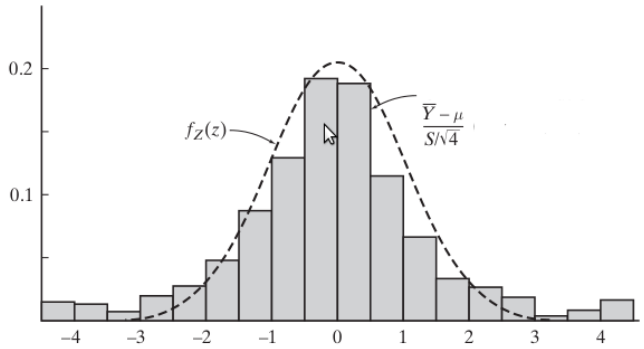
Orneklem Dagilimlari (Sampling Distributions)

$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ ve $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$ Karsilastirmasi

Diyelim ki normal olarak dagildigini bildigimiz bir nufustan $Y_1, .., Y_n$ rasgele orneklemini topladik, ve amacimiz bilinmeyen gercek μ hakkinda bazi sonuclara varmak. Eger varyans σ^2 biliniyorsa, bu noktadan sonra ne yapacagimiz gayet acik: daha once gordugumuz gibi bir karar kurali ortaya cikartmak, ya da guven araligi hesaplamak cok kolay, ki bu tekniklerin temelinde $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ dagiliminin standart normal $f_Z(z)$'ye yaklasmasi yatiyor.

Fakat pratikte σ^2 genellikle bilinmez, o zaman nufus varyansinin tahmin edicisi $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ kullanilir, ki bu maksimum olurluk tahmin edicisinin yansiz (unbiased) versiyonu. Fakat buradaki onemli soru su: σ^2 yerine S^2 koyma Z oranini nasil etkiler? Daha once buyuk orneklemler icin bir fark olmadigindan bahsettik. Peki kucuk orneklemler icin?

Kucuk n icin bu iki oraninin birbirinden farkli oldugununun kesfi William Sealy Gossett adli arastirmaciya ait. 1899'da Oxford'dan Kimya ve Matematik bolumunden mezun olduktan sonra Gossey, Guiness adli sirkette calismaya basladi. Urunlerin uzerinde yapacagi deneylerden aldigi veriler lojistik bazi sebepler dolasisiyla cok azdi, ve "gercek" σ^2'nin bilinmesi mumkun degildi. Cogu zaman n 4 ya da 5'den bile az oluyordu. Bu gibi durumlarla ugrasa ugrasa Gossey $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$'nin beklendigi gibi can egrisi $f_Z(z)$ seklinde degil, daha "etekleri kabarik" baska bir dagilim gibi gozuktugunu farketti, yani sifirdan cok kucuk ya da ondan cok buyuk oranlarin ihtimali cok dusuk degildi.



Ustteki histogram S kullanarak hesaplanmistir, $n = 4$ olmak uzere 500 orneklem uzerinden hesap yapilmistir. Iki dagilimin birbirinden uzaklastigi goruluyor.

Turetmek

Genel olarak olasilik dagilimlari iki buyuk kategori altina duser. Asagi yukari bir duzine kadari gercek dunyadan alinabilecek her olcumu oldugu haliyle iyi modelleme kabiliyetine sahiptir; mesela normal, binom, Poisson, ustel dagilimlar gibi. Diger yandan daha az sayida (ama bir o kadar onemli) dagilimlar n tane rasgele degiskenin uzerinden hesaplanan *fonksiyonlarin* nasil davrandigini cok iyi

modeller. Iste bu dagilimlara orneklem dagilimlari ismi verilir ve tipik kullanim alanlari cikarsama (inference) yapmaktir.

Normal dagilimi her iki kategoriye de aittir. Hem ayri ayri olcumleri modellemek, hem de $T = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$'in olasiliksal davranisini modellemek icin kullanilir. Ikinci kullanimi normal dagilimin bir orneklem dagilimi olarak kullanilmasina ornektir.

Normal dagilimdan sonra en onemli uc orneklem dagilimi Ogrenci t Dagilimi, chi kare dagilimi ve F dagilimidir. Son iki dagilim t oranini temsil eden $f_T(t)$'yi, yani $T = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$'yi turetmek icin gerekli.

χ^2 Dagilimi

Tanim

$Z_1, .., Z_p$ bagimsiz standart Normal rasgele degiskenler ise, $U = \sum_{i=1}^{p} Z_p^2 \sim \chi_p^2$ ki bu dagilima p derecede serbestlige (degrees of freedom) olan chi kare dagilimi (chi square distribution) ismi verilir.

Teori

U, p derece serbestlige sahip bir χ^2 dagilima sahip ise, ki $U \sim \chi_p^2$ olarak gosterilir, yogunluk

$$f_U(u) = \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} u^{(p/2)-1} e^{-u/2}$$

$$u \geqslant 0$$

formulune esittir. Ustteki yogunlugun $r = m/2$ ve $\lambda = 1/2$ olan bir Gamma dagilimi oldugu da soylenebilir. Ispat icin [1]'e bakiniz.

Teori

$Y_1, .., Y_n$ ortalamasi μ, varyansi σ^2 olan bir normal dagilimdan alinan n orneklem olsun. O zaman

a. S^2 ve $\bar{Y}$ birbirinden bagimsizdir

b. $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2)$ hesabi $n-1$ derece serbestligike sahip bir chi kare dagilimidir.

Ispat icin [1].

Tanim

Z bir standart normal rasgele degisken, U ise n derece serbestlikteki bir chi kare rasgele degisken olsun. O zaman n derece serbestligi olan Ogrenci t orani (Student's t ratio)

$$T_n = \frac{Z}{\sqrt{\frac{U}{n}}}$$

olarak belirtilir.

t Dagilimi (Student's t)

X, n derece bagimsizlikta t dagilimina sahiptir, ve dagilimi

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}$$

Aslinda Normal dagilimi t dagiliminin $\nu = \infty$ oldugu hale tekabul eder. Cauchy dagilimi da t'nin ozel bir halidir, $n = 1$ halidir. Bu durumda yogunluk fonksiyonu

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

Bu formul hakikaten bir yogunluk mudur? Kontrol icin entegralini alalim,

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

Cogunlukla entegre edilen yerde "1 arti ya da eksi bir seyin 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

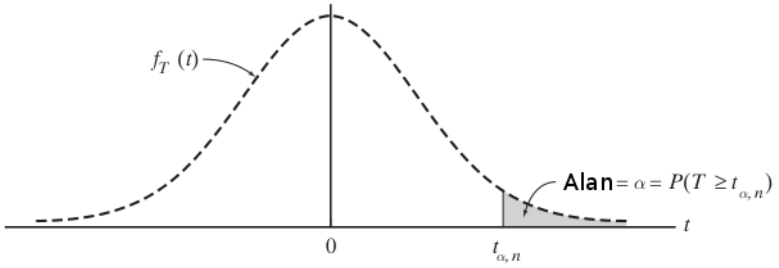
$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{dx}{1+x^2} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{\sec^2\theta}\sec^2\theta d\theta = \frac{1}{\pi}\int_{-\infty}^{\infty}1\,d\theta =$$

$$= \frac{1}{\pi}\theta|_{-\infty}^{\infty} = \frac{1}{\pi}[\arctan(\infty) - \arctan(-\infty)]$$

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

Guven Araliklari

Daha once Z oranini temel alarak guven araliklari ya da hipotez testleri olusturmustuk. Bu islemler icin standart normal dagilimin ust ve alt yuzdelikleri hakkinda bazi bilgiler gerekmisti. Bu bilgiler bir tablodan bakilan degerlerdi ya da istatistik yazilimimizda gerekli bir cagri ile halledilmisti.



Ogrenci t'nin Z'ye gore farkli bir tarafi belli bir degeri bulmak icin iki parametreye ihtiyac olmasi, bunlardan biri α digeri ise serbestlik derecesi. t icin artik tablolarla ugrasmayacagiz, 21. yuzyildayiz, yazilim ile bu isi halledelim! Mesela T bir Ogrenci t dagilimi ise, ve serbestlik derecesi 3 ise, $\alpha = 0.01$ icin icin $f_T(t)$'nin $100(1-\alpha)$ yuzdeligi nedir?

```
from scipy.stats.distributions import t
df = 3
print t.ppf(0.99,df)
print 1-t.cdf(4.541,df)
```

```
4.5407028587
0.00999823806449
```

Yani

$$P(T_3 \geqslant 4.541) = 0.01$$

Kaynaklar

Larsen, *Introduction to Mathematical Statistics and Its Applications*