Ozet Istatistikleri, Grafikleri

Beklenti (Expectation)

Bu deger, dagilim $f(x)$'in tek sayilik bir ozetidir. Yani beklenti hesabina bir taraftan bir dagilim fonksiyonu girer, diger taraftan tek bir sayi disari cikar.

Tanim

Surekli dagilim fonksiyonlari icin $E(X)$

$$E(X) = \int x f(x) dx$$

ayriksal dagilimlar icin

$$E(X) = \sum_x x f(x)$$

Hesabin, her x degerini onun olasiligi ile carpip topladigina dikkat. Bu tur bir hesap dogal olarak tum x'lerin ortalamasini verecektir, ve dolayli olarak dagilimin ortalamasini hesaplayacaktir. Ortalama μ_x olarak ta gosterilebilir.

$E(X)$'in bir tanim olduguna dikkat, yani bu ifade tamamen bizim yarattigimiz, ortaya cikarttigimiz bir sey, matematigin baz kurallarindan gelerek turetilen bir kavram degil.

Notasyonel basitlik icin ustteki toplam / entegral yerine

$$= \int x \ dF(x)$$

diyecegiz, bu notasyonel bir kullanim sadece, unutmayalim, reel analizde $\int x \ dF(x)$'in ozel bir anlami var (hoca tam diferansiyel dF'den bahsediyor).

Beklentinin taniminin kapsamli / eksiksiz olmasi icin $E(X)$'in "mevcudiyeti" icin de bir sart tanimlamak gerekir, bu sart soyle olsun,

$$\int_x |x| dF_X(x) < \infty$$

ise beklenti mevcut demektir. Tersi sozkonusu ise beklenti mevcut degildir.

Ornek

$X \sim Unif(-1, 3)$ olsun. $E(X) = \int x dF(x) = \int x f_X(x) dx = \frac{1}{4} \int_{-1}^{3} x dx = 1$.

Ornek

Cauchy dagiliminin $f_X(x) = \{\pi(1+x^2)\}^{-1}$ oldugunu soylemistik. Simdi beklentiyi hesaplayalim. Parcali entegral teknigi lazim, $u = x$, $dv = 1/1 + x^2$ deriz, ve o zaman $v = \tan^{-1}(x)$ olur, bkz *Ters Trigonometrik Formuller* yazimiz. Demek ki

$$\int |x| dF(x) = \frac{2}{\pi} \int_0^\infty \frac{x dx}{1+x^2}$$

2 nereden cikti? Cunku $|x|$ kullaniyoruz, o zaman sinir degerlerinde sadece sifirin sagina bakip sonucu ikiyle carpmak yeterli. Bir sabit oldugu icin π ile beraber disari cikiyor. Simdi

$$\int u dv = uv - \int v du$$

uzerinden

$$= [x \tan^{-1}(x)]_0^\infty - \int_0^\infty \tan^{-1}(x) dx = \infty$$

Yani ustteki hesap sonsuzluga gider. O zaman ustteki tanimimiza gore Cauchy dagiliminin beklentisi yoktur.

Tanim

$x_1, .., x_n$ verilerini iceren orneklemin (sample) ortalamasi

$$\bar{x} = \frac{1}{n} \sum x_i \tag{1}$$

Dikkat bu orneklemdeki verinin ortalamasi. Hicbir dagilim hakkinda hicbir faraziye yapmadik. Ayrica tanim kullandik, yani bu ifadenin ne oldugu tamamen bize bagli.

Orneklem ortalamasi sadece tek merkezi bir tepesi olan (unimodal) dagilimlar icin gecerlidir. Eger bu temel varsayim gecerli degilse, ortalama kullanarak yapilan hesaplar bizi yanlis yollara goturur. Ayrica bir dagilimi simetrik olup olmadigi da ortalama ya da medyan kullanilip kullanilmamasi kararinda onemlidir. Eger simetrik, tek tepeli bir dagilim var ise, ortalama ve medyan birbirine yakin olacaktir. Fakat veri baska turde bir dagilim ise, o zaman bu iki olcut birbirinden cok farkli olabilir.

Tanim

Y rasgele degiskeninin varyansi (variance)

$$Var(Y) = E((Y - E(Y))^2)$$

Ifadede toplama ve bolme gibi islemler olmadigina dikkat; onun yerine kare ifadeleri uzerinde beklenti ifadesi var. Yani Y'nin beklentisini rasgele degiskenin kendisinden cikartip kareyi aliyoruz, ve bu islemin Y'den gelen tum zar atislari uzerinden beklentisi bize varyansi veriyor. Bir rasgele degisken gorunce onun yerine "dagilimdan uretilen sayi" dusunmek faydalidir, ki bu gercek dunya sartlarindan (ve buyuk miktarda olunca) veri noktalarini temsil eder.

Varyans formulunu acarsak, ileride isimize yarayacak baska bir formul elde edebiliriz,

$$Var(Y) = E(Y^2 - 2YE(Y) + (E(Y)^2))$$

$$= E(Y^2) - 2E(Y)E(Y) + (E(Y)^2)$$

$$Var(Y) = E(Y^2) - (E(Y)^2)$$

Tanim

$y_1, .., y_n$ ornekleminin varyansi (literaturde S^2 olarak gecebiliyor,

$$S^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 \tag{2}$$

Standart sapma veri noktalarin "ortalamadan farkinin ortalamasini" verir. Tabii bazen noktalar ortalamanin altinda, bazen ustunde olacaktir, bizi bu negatiflik, pozitiflik ilgilendirmez, biz sadece farkla alakaliyiz. O yuzden her sapmanin karesini aliriz, bunlari toplayip nokta sayisina boleriz.

Ilginc bir cebirsel islem sudur ve bize verinin uzerinden tek bir kez gecerek (one pass) hem sayisal ortalamayi hem de sayisal varyansi hesaplamamizi saglar. Eger $\bar{y}$ tanimini ustteki formule sokarsak,

$$= \frac{1}{n} \sum_i y_i^2 + \frac{1}{n} \sum_i m^2 - \frac{2}{n} \sum_i y_i \bar{y}$$

$$= \frac{1}{n} \sum_i y_i^2 + \frac{\bar{y}^2 n}{n} - \frac{2\bar{y}n}{n} \bar{y}$$

$$= \frac{1}{n} \sum_i y_i^2 + \bar{y}^2 - 2\bar{y}^2$$

$$= \frac{1}{n} \sum_i y_i^2 - \bar{y}^2$$

Bu arada standard sapma varyansin karekokudur, ve biz karekok olan versiyon ile calismayi tercih ediyoruz. Niye? Cunku o zaman veri noktalarinin ve yayilma olcusunun birimleri birbiri ile ayni olacak. Eger veri setimiz bir alisveris sepetindeki malzemelerin lira cinsinden degerleri olsaydi, varyans bize sonucu ”kare lira” olarak verecekti ve bunun pek anlami olmayacakti.

Kovaryans ve Korelasyon (Covariance and Correlation)

Verisel Kovaryans (Empirical Covariance)

Eger verinin kolonlari arasindaki iliskiyi gormek istersek, en hizli yontem matristeki her kolonun (degiskenin) ortalamasini kendisinden cikartmak, yani onu “sifirda ortalamak” ve bu matrisin devrigini alarak kendisi ile carpmaktir. Bu islem her kolonu kendisi ve diger kolonlar ile noktasal carpimdan gecirecektir ve carpim, toplama sonucunu nihai matrise yazacaktir. Carpimlarin bildigimiz ozelligine gore, arti deger arti degerle carpilinca arti, eksi ile eksi arti, eksi ile arti eksi verir, ve bu bilgi bize ilinti bulma hakkinda guzel bir ipucu sunar. Pozitif sonucun pozitif korelasyon, negatif ise tersi sekilde ilinti oldugu sonucuna boylece kolayca erisebiliriz.

Tanim

$$S = \frac{1}{n}(X - E(X))^T(X - E(X)))$$

Pandas ile `cov` cagrisi bu hesabi hizli bir sekilde yapar,

```
print df.cov()
```

```
              Sepal Length  Sepal Width  Petal Length  Petal Width
Sepal Length      0.685694    -0.039268      1.273682     0.516904
Sepal Width      -0.039268     0.188004     -0.321713    -0.117981
Petal Length      1.273682    -0.321713      3.113179     1.296387
Petal Width       0.516904    -0.117981      1.296387     0.582414
```

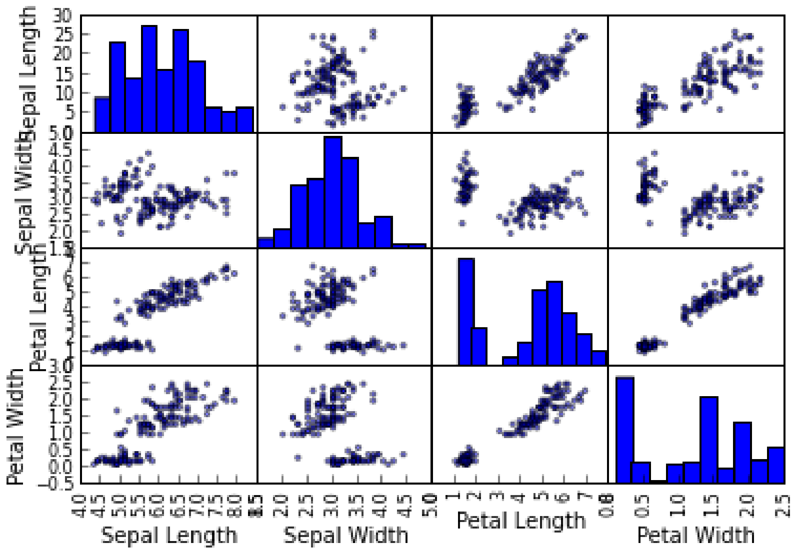
Eger kendimiz bu hesabi yapmak istersek,

```
means = df.mean()
n = df.shape[0]
df2 = df.apply(lambda x: x - means, axis=1)
print np.dot(df2.T,df2) / n
```

```
[[ 0.68112222 -0.03900667  1.26519111  0.51345778]
 [-0.03900667  0.18675067 -0.319568   -0.11719467]
 [ 1.26519111 -0.319568    3.09242489  1.28774489]
 [ 0.51345778 -0.11719467  1.28774489  0.57853156]]
```

Verisel kovaryansin sayisal gosterdigini grafiklemek istersek, yani iki veya daha fazla boyutun arasindaki iliskileri grafiklemek icin yontemlerden birisi verideki mumkun her ikili iliskiyi grafiksel olarak gostermektir. Pandas `scatter_matrix` bunu yapabilir. Iris veri seti uzerinde gorelim, her boyut hem y-ekseni hem x-ekseninde verilmis, iliskiyi gormek icin eksende o boyutu bulup kesisme noktalarindaki grafige bakmak lazim.

```
import pandas as pd
df = pd.read_csv('iris.csv')
df = df.ix[:,0:4]
pd.scatter_matrix(df)
plt.savefig('stat_summary_01.png')
```



Iliski oldugu zaman o iliskiye tekabul eden grafikte "duz cizgiye benzer" bir goruntu olur, demek ki degiskenlerden biri artinca oteki de artiyor (eger cizgi soldan sage yukari dogru gidiyorsa), azalinca oteki de azaliyor demektir (eger cizgi asagi dogru iniyorsa). Eger ilinti yok ise bol gurultulu, ya da yuvarlak kureye benzer bir sekil cikar. Ustteki grafige gore yaprak genisligi (petal width) ile yaprak boyu (petal length) arasinda bir iliski var.

Tanim

X, Y rasgele degiskenlerin arasindaki kovaryans,

$$Cov(X, Y) = E(X - E(X))(Y - E(Y))$$

Yani hem X hem Y'nin beklentilerinden ne kadar saptiklarini her veri ikilisi icin, cikartarak tespit ediyoruz, daha sonra bu farklari birbiriyle carpiyoruz, ve beklentisini aliyoruz (yani tum olasilik uzerinden ne olacagini hesapliyoruz).

Ayri ayri X, Y degiskenleri yerine cok boyutlu X kullanirsak, ki boyutlari m, n olsun yani m veri noktasi ve n boyut (ozellik, oge) var, tanimi soyle ifade edebiliriz,

$$\Sigma = Cov(X) = E((X - E(X))^T (X - E(X)))$$

Medyan ve Yuzdelikler (Percentile)

Ustteki hesaplarin cogu sayilari toplayip, bolmek uzerinden yapildi. Medyan ve diger yuzdeliklerin hesabi (ki medyan 50. yuzdelige tekabul eder) icin eldeki tum degerleri "siraya dizmemiz" ve sonra 50. yuzdelik icin ortadakine bakmamiz gerekiyor. Mesela eger ilk 5. yuzdeligi ariyorsak ve elimizde 80 tane deger var ise, bastan 4. sayiya / vektor hucresine / ogeye bakmamiz gerekiyor. Eger 100 eleman var ise, 5. sayiya bakmamiz gerekiyor, vs.

Bu siraya dizme islemi kritik. Kiyasla ortalama hesabi hangi sirada olursa olsun, sayilari birbirine topluyor ve sonra boluyor. Zaten ortalama ve sapmanin istatistikte daha cok kullanilmasinin tarihi sebebi de aslinda bu; bilgisayar oncesi cagda sayilari siralamak (sorting) zor bir isti. Bu sebeple hangi sirada olursa olsun, toplayip, bolerek hesaplanabilecek ozetler daha makbuldu. Fakat artik siralama islemi kolay, ve veri setleri her zaman tek tepeli, simetrik olmayabiliyor. Ornek veri seti olarak unlu `dellstore2` tabanindaki satis miktarlari kullanirsak,

```
print np.mean(data)
```

```
213.948899167
```

```
print np.median(data)
```

```
214.06
```

```
print np.std(data)
```

```
125.118481954
```

```
print np.mean(data)+2*np.std(data)
```

```
464.185863074
```

```
print np.percentile(data, 95)
```

```
410.4115
```

Goruldugu gibi uc nokta hesabi icin ortalamadan iki sapma otesini kullanirsak, 464.18, fakat 95. yuzdeligi kullanirsak 410.41 elde ediyoruz. Niye? Sebep ortalamanin kendisi hesaplanirken cok uc degerlerin toplama dahil edilmis olmasi ve bu durum, ortalamanin kendisini daha buyuk seviyeye dogru itiyor. Yuzdelik hesabi ise sadece sayilari siralayip belli bazi elemanlari otomatik olarak uc nokta olarak addediyor.

Box Whisker Grafikleri

Tek boyutlu bir verinin dagilimini gormek icin Box ve Whisker grafikleri faydali araclardir; medyan (median), dagilimin genisligini ve siradisi noktalari (outliers) acik sekilde gosterirler. Isim nereden geliyor? Box yani kutu, dagilimin agirlig-inin nerede oldugunu gosterir, medyanin sagindada ve solunda olmak uzere iki ceyregin arasindaki kisimdir, kutu olarak resmedilir. Whiskers kedilerin biyik-larina verilen isimdir, zaten grafikte birazcik biyik gibi duruyorlar. Bu uzantilar medyan noktasindan her iki yana kutunun iki kati kadar uzatilir sonra verideki

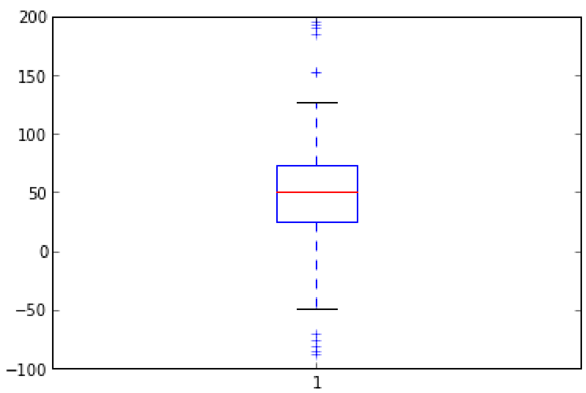
”ondan az olan en buyuk” noktaya kadar geri cekilir. Tum bunlarin disinda kalan veri ise teker teker nokta olarak grafikte basilir. Bunlar siradisi (outlier) olduklari icin daha az olacaklari tahmin edilir.

BW grafikleri iki veriyi dagilimsal olarak karsilastirmak icin birebirdir. Mesela Larsen and Marx adli arastirmacilar cok az veri iceren Quintus Curtius Snodgrass veri setinin degisik oldugunu ispatlamak icin bir suru hesap yapmislardir, bir suru matematiksel isleme girmislerdir, fakat basit bir BW grafigi iki setin farkliligini hemen gosterir.

BW grafikleri iki veriyi dagilimsal olarak karsilastirmak icin birebirdir. Mesela Larsen and Marx adli arastirmacilar cok az veri iceren Quintus Curtius Snodgrass veri setinin degisik oldugunu ispatlamak icin bir suru hesap yapmislardir, bir suru matematiksel isleme girmislerdir, fakat basit bir BW grafigi iki setin farkliligini hemen gosterir.

Python uzerinde basit bir BW grafigi

```
spread= rand(50) * 100
center = ones(25) * 50
flier_high = rand(10) * 100 + 100
flier_low = rand(10) * -100
data =concatenate((spread, center, flier_high, flier_low), 0)
plt.boxplot(data)
plt.savefig('05_03.png')
```

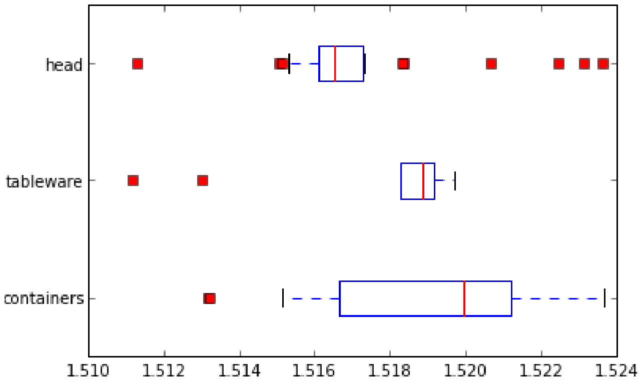


Bir diger ornek Glass veri seti uzerinde

```
data = loadtxt("glass.data",delimiter=",")
head = data[data[:,10]==7]
tableware = data[data[:,10]==6]
containers = data[data[:,10]==5]

print head[:,1]

data =(containers[:,1], tableware[:,1], head[:,1])
```

```
plt.yticks([1, 2, 3], ['containers', 'tableware', 'head'])

plt.boxplot(data,0,'rs',0,0.75)
plt.savefig('05_04.png')
```

```
[ 1.51131  1.51838  1.52315  1.52247  1.52365  1.51613  1.51602  1.51623
  1.51719  1.51683  1.51545  1.51556  1.51727  1.51531  1.51609  1.51508
  1.51653  1.51514  1.51658  1.51617  1.51732  1.51645  1.51831  1.5164
  1.51623  1.51685  1.52065  1.51651  1.51711]
```



Parametre Tahmin Ediciler (Estimators)

Maksimum Olurluk (maximum likelihood) kavramini kullanarak ilginc bazi sonuclara erismek mumkun; bu sayede dagilim fonksiyonlari ve veri arasinda bazi sonuclar elde edebiliriz. Maksimum olurluk nedir? MO ile verinin her noktasi teker teker olasilik fonksiyonuna gecilir, ve elde edilen olasilik sonuclari birbiri ile carpilir. Cogunlukla formul icinde bilinmeyen bir(kac) parametre vardir, ve bu carpim sonrasi, icinde bu parametre(ler) olan yeni bir formul ortaya cikar. Bu nihai formulun kismi turevi alinip sifira esitlenince cebirsel bazi teknikler ile bilinmeyen parametre bulunabilir. Bu sonuc eldeki veri baglaminda en mumkun (olur) parametre degeridir. Oyle ya, mesela Gaussian $N(10, 2)$ dagilimi var ise, 60,90 gibi degerlerin "olurlugu" dusuktur. Gaussin uzerinde ornek,

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

Carpim sonrasi

$$f(x_1, .., x_n; \mu, \sigma) = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i-\mu)^2 \right\}$$

$$= \frac{(2\pi)^{-n/2}}{\sigma^n} \exp\left\{ -\frac{\sum(x_i-\mu)^2}{2\sigma^2} \right\}$$

Ustel kisim $-n/2$ nereden geldi? Cunku bolen olan karekoku uste cikardik, boylece $-1/2$ oldu, n cunku n tane veri noktasi yuzunden formul n kere carpiliyor. Veri noktalari x_i icinde. Eger log, yani ln alirsak exp'den kurtuluruz, ve biliyoruz ki log olurlugu maksimize etmek normal olurlugu maksimize etmek ile ayni seydir, cunku ln transformasyonu monoton bir transformasyondur. Ayrica olurluk icbukeydir (concave) yani kesin tek bir maksimumu vardir.

$$\ln f = -\frac{1}{2} n \ln(2\pi) - n \ln \sigma - \frac{\sum (x_i - \mu)^2}{2\sigma^2}$$

Turevi alip sifira esitleyelim

$$\frac{\partial (\ln f)}{\partial \mu} = \frac{\sum (x_i - \mu)^2}{2\sigma^2} = 0$$

$$\hat{\mu} = \frac{\sum x_i}{n}$$

Bu sonuc (1)'deki formul, yani orneklem ortalamasi ile ayni! Fakat buradan hemen bir baglantiya ziplamadan once sunu hatirlayalim - orneklem ortalamasi formulunu *biz* tanimladik. "Tanim" diyerek bir ifade yazdik, ve budur dedik. Simdi sonradan, verinin dagiliminin Gaussian oldugunu farzederek, bu verinin mumkun kilabilecegi en optimal parametre degeri nedir diye hesap ederek ayni formule eristik, fakat bu bir anlamda bir guzel raslanti oldu.. Daha dogrusu bu aynilik Gaussian / Normal dagilimlarinin "normalligi" ile alakali muhakkak, fakat ornekleme ortalamasi hicbir dagilim faraziyesi yapmiyor, herhangi bir dagilimdan geldigi bilinen ya da bilinmeyen bir veri uzerinde kullanilabiliyor. Bunu unutmayalim. Istatistikte matematigin lakaytlasmasi (sloppy) kolaydir, o sebeple neyin tanim, neyin hangi faraziyeye gore optimal, neyin nufus (population) neyin orneklem (sample) oldugunu hep hatirlamamiz lazim.

Devam edelim, maksimum olurluk ile $\hat{\sigma}$ hesaplayalim,

$$\frac{\partial (\ln f)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum (x_i - \mu)^2}{2\sigma^3} = 0$$

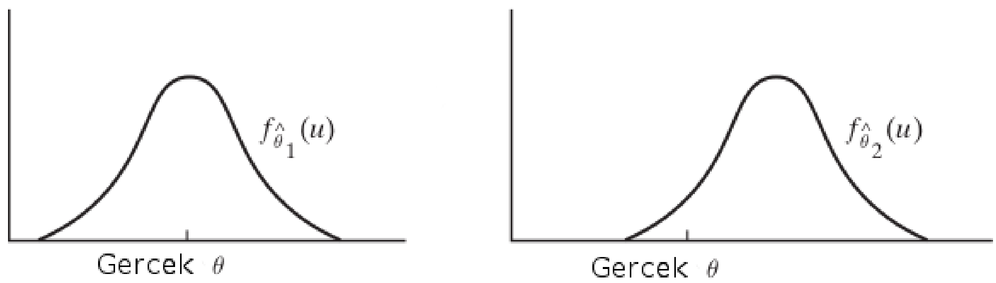
Cebirsel birkac duzenleme sonrasi ve μ yerine yeni hesapladigimiz $\hat{\mu}$ kullanarak,

$$\hat{\sigma}^2 = \frac{\sum (x_i - \hat{\mu})^2}{n}$$

Bu da orneklem varyansi ile ayni!

Yansizlik (Unbiasedness)

Tahmin edicilerin kendileri de birer rasgele degisken oldugu icin her orneklem icin degisik degerler verirler. Diyelim ki θ icin bir tahmin edici $\hat{\theta}$ hesapliyoruz, bu $\hat{\theta}$ gercek θ icin bazi orneklemler icin cok kucuk, bazi orneklemler icin cok buyuk sonuclar (tahminler) verebilecektir. Kabaca ideal durumun, az cikan tahminlerin cok cikan tahminleri bir sekilde dengelemesi oldugunu tahmin edebiliriz, yani tahmin edicinin uretecegi pek cok degerin θ'yi bir sekilde "ortalamasi" iyi olacaktir.



Bu durumu soyle aciklayalim, madem tahmin ediciler birer rasgele degisken, o zaman bir dagilim fonksiyonlari var. Ve ustteki resimde ornek olarak $\hat{\theta_1}, \hat{\theta_2}$ olarak iki tahmin edici gosteriliyor mesela ve onlara tekabul eden yogunluklar $f_{\hat{\theta_1}}, f_{\hat{\theta_1}}$. Ideal durum soldaki resimdir, yogunlugun fazla oldugu yer gercek θ'ya yakin olmasi. Bu durumu matematiksel olarak nasil belirtiriz? Beklenti ile!

Tanim

$Y_1, .., Y_n$ uzerindeki θ tahmin edicisi $\hat{\theta}$'den alinmis rasgele orneklem. Eger tum θ'lar icin $E(\hat{\theta}) = \theta$ ise, bu durumda tahmin edicinin yansiz oldugu soylenir.

Ornek olarak maksimum olurluk ile onceden hesapladigimiz $\hat{\sigma}$ tahmin edicisine bakalim. Bu ifade

$$\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{\mu})^2$$

ya da

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (Y_i - \bar{Y})^2$$

ile belirtildi. Tahmin edici $\hat{\sigma}^2$, σ^2 icin yansiz midir? Tanimimiza gore eger tahmin edici yansiz ise $E(\hat{\sigma}^2) = \sigma^2$ olmalidir.

Not: Faydali olacak bazi esitlikler, daha onceden gordugumuz

$$Var(X) = E(X^2) - (E(X)^2)$$

ve sayisal ortalama $\bar{Y}$'nin beklentisi $E(\bar{Y}) = E(Y_i)$, ve $Var(\bar{Y}) = 1/nVar(Y_i)$.

Baslayalim,

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n}\sum_i (Y_i - \bar{Y})^2\right)$$

Parantez icindeki $1/n$ sonrasindaki ifadeyi acarsak,

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)$$

$$= \sum_i Y_i^2 - 2\sum_i Y_i\bar{Y} + n\bar{Y}^2$$

$\sum_i Y_i$'nin hemen yaninda $\bar{Y}$ goruyoruz. Fakat $\bar{Y}$'nin kendisi zaten $1/n \sum_i Y_i$ demek degil midir? Ya da, toplam icinde her i icin degismeyecek $\bar{Y}$'yi toplam disina cekersek, $\bar{Y}\sum_i Y_i$ olur, bu da $\bar{Y} \cdot n\bar{Y}$ demektir ya da $n\bar{Y}^2$,

$$= \sum_i Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2$$

$$= \sum_i Y_i^2 - n\bar{Y}^2$$

Dikkat, artik $-n\bar{Y}^2$ toplama isleminin *disinda*. Simdi beklentiye geri donelim,

$$= E\left(\frac{1}{n}\left(\sum_i Y_i^2 - n\bar{Y}^2\right)\right)$$

$1/n$ disari cekilir, beklenti toplamdan iceri nufuz eder,

$$= \frac{1}{n}\left(\sum_i E(Y_i^2) - nE(\bar{Y}^2)\right)$$

Daha once demistik ki (genel baglamda)

$$Var(X) = E(X^2) - (E(X)^2)$$

Bu ornek icin harfleri degistirirsek,

$$Var(Y_i) = E(Y_i^2) - E(Y_i)^2$$

Yani

$$E(Y_i^2) = Var(Y_i) + E(Y_i)^2$$

$E(Y_i) = \mu$ oldugunu biliyoruz,

$$E(Y_i^2) = Var(Y_i) + \mu^2$$

Aynisini $E(\bar{Y}^2)$ icin kullanirsak,

$$E(\bar{Y}^2) = Var(\bar{Y}) + E(\bar{Y})^2$$

$E(\bar{Y}) = \mu$,

$$E(\bar{Y}^2) = Var(\bar{Y}) + \mu^2$$

$$= \frac{1}{n}\bigg(\sum_i Var(Y_i) + \mu^2 - n(Var(\bar{Y}) + \mu^2)\bigg)$$

$Var(Y_i) = \sigma$, ve basta verdigimiz esitlikler ile beraber

$$= \frac{1}{n}\bigg(\sum_i (\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\bigg)$$

Tekrar hatirlatalim, $\sum_i$ sadece ilk iki terim icin gecerli, o zaman, ve sabit degerleri n kadar topladigimiza gore bu aslinda bir carpim islemi olur,

$$= \frac{1}{n}\bigg(n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n} + \mu^2)\bigg)$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n}$$

$$= \frac{n\sigma^2 - \sigma^2}{n}$$

$$= \frac{\sigma^2(n-1)}{n}$$

$$= \sigma^2\frac{n-1}{n}$$

Goruldugu gibi eristigimiz sonuc σ^2 degil, demek ki bu tahmin edici yansiz degil. Kontrol tamamlandi.

Fakat eristigimiz son denklem bize baska bir sey gosteriyor, eger ustteki sonucu $\frac{n}{n-1}$ ile carpsaydik, σ^2 elde etmez miydik? O zaman yanli tahmin ediciyi yansiz hale cevirmek icin, onu $\frac{n}{n-1}$ ile carpariz ve

$$\frac{n}{n-1}\frac{1}{n}\sum_i (Y_i - \bar{Y})^2$$

$$= \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$$

Ustteki ifade σ^2'nin yansiz tahmin edicisidir.

Hesap icin kullandiginiz kutuphanelerin yanli mi yansiz mi hesap yaptigini bilmek iyi olur, mesela Numpy versiyon 1.7.1 itibariyle yanli standart sapma hesabi yapiyor, fakat Pandas yansiz olani kullaniyor (Pandas versiyonu daha iyi)

```
import pandas as pd
arr = np.array([1,2,3])
print 'numpy', np.std(arr)
print 'pandas', float(pd.DataFrame(arr).std())
```

```
numpy 0.816496580928
pandas 1.0
```

Büyük Sayılar Kanunu (Law of Large Numbers)

Olasılık kuramında önemli matematiksel bir denklem, büyük sayılar kanunudur. Bu kanun, orneklem (sample) ile rasgele degiskenler, yani matematiksel olasilik dagilimlari olan dunya arasinda bir baglanti gorevi gorur.

Kanun kabaca bildiğimiz günlük bir gerçeğin matematiksel ispatıdır da denebilir. Yazı-tura atarken yazı çıkma ihtimalinin 1/2 olduğunu biliyoruz (cunku zar atisini bir dagilim gibi goruyoruz). Herhalde çoğumuz da bu yazı-tura işleminin "bir çok kere" tekrarlandığı durumda, toplam sonucun aşağı yukarı yarısının yazı olacağını tahmin biliyoruz.

Matematiksel olarak, farzedelim ki her yazı-tura atışı bir deney olsun. Her ayrı deneyin sonucu $X_1, X_2 ... X_n$ olarak rasgelen değişkenlerle tanımlanmış olsun, bu degiskenlerin dagilimi ayni (cunku ayni zar), ama birbirlerinden bagimsizlar (cunku her deney digerinden alakasiz). Değişkenlerin sonucu 1 ya da 0 değeri taşıyacak, Yazı=1, Tura=0.

Buyuk Sayilar Kanunu tum bu deney sonuclarinin, yani rasgele degiskenlerin averaji alinirsa, yani $\bar{X} = X_1 + .. + X_n$ ile, elde edilen sonucun X_i'lerin (ayni olan) beklentisine yaklasacaginin soyler, yani n büyüdükçe $\bar{X}_n$'in 1/2'ye yaklaştığını ispatlar, yani $E[X_i] = 1/2$ degerine. Notasyonel olarak $E(X_i) = \mu$ olarak da gosterilebilir.

Formulsel olarak, herhangi bir $\epsilon > 0$ icin,

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \leqslant \epsilon) = 1$$

ya da

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

ya da

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0$$

Burada ne soylendigine dikkat edelim, X_i dagilimi *ne olursa olsun*, yani ister Binom, ister Gaussian olsun, *orneklem* uzerinden hesaplanan sayisal ortalamanin (empirical mean) formulsel olasilik beklentisine yaklastigini soyluyoruz! X_i'ler en absurt dagilimlar olabilirler, bu dagilimlarin fonksiyonu son derece cetrefil, tek tepeli (unimodal) bile olmayabilir, o formuller uzerinden beklenti icin gereken entegralin belki analitik cozumu bile mevcut olmayabilir! Ama yine de ortalama, o dagilimlarin beklentisine yaklasacaktir. Istatistik ile olasilik teorisi arasindaki cok onemli bir baglanti bu.

Sonuc sasirtici, fakat bir ek daha yapalim, sezgisel (intuite) olarak bakarsak aslinda sonuc cok sasirtici olmayabilir. Niye? Diyelim ki genel veri $N(\mu, \sigma^2)$ seklinde bir Normal dagilimdan geliyor ve orneklem de bu sebeple ayni dagilima sahip. Bu durumda orneklemdeki veri noktalarinin μ'ya yakin degerler olmasini beklemek mantikli olmaz mi? Cunku bu dagilim "zar atinca" ya da bir genel nufustan bir "ornek toplayinca" (ki bunu bir anlamda istatistiksel bir zar atisi olarak gorebiliriz) onu μ, σ^2'e gore atacak. Orneklemi zar atisi sonuclari olarak gordugumuze gore elde edilen verilerin bu sekilde olacagi sasirtici olmamali. Ve bu zar atislarinin ortalamasinin, son derece basit bir aritmetik bir islemle hesaplaniyor olsa bile, μ'ye yaklasmasi normal olmali.

Bu arada, bu argumana tersten bakarsak Monte Carlo entegralinin niye isledigine gorebiliriz (bkz *Monte Carlo, Entegraller, MCMC* yazisi).

Ozellikle orneklem ile genel nufus (population) arasinda kurulan baglantiya dikkat edelim. Istatigin onemli bir bolumunun bu baglanti oldugu soylenebilir. Her orneklem, bilmedigimiz ama genel nufusu temsil eden bir dagilimla ayni dagilima sahip olan X_i'dir dedik, ve bu ayniliktan ve bagimsizliktan yola cikarak bize genel nufus hakkinda bir ipucu saglayan bir kanun gelistirdik (ve birazdan ispatlayacagiz).

Ispata başlayalım.

$X_1, X_2, .., X_n$ bagimsiz degiskenler olsun.

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\bar{X}_n$ de bir rasgele değişkendir, çünku $\bar{X}_n$ değişkeni her X_i dağılımıyla alakali.

İspat devam etmek için, $\bar{X}_n$ dağılımının beklentisini bulmamız gerekiyor.

$$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i)$$

E dogrusal bir islec (linear operator) oldugu icin disaridan iceri dogru nufuz eder.

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu$$

$$= \mu$$

Dikkat edelim, bu *ortalamanin* beklentisi, ortalamanin kendisinin hangi degere yaklasacagini hala gostermiyor. Eger oyle olsaydi isimiz bitmis olurdu :) Daha yapacak cok is var.

Simdi $\bar{X}_n$ dağılımının standart sapmasını da bulalım. Diger bir olasilik kuramina gore

$$Y = a + bX$$

$$Var(Y) = b^2 Var(X)$$

oldugunu biliyoruz. O zaman,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Var(\bar{X}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \quad (3)$$

Artık Çebişev kuramını kullanmaya hazırız. Ispatlamaya calistigimiz neydi? $n \to \infty$ iken,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0$$

Cebisev'den

$$P(|\bar{X}_n - \mu| > \epsilon) \leqslant \frac{Var(\bar{X}_n)}{\epsilon^2}$$

$$P(|\bar{X}_n - \mu| > \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2} \to 0$$

$\sigma^2/n\epsilon^2$'in sifira gitmesi normal cunku n sonsuza gidiyor.

Peki $P(|\bar{X}_n - \mu| > \epsilon)$'nin sifira gittigini gosterdik mi?

$\sigma^2/n\epsilon^2$'nin sifira gittigini gosterdik. $\sigma^2/n\epsilon^2$ de $P(|\bar{X}_n - \mu| > \epsilon)$'den buyuk olduguna gore, demek ki o da sifira iner.

□

Buyuk Sayilar Kanunu orneklem ortalamasinin ve varyansinin X_i'in beklentisi ve varyansi ile baglanti kurar. Merkezi Limit Teorisi bir adim daha atar, ve der ki "$\bar{X}$'in dagilimi Gaussian dagilim olmalidir yani normal egrisi seklinde cikmalidir!". Teorinin detaylari bu bolumde bulunabilir.

Çebişev Eşitsizliği

Olasılık matematiğinde, büyük sayılar kuramı adında anılan ve olasılık matematiğinin belkemiğini oluşturan kuramı ispatlamak için, diğer bir kuram olan Çebişev eşitsizliğini de anlamamız gerekiyor. Çebişev eşitsizliği bir rasgele değişken, onun ortalaması (beklentisi) ve herhangi bir sabit sayı arasındaki üçlü arasında bir 'eşitsizlik' bağlantısı kurar, ve bu bağlantı diğer olasılık işlemlerimizde ispat verisi olarak işimize yarar.

Teori: Herhangi bir t değeri için,

$$P(|X - \mu| > t) \leqslant \frac{\sigma^2}{t^2}$$

İspata başlayalım. Entegral ile olasılık hesabı yapmak için bize bir x uzayı lazım.

$$\mathbb{R} = x : |x - \mu| > t$$

Yani $\mathbb{R}$ uzayı, x ile ortalamasının farkının, t'den büyük olduğu bütün sayıların kümesidir.

O zaman,

$$P(|X-\mu| > t) = \int_R f(x)dx$$

Dikkat edelim $P(..)$ içindeki formül, küme tanımı ile aynı. O yüzden $P()$ hesabı ortada daha olmayan, ama varolduğu kesin bir dağılım fonksiyonu tanımlamış da oluyor. Buna $f(x)$ deriz. $P()$'in, $f(x)$ fonksiyonunun R üzerinden entegral olduğunu olasılığa giriş dersinden bilmemiz lazım.

Eger $x \in R$ dersek o zaman

$$\frac{|x-\mu|^2}{t^2} \geqslant 1$$

t'nin denkleme bu şekilde nereden geldiği şaşkınlık yaratabilir. Daha önce tanımlanan şu ibareye dikkat edelim, $x : |x-u| > t$ diye belirtmiştik. Bu ifadeyi değiştirerek, yukarıdaki denkleme gelebiliriz.

Devam edersek, elimizdeki 1'den büyük bir değer var. Bu değeri kullanarak, aşağıdaki tanımı yapmamız doğru olacaktır.

$$\int_R f(x)dx \leqslant \int_R \frac{(x-\mu)^2}{t^2} f(x)dx \leqslant \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{t^2} f(x)dx$$

Ortadaki entegral niye birinci entegralden büyük? Çünkü ortadaki entegraldeki $f(x)dx$ ibaresinden önce gelen kısmın, her zaman 1'den büyük olacağını belirttiğimize göre, ikinci entegralin birinciden büyük olması normaldir, cunku birinci entegral $f(x)$ olasilik dagilimina bagli, entegral ise bir alan hesabidir ve olasilik dagilimlarinin sonsuzlar arasindaki entegrali her zaman 1 cikar, kaldi ki ustteki x'in uzayini daha da daralttik.

Evet...Üçüncü entegral ispata oldukça yaklaştı aslında. Standart sapma işaretini hala ortada göremiyoruz, fakat son entegraldeki ibare standart sapma değerini zaten içeriyor. Önce daha önceki olasılık natematiği bilgimize dayanarak, standart sapmanın tanımını yazıyoruz. Dikkat edelim, bu ibare şu anki ispatımız dahilinden değil, haricinden önceki bilgimize dayanarak geldi. Standart sapmanın tanımı şöyledir.

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

O zaman

$$\frac{\sigma^2}{t^2} = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{t^2} f(x)dx$$

yani

$$\int_R f(x)dx \leqslant \frac{\sigma^2}{t^2} = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{t^2} f(x)dx$$

ki $\int_R f(x)dx$ zaten $P(|X-\mu| > t)$ olarak tanimlanmisti.

□

Merkezi Limit Teorisi (Central Limit Theorem -CLT-)

Buyuk Sayilar Kanunu orneklem ortalamasinin gercek nufus beklentisine yaklasacagini ispatladi. Orneklem herhangi bir dagilimdan gelebiliyordu. CLT bu teoriyi bir adim ilerletiyor ve diyor ki kendisi de bir rasgele degisken olan orneklem ortalamasi $\bar{X}$ Normal dagilima sahiptir! Daha detaylandirmal gerekirse,

Diyelim ki $X_1, .., X_i$ orneklemi birbirinden bagimsiz, ayni dagilimli ve ortalamasi μ, standart sapmasi σ olan (ki o da ayni dagilima sahip) bir nufustan geliyorlar. Orneklem ortalamasi $\bar{X}$, ki bu rasgele degiskenin beklentisinin μ, ve (3)'e gore standart sapmasinin $\sigma/\sqrt{n}$ oldugunu biliyoruz. Dikkat: $\bar{X}$'in kendisinden degil, *beklentisinden* bahsediyoruz, BSK'deki ayni durum, yani ortalama dagiliminin ortalamasi. Teori der ki n buyudukce $\bar{X}$ dagilimi (bu sefer kendisi) bir $N(\mu, \sigma/\sqrt{n})$ dagilimina yaklasir.

Bu ifade genelde standart normal olarak gostrilir, herhangi bir normal dagilimi standart normal'e donusturmeyi daha once gormustuk zaten, beklentiyi cikartip standart sapmaya boluyoruz, o zaman orneklem dagilimi $\bar{X}$,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

dagilimina yaklasir diyoruz, ki $Z = N(0,1)$ dagilimidir, beklentisi sifir, standart sapmasi 1 degerindedir.

Bu teorinin ispatini simdilik vermeyecegiz.

Kaynaklar

[1] `http://mathworld.wolfram.com/MaximumLikelihood.html`

[2] Introduction to Probability and Statistics Using R