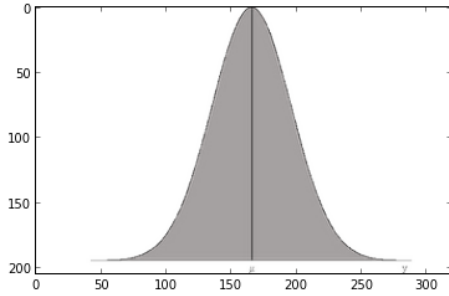


Bu notlar makine öğrenimi, veri madenciliği gibi konularda gerekli olasılık ve istatistik bilgisini paylaşmak için hazırlanıyor. Notlarda olasılık ve istatistik aynı anda anlatılacak, ve uygulamalara ağırlık verilecek.

Dağılımlar Hakkında

Doğadan yapılan çoğu ölçümlerin, sıklık grafiğini alınca sonucun aşağıda gibi çıkması ilginçtir.



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 ekseninde 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, üstte görülen 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 çıkabilir: 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 çok diğer faktör olduğu için bu tek olcutleri dağılımlarının normal çıktığı iddia edilebilir.

Toplamların dağılımının çan eğrisine yaklaşması durumu İstatistikte Merkezi Limit Teorisi ile ispatlanmıştır.

Simulasyon

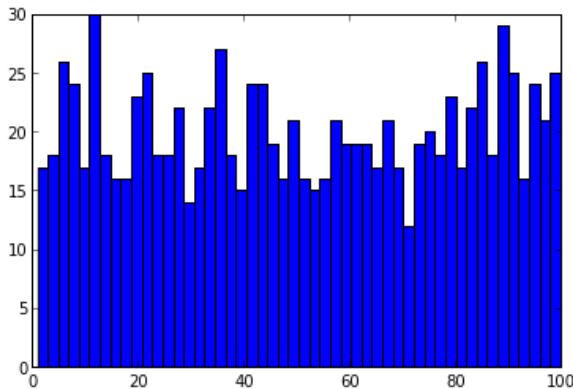
Eğer bu kavramları simulasyon ortamında göstermek istersek, Python ile bunu yapabiliriz.

İlk önce, Random.org sitesinden rasgele sayı üretip bilgisayarımıza 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 seçenek olabilir. Ama sunu da söylemek lazım, simulasyon tekniklerinin tamamı için yarı-rasgele (pseudorandom) sayılar 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 = []
```

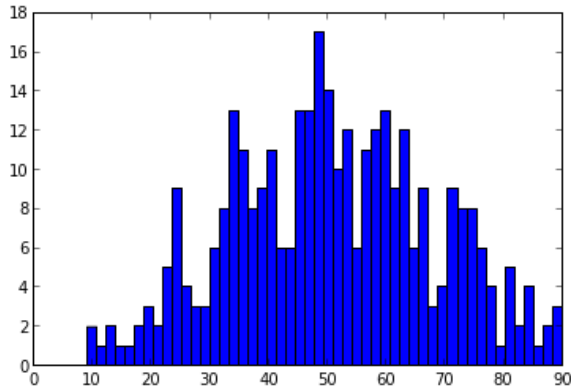
```

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. $A = \text{"en az bir tura gelme"}$ 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.

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(\text{hepsi yazi}) \\ &= 1 - P(T_1)P(T_2)\dots P(T_{10}) \\ &= 1 - \left(\frac{1}{2}\right)^{10} \approx .999 \end{aligned}$$

Rasgele Degiskenler (Random Variables)

Bir rasgele degisken X bir eslemedir, ki bu esleme $X : \Omega \rightarrow \mathbb{R}$ 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 $\omega = \text{HHTHHTHHTT}$ ise $X(\omega) = 6$. Tura sayisi eslemesi ω sonucunu 6 sayisina esledi.

Ornek

$\Omega = \{(x, y); x^2 + y^2 \leq 1\}$, 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 $F_X : \mathbb{R} \rightarrow [0, 1]$ tanimi

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

Eger X ayrık ise, yani sayılabilir bir küme $\{x_1, x_2, \dots\}$ icinden değerler aliyorsa olasılık fonksiyonu (probability function), ya da olasılık kütle fonksiyonu (probability mass function -PMF-)

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

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

Tanım

Eger X sürekli (continuous) ise, yani tüm x 'ler için $f_X(x) > 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$ olacak şekilde bir f_X mevcut ise, o zaman her $a \leq b$ için

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

Bu durumda f_X olasılık yoğunluk fonksiyonudur (probability density function -PDF-).

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

Ayrıca $F_X(x)$ 'in türevi alınabildiği her x noktasında $f_X(x) = F'_X(x)$ demektir.

Dikkat! Eger X sürekli ise o zaman $P(X = x) = 0$ değerindedir. $f(x)$ fonksiyonunu $P(X = x)$ olarak görmek hatalıdır. Bu sadece ayrık rasgele değişkenler için işler. Sürekli durumda olasılık hesabı için belli iki nokta arasında integral hesabı yapmamız gereklidir. Ek olarak PDF 1'den büyük olabilir, ama PMF olamaz. PDF'in 1'den büyük olabilmesi integrali bozmaz mı? Unutmayalım, integral hesabı yapıyoruz, noktasal değerlerin 1 olması tüm 1'lerin toplandığı anlamına gelmez. Bakınız *Entegralleri Nasıl Düşünelim* yazımız.

Tanım

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

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

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

Eger inf kavramını bilmiyorsak simdilik onu minimum olarak düşünebiliriz.

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

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

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

olarak bilinir.

İki rasgele değişken X ve Y dağılımsal olarak birbirine eşitliği, yani $X \stackrel{d}{=} Y$ eğer $F_X(x) = F_Y(x)$, $\forall x$. Bu X, Y birbirine eşit, birbirinin aynisi demek değildir. Bu değişkenler hakkındaki tüm olasılıksal işlemler, sonuçlar aynı olacak demektir.

Uyari! “ X ’in dağılımı F ’tir” beyanını $X \sim F$ şeklinde yazmak bir gelenek. Bu biraz kötü bir gelenek aslında çünkü \sim sembolü aynı zamanda yaklaşıksallık kavramını belirtmek için de kullanılıyor.

Dagilimler

Bernoulli Dağılımı

X ’in bir yazı-tura atısını temsil ettiğini düşünelim. O zaman $P(X = 1) = p$, ve $P(X = 0) = 1 - p$ olacaktır, ki $p \in [0, 1]$ olmak üzere. O zaman X ’in dağılımı Bernoulli deriz, ve $X \sim \text{Bernoulli}(p)$ diye gösteririz. Olasılık fonksiyonu $f(x) = p^x(1 - p)^{(1-x)}$, $x \in \{0, 1\}$.

Yani x ya 0, ya da 1. Parametre p , 0 ile 1 arasındaki herhangi bir reel sayı.

Beklenti ve varyans

$$E(X) = p$$

$$\text{Var}(X) = p(1 - p)$$

Uyari!

X bir rasgele değişken; x bu değişkenin alabileceği spesifik bir değer; p değeri ise bir **parametre**, yani sabit, önceden belirlenmiş reel sayı. Tabii istatistikî problemlerde (olasılık problemlerinin tersi olarak düşünersek) yoğunlukla o sabit parametre bilinmez, onun veriden hesaplanması, kestirilmesi gerekir. Her halukarda, coğu istatistikî modelde rasgele değişkenler vardır, ve onlardan ayrı olarak parametreler vardır. Bu iki kavramı birbiriyle karıştırmayalım.

Binom Dağılımı (Binomial Distribution)

Her biri birbirinden bağımsız ve birbiriyle aynı Bernoulli Dağılımına sahip deneylerden n tane yapıldığını farzedelim, ki bu deneylerin sadece iki sonucu olacak (1/0. başarı/başarısızlık, vs). Bu deneylerin p ’si aynı olacak. O zaman n deney içinden toplam kaç tanesinin başarılı olduğunu gösteren X rasgele değişkeni Binom Dağılımına sahiptir denir.

Bu dağılımın yoğunluğu

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

Bu fonksiyonun parametreleri p, n degerleridir. Beklenti ve varyans

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

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

Düz (Uniform) Dağılım

X düz, $\text{Uniform}(a, b)$ olarak dağılmış deriz, ve bu $X \sim \text{Uniform}(a, b)$ olarak yazılır eğer

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

ise ve $a < b$ olacak şekilde. CDF hesabi olasılık egrisinin integralini temel alır, düz dağılım bir a, b arasında $1/b - a$ yüksekliğinde bir dikdörtgen şeklinde olacaktır için, bu dikdörtgendeki herhangi bir x noktasında CDF dağılımı, yani o x 'in başlayıp sol tarafın alanının hesabi basit bir dikdörtgensel alan hesabıdır, 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) Dağılım

$X \sim N(\mu, \sigma^2)$ 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 şekilde. Bazıları bu dağılımı

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

olarak gösterebiliyor, çünkü bu şekilde (birazdan göreceğimiz) çok boyutlu Gaussian formülü ile alaka daha rahat gözüküyor.

İleride göreceğiz ki μ bu dağılımın “ortası”, ve σ onun etrafa ne kadar “yayıldığı” (spread). Normal dağılım olasılık ve istatistikte çok önemli bir rol oynar. Doğadaki pek çok olay yaklaşıksal olarak Normal dağılıma 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. Gele-
nege 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 bulunamamaktadır, bunun sebebi ise Nor-
mal PDF'in integralinin analitik olarak alinamiyor olusudur.

Bazi faydali puf noktaları

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

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

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

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

$$\begin{aligned} &= P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) \\ &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

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 hesaplayabiliy-
orsak, istedigimiz Normal olasilik hesabini yapabiliriz demektir, cunku artik X
iceren bir hesabın Z 'ye nasil tercume edildigini goruyoruz.

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

Ornek

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

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

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

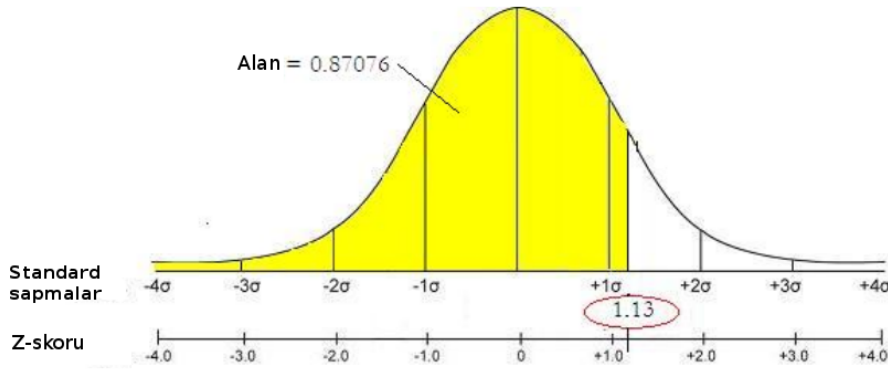
Soru $P(a < X < b)$ formunda a kullanmadi, sadece b oldugu icin yukaridaki form ortaya cikti. Python ile

```
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 bakmanın bir diğer 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 formül 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

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

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

t (Student's t) ve Cauchy Dağılımı

X , ν derece bağımsızlıkta t dağılımına sahiptir, ki bu $X \sim t_\nu$ diye yazılır eğer

$$f(x) = \frac{\Gamma(\nu + 1)/2}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

t dağılımı Normal dağılıma benzer ama daha kuyruğu daha kalındır. Aslında Normal dağılımı t dağılımının $\nu = \infty$ olduğu hale tekabül eder. Cauchy dağılımı da t'nin özel bir halidir, $\nu = 1$ halidir. Bu durumda yoğunluk fonksiyonu

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Bu formül hakikaten bir yoğunluk mudur? Kontrol için integralini alalım,

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

Cogunlukla entegre edilen yerde “1 artı ya da eksi bir şeyin karesi” turunde bir ifade gorulurse, yerine gecirme (subsitution) islemi trigonometrik olarak yapilir.

$$x = \tan \theta, \theta = \arctan x$$

$$1 + x^2 = 1 + \tan^2 \theta = \sec^2 \theta$$

$$dx/d\theta = \sec^2 \theta$$

O zaman

$$\begin{aligned} &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dx}{1 + x^2} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{\sec^2 \theta} \sec^2 \theta d\theta = \frac{1}{\pi} \int_{-\infty}^{\infty} 1 d\theta = \\ &= \frac{1}{\pi} \theta \Big|_{-\infty}^{\infty} = \frac{1}{\pi} [\arctan(\infty) - \arctan(-\infty)] \end{aligned}$$

$$= \frac{1}{\pi} \left[\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) \right] = 1$$

χ^2 Dagilimi

X 'in p derece serbestlige sahip bir χ^2 dagilima sahip ise $X \sim \chi_p^2$ olarak gosterilir, yogunluk

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, x > 0$$

Eger Z_1, \dots, Z_p bagimsiz standart Normal rasgele degiskenler ise, $\sum_{i=1}^p Z_i^2 \sim \chi_p^2$ esitligi dogrudur.

İki Degiskenli Dagilimler

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 $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ 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

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

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

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

Niye c bilinmiyor? Belki problemin modellemesi sirasinda bu bilinmez olarak

ortaya cikmistir. Olabilir. Bu degeri hesaplayabiliriz, cunku $f(x, y)$ yogunluk olmalı, ve yogunluk olmanın sarti $f(x, y)$ entegre edilince sonucun 1 olması.

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

$$\begin{aligned}
 1 &= \int \int f(x, y) dy dx = c \int_{-1}^1 \int_{x^2}^1 x^2 y \\
 &= c \int_{-1}^1 x^2 \int_{x^2}^1 y dy dx = \int_{-1}^1 x^2 \left(\frac{1}{2} - \frac{x^4}{2} \right) dx = 1 \\
 &= c \int_{-1}^1 x^2 \left(\frac{1 - x^4}{2} \right) dx = 1 \\
 &= \frac{c}{2} \int_{-1}^1 x^2 - x^6 dx = 1
 \end{aligned}$$

Devam edersek $c = 21/4$ buluruz.

Simdi, diyelim ki bizden $P(X \geq Y)$ 'yi hesaplamamız isteniyor. Bu hangi A bolgesine tekabül eder? Elimizdekiler

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

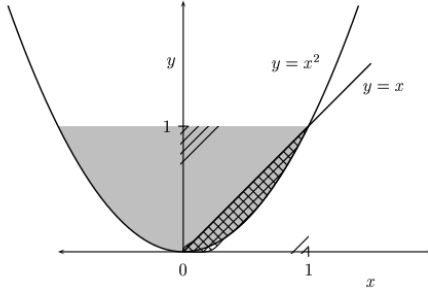
Simdi bunlara bir de $y \leq x$ eklememiz lazim. Yani ortadaki esitsizlige bir öge daha eklenir.

$$-1 \leq x \leq 1$$

$$x^2 \leq y \leq x$$

$$y \leq 1$$

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



Aynı şekilde $y \leq x$ için $y = x$ 'i düşünelim, ki bu 45 derece açıyla çizilmiş düz bir çizgi. Çizginin altı $y \leq x$ olur. Bu iki bölgenin kesişimi yukarıdaki resimdeki gölgeli kısım.

Ek bir bölge şartı $0 \leq x \leq 1$. Bu şart resimde bariz görülüyor, ama cebirsel olarak bakarsak $y \geq x^2$ olduğunu biliyoruz, o zaman $y \geq 0$ çünkü x^2 muhakkak bir pozitif sayı olmalı. Diğer yandan $x \geq y$ verilmiş, tüm bunları yanyana koyarsak $x \geq 0$ şartı ortaya çıkar.

Artık $P(X \geq Y)$ hesabı için hazırız,

$$\begin{aligned} P(X \geq Y) &= \frac{21}{4} \int_0^1 \int_{x^2}^x x^2 y \, dy \, dx = \frac{21}{4} \int_0^1 x^2 \left[\int_{x^2}^x y \, dy \right] dx \\ &= \frac{21}{4} \int_0^1 x^2 \frac{x^2 - x^4}{2} dx = \frac{3}{20} \end{aligned}$$

“Hafızasız” Dağılım, Ustel (Exponential) Dağılım

Ustel dağılımın hafızasız olduğu söylenir. Bunun ne anlama geldiğini anlatmaya uğrasalım. Diyelim ki rasgele değişken X bir aletin omrunu temsil ediyor, yani bir $p(x)$ fonksiyonuna bir zaman “sordugumuz” zaman bize dondurulmuş olasılık, o aletin x zamanı kadar daha işlemesinin olasılığı. Eğer $p(2) = 0.2$ ise, aletin 2 yıl daha yaşamasının olasılığı 0.2.

Bu hafızasızlığı, olasılık matematiği ile nasıl temsil ederiz?

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

Yani öyle bir dağılım var ki elimizde, $X > t$ bilgisi veriliyor, ama (kalan) zamanı hala $P(X > s)$ olasılığı veriyor. Yani t kadar zaman geçtiği bilgisi hiçbir şeyi değiştirmiyor. Ne kadar zaman geçmiş olursa olsun, direkt s ile gidip aynı olasılık hesabını yapıyoruz.

Sartsal (conditional) formülünü uygularsak üstteki şöyle 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 olması için X ne şekilde dağılmış olmalıdır? Üstteki denklem sadece X dağılım fonksiyonu üstel (exponential) olursa mümkündür, çünkü sadece o zaman

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

gibi bir ilişki kurulabilir.

Örnek

Diyelim ki bir bankadaki bekleme zamanı ortalama 10 dakika ve üstel olarak dağılmış. Bir müşterinin i) bu bankada 15 dakika beklemesinin ihtimali nedir? ii) Bu müşterinin 10 dakika bekledikten sonra toplam olarak 15 dakika (ya da daha fazla) beklemesinin olasılığı nedir?

Cevap

i) Eğer X müşterinin bankada beklediği zamanı temsil ediyorsa

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

ii) Sorunun bu kısmı müşteri 10 dakika gecirdikten sonra 5 dakika daha gecirmesinin olasılığını soruyor. Fakat üstel dağılım “hafızasız” olduğu için kalan zamanı alıp yine direkt aynı fonksiyona geçiyoruz,

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

Bilesen (Marginal) Dağılımlar

Sürekli rasgele değişkenler için bilesen yoğunluk

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

ve

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

Üstteki integraller gerçek bir dağılım fonksiyonu $f(x, y)$ verilince alt ve üst limitleri tanımlamak zorundadır. Çünkü bilesen yoğunluk için bir veya daha fazla

degiskeni “integralle disari atmak (integrate out)” ettigimiz soylenir, eger ayriskal (discrete) ortamda olsaydik bu atilan degiskenin tum degerlerini goze alarak toplama yapan bir formül 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 gerektiği için $-\infty, +\infty$ olmalıdır. 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 $x, y \geq 0$. O zaman $f_X(x)$

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

Ornek

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

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

Tanim

İki rasgele degisken A, B bagimsizdir eger tum A, B degerleri için

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

esitligi dogru ise. Bu durumda $X \perp Y$ 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 $X \perp Y$ dogrudur.

Ornek

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

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

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

Cevap

Bagimsizligi kullanarak birlesik dagilimi hesaplayabiliriz

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

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

$$P(X + Y \leq 1) = \iint_{x+y \leq 1} f(x, y) dy dx$$

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

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

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 Dagilimler (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,

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

$P(X < 1/4 | Y = 1/3)$ 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) Dagilimler ve IID Orneklemler (Samples)

$X = (X_1, \dots, X_n)$ olsun, ki (X_1, \dots, X_n) 'lerin herbiri bir rasgele degisken, o zaman X 'e rasgele vektor (random vector) ismi verilir. $f(x_1, \dots, x_n)$ 'in PDF'i temsil ettigini dusunelim. Bu PDF'i baz alarak aynen iki degiskenli (bivariate) orneklerde oldugu gibi, benzer tekniklerle bilesenleri, kosullu dagilimleri, 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 $Z_1, \dots, Z_k \sim N(0,1)$. 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* çok 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 çok 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. Hatırlayalım, bir matris pozitif kesindir eger tum sifir olmayan x vektorleri icin $x^T \Sigma x > 0$ ise.

Not: Karekok kavrami tekil sayılardan matrislere de aktarılabilir. 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 $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

Hatırlama Numarası

Normal Dagilimin formülünü bazen hatırlayamayabiliriz. Peki daha basit bir formülden başlayarak onu türetebilir miyiz? Bu mümkün. Daha önce e^{-x^2} *Nasil Entegre Edilir?* yazısında

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

oldugunu gormustuk. Dikkat edersek bu integral bir formülün 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 tarafı $\sqrt{\pi}$ 'ye bolerseniz, sag taraf 1 olur ve böylece solda bir dagilim elde ederiz. Yani

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

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

İlk amacımız $\sigma = 1$ '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,

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

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

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

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

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

/

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 cogenlukla 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 olasiliklerini 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 attirmek icin ozel kodlamalar, yari-rasgele sayi uretimi gereklidir, biz kavramsal ve cebirsel olarak onlari neyi temsil ettiginden bahsediyoruz.

İki kavramdan daha bahsetmek bu noktada faydalı. 1) Nüfus (Population) 2) Örneklem (Sample). Nüfus, üzerinde istatistiksel analiz yaptığımız kitlenin tamamı. Eğer insanların boyları hakkında istatistiksel analiz yapıyor olsaydık tüm insanlar nüfus olurdu. Nüfusun cinsiyetle ilgili dağılımı bilinmez, biliniyor olsa da dağılımın parametreleri bilinmiyor olabilir. Örneklem, nüfus içinden alınan rasgele ölçümlere verilen isimdir, X_1, \dots, X_n olarak gösterilebiliyor, bu durumda nüfusun dağılımının “zar attığı” ve her zar atışının rasgele değişkenlerden birinin içini doldurduğu düşünülebilir. Örneklem nüfustan geldiği için dağılımının aynen nüfus gibi olduğu kabul edilir. Bu bağlantıdan yola çıkılarak birçok istatistiksel analiz yapmak mümkündür.

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