cevap

Bagimsizligi kullanarak birlesik dagilimi hesaplayabiliriz

$$f(x,y) = f_X(x) f_Y(y) = \begin{cases} 4xy & eger\ 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1 \\ 0 & digerleri \end{cases}$$

Simdi bu birlesik yogunluk uzerinden istedigimiz bolgeyi hesaplariz, bolgeyi tanimlayan $X+Y \leqslant 1$ ifadesi.

$$P(X+Y \leqslant 1) = \int\int_{x+y \leqslant 1} f(x,y) dy dx$$

Entegralin limiti ustteki hali sembolik, hesap icin bu yeterli degil, eger $x+y \leqslant 1$ ise, $y \leqslant 1-x$ demektir, ve bolge $y = 1-x$ cizgisinin alti olarak kabul edilebilir. x, y zaten sifirdan buyuk olmali, yani sola dogru yatik cizginin alti ve y, x eksenlerinin ustu kismini olusturan bir ucgen,

$$= \int_0^1 \int_0^{1-x} 4yx \ dy dx = 4 \int\int_0^1 x \left[\int_0^{1-x} y dy \right] dx$$

Numaraya dikkat, hangi degisken uzerinden entegral aldigimiza bakarak, onun haricindekileri sabit kabul ederek bu "sabitleri" entegral disina atiyoruz, boylece isimizi kolaylastiriyoruz. Hesabi tamamlarsak,

$$4 \int_0^1 x \frac{(1-x)^2}{2} dx = \frac{1}{6}$$

Kosullu Dagilimlar (Conditional Distributions)

Surekli rasgele degiskenler icin kosullu olasilik yogunluk fonksiyonlari

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Devam edelim, eger kosullu yogunluk uzerinden olay hesabi yapmak istersek, ve $f_Y(y) > 0$ oldugunu farzederek,

$$P(X \in A | Y = y) = \int_A f_{X|Y}(x|y) dx$$

Ornek

(1) sonucunu aldigimiz ornege donelim,

$$f(x,y) = \begin{cases} x+y & eger\ 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1 \\ 0 & diger \end{cases}$$

$P(X < 1/4 | Y = 1/3)$ nedir?

Cevap

Ustteki olasilik hesabi icin $f_{X|Y}$ fonksiyonuna ihtiyacimiz var. (1)'de gordugumu uzere,

$$f_Y(y) = \frac{1}{2} + y$$

Ana formulumuz neydi?

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$= \frac{x+y}{\frac{1}{2}+y}$$

$$P(X < 1/4|Y = 1/3) = \int_0^{1/4} \frac{x + \frac{1}{3}}{\frac{1}{2} + \frac{1}{3}} dx = \frac{\frac{1}{32} + \frac{1}{3}}{\frac{1}{2} + \frac{1}{3}} = \frac{14}{32}$$

Cok Degiskenli (Multivariate) Dagilimlar ve IID Orneklemler (Samples)

$X = (X_1, ..., X_n)$ olsun, ki $(X_1, ..., X_n)$'lerin herbiri bir rasgele degisken, o zaman X'e rasgele vektor (random vector) ismi verilir. $f(x_1, ..., x_n)$'in PDF'i temsil ettigini dusunelim. Bu PDF'i baz alarak aynen iki degiskenli (bivariate) orneklerde oldugu gibi, benzer tekniklerle bilesenleri, kosullu dagilimlari, vs. hesaplamak mumkundur.

Cok Degiskenli Normal

Tek degiskenli Normal dagilimin iki parametresi vardi, μ, σ. Cok degiskenli formda μ bir vektor, σ yerine ise Σ matrisi var. Once rasgele degiskeni tanimlayalim,

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix}$$

ki $Z_1, ..., Z_k \sim N(0,1)$. Z'nin yogunlugu

$$f(z) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{k} z_j^2 \right\}$$

$$= \frac{1}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} z^T z \right\}$$

Bu durumda Z'nin *standart* cok degiskenli Normal dagilima sahip oldugu soylenir, ve $Z \sim N(0, I)$ olarak gosterilir. Buradaki 0 degeri icinde k tane sifir olan bir vektor olarak, I ise $k \times k$ birim (identity) matrisi olarak anlasilmalidir.

Daha genel olarak bir vektor X'in cok degiskenli Normal dagilimina sahip oldugunu soyleriz, ve bunu $X \sim N(\mu, \Sigma)$ olarak gosteririz, eger dagilimin yogunlugu

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp \left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

Σ pozitif kesin (positive definite) bir matristir. Hatirlayalim, bir matris pozitif kesindir eger tum sifir olmayan x vektorleri icin $x^T \Sigma x > 0$ ise.

Not: Karekok kavrami tekil sayilardan matrislere de aktarilabilir. Bir matris B'nin A'nin karekoku oldugu soylenir, eger $B \cdot B = A$ ise.

Devam edersek, eger Σ pozitif kesin ise bir $\Sigma^{1/2}$ matrisini oldugu gosterilebilir, ki bu matrise Σ'nin karekoku ismi verilir, ve bu karekokun su ozellikleri vardir, (i) $\Sigma^{1/2}$ simetriktir, (ii) $\Sigma = \Sigma^{1/2}\Sigma^{1/2} = I$ ve $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

Hatirlama Numarasi

Normal Dagilimin formulunu bazen hatirlayamayabiliriz. Peki daha basit bir formulden baslayarak onu turetebilir miyiz? Bu mumkun. Daha once e^{-x^2} *Nasil Entegre Edilir?* yazisinda

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

oldugunu gormustuk. Dikkat edersek bu integral bir formulun olasiliksal dagilim olup olmadigini kontrol etmek icin kullandigimiz integrale benziyor. Eger integral 1 cikarsa onun olasiliksal dagilim oldugunu biliyoruz. Ustteki sonuc $\sqrt{\pi}$, fakat iki tarafi $\sqrt{\pi}$'ye bolersek, sag taraf 1 olur ve boylece solda bir dagilim elde ederiz. Yani

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{-x^2} dx = 1$$

formulunde integralin sagindaki kisim bir dagilimdir. Bu formulu donusturerek Gaussian'a erisebiliriz. Ustteki formulun orta noktasi (mean) sifir, varyansi (variance), yani $\sigma^2 = 1/2$ (bunu da ezberlemek lazim ama o kadar dert degil). O zaman $\sigma = 1/\sqrt{2}$.

Ilk amacimiz $\sigma = 1$'e erismek olsun (cunku oradan herhangi bir σ'ya atlayabiliriz), bunun icin x'i $\sqrt{2}$'e bolmek lazim, tabii ayni anda onun etkisini sifirlamak icin normalize eden sabiti dengelemek amaciyla $\sqrt{2}$'ye bolmek lazim,

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(\frac{x}{\sqrt{2}})^2} dx$$

$\sigma = 1$'e erisince oradan herhangi bir σ icin, σ degiskenine bolelim, yine hem e ustune hem sabite bu eki yapalim,

$$= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x}{\sqrt{2}\sigma})^2} dx$$

Simdi herhangi bir ortalama μ icin bu degiskeni formule sokalim, bunun icin μ'yu x'den cikarmak yeterli

$$= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2} dx$$

/

e ustundeki kare alma islemini acarsak,

$$= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Boylece integral icindeki kisim tek boyutlu Gaussian formuna erismis oluyor.

Rasgele Degiskenler, Yogunluklar

Simdi konularin uzerinden bir daha gecelim; rasgele degisken, X, Y gibi buyuk harflerle gosterilen buyuklukler "bir zar atis sonucu icleri doldurulan" degiskenlerdir. Bu zar atisi her zaman X'in, Y'nin bagli oldugu dagilima gore olacaktir. Eger $X \sim N(10, 2)$ ise, bir formulun / hesabin icinde X gordugumuz zaman cogunlukla o noktaya 10'a yakin degerler olacagini biliriz. Tabii ki "kesin" her zaman ne olacagini bilmeyiz, zaten bir modelde noktasal deger (tipik cebirsel degiskenler) yerine rasgele degisken kullanmanin sebeplerinden biri budur.

Rasgele degiskenlerin matematiksel formullerde kullanilmasi $Z = X + Y$ seklinde olabilir mesela. O zaman elde edilen yeni degisken de bir rasgele degisken olur. Bu tur formuller envai sekle girebilir, hatta rasgele degisken iceren formullerin turevi bile alinabiliyor, tabii bunun icin ozel bir Calculus gerekli, Ito'nun Calculus'u bu tur islerle ugrasiyor.

Elimizde sunlar var; olasilik fonksiyonu bir matematiksel denklem, one degerler geciyoruz, ve bu degerlerin olasiliklarini gayet direk, mekanik bir formulden cevap olarak aliyoruz. Rasgele degiskenler ise bu yogunluk fonksiyonlarini bir anlamda "tersten isletiyor", o dagilima "zar attiriyor", ve olasilik fonksiyonuna gecilen degerler bu sefer disari cikiyor. Tabii yogunlugun ne olduguna gore bazi degerler daha cok, bazilari daha az cikiyor. Hesapsal olarak bir rasgele degiskene / dagilima zar attirmak icin ozel kodlamalar, yari-rasgele sayi uretimi gereklidir, biz kavramsal ve cebirsel olarak onlarin neyi temsil ettiginden bahsediyoruz.

Iki kavramdan daha bahsetmek bu noktada faydali. 1) Nufus (Population) 2) Orneklem (Sample). Nufus, uzerinde istatistiksel analiz yaptigimiz kitlenin tamami. Eger insanlarin boylari hakkinda istatistiki analiz yapiyor olsaydik tum insanlar nufus olurdu. Nufusun bazen hangi dagilimda oldugu bilinmiyor olabilir, biliniyor olsa da bazen bu dagilimin parametreleri bilinmiyor olabilir. Orneklem, nufus icinden alinan rasgele olcumlere verilen isimdir, $X_1, .., X_n$ olarak gosterilebiliyor, bu durumda nufusun dagiliminin "zar attigi" ve her zar atisinin rasgele degiskenlerden birinin icini doldurdugu dusunebilir. Orneklem nufustan geldigi icin dagiliminin aynen nufus gibi oldugu kabul edilir. Bu baglantidan yola cikilarak bircok istatistiki analiz yapmak mumkundur.

Kaynaklar

[1] http://en.wikipedia.org/wiki/Confidence_interval

[2] Janert, P., Data Analysis with Open Source Tools

[3] Introduction to Probability Models, Sheldon Ross, 8th Edition, sf. 273