

Guven Araliklari, Hipotez Testleri

Guven Araliklari

Diyelim ki X_1, \dots, X_n orneklemi birbirinden bagimsiz, ayni dagilimli ve ortalamasi μ , standart sapmasi σ ve yine ayni olan bir nufus dagilimindan geliyor. O zaman biliyoruz ki, Merkezi Limit Teorisi (Central Limit Theorem) teorisine gore, orneklem ortalamasi $\bar{X} = \frac{1}{n}X_1 + \dots + X_n$, ortalamasi μ , standart sapmasi σ/\sqrt{n} olan bir normal dagilima yaklasiyor.

Peki veriyi (yani orneklemi) ve CLT'yi kullanarak μ hakkında bir tahmin yapabilir miyiz? Yani Buyuk Sayilar Kanununa gore μ hakkında noktasal tahmin yapabiliriz fakat, belki ondan bir adim otesi, bir "guven araligi" hesaplamaktan bahsediyoruz. Bu tahmin "gercek μ , %95 ihtimalde su iki deger arasindadir" turunde bir tahmin olacak.

Bu araligin hesabi icin once \bar{X} 'i standardize edelim, yani $N(0,1)$ haline cevirelim,

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

Z-skorlarini isledigimiz yazida

$$P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$$

gibi bir ifade gorduk. Esitligin sag tarafi aslinda bir alan hesabidir, surekli fonksiyonlarda olasilik bir entegral, ya da iki kumulatif yogunluk fonksiyonunun farki. Guven araligi icin bize lazim olan da bir olasilik, hatta "kesin" bir olasilik, %95 olasiligi. Demek ki esitligin sag tarafi .95 olacak. .95 hesabi icin, normal egrisini dusunursek, sagindan ve solundan 0.025 buyuklugunde iki parçayı "kirpmamız" lazim. O zaman 0.975 olasiliginin z degeri ile, 0.025 olasiliginin z degeri arasindaki olasilikta olmamız lazim. Bu hesaplarda baz alinan $z_{\alpha/2}$ degeri ve bu $100 \cdot \alpha/2$ ust yuzdelik kismina, ornegimizde 0.975 kismina tekabul ediyor. Normal dagilimin simetrisi sebebiyle onun eksisi alinmis hali oteki (soldaki) parçayı verir, yani $-z_{\alpha/2}$.



z-skoru hesaplarırken tabloya danismistik, simdi tabloya tersinden bakacagiz, kesisme noktasinda 0.975 diyen yeri bulup kordinatlari alacagiz, ki bu deger 1.96. Istatistik kaynaklarinda “sihirli deger” seklinde tarif edilen bir deger bu, gozlerimiz kamasmasin, geldiği yer burasi iste. Simdi formulu buna gore degistirelim,

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$P(\cdot)$ icinde biraz duzenleme, tum terimleri σ/\sqrt{n} ile carpalim, \bar{X} cikartalim, ve -1 ile carpalim,

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Guven araligi ifadesine aslina erismis olduk. Eger %95 kesinlikten bahsediyor olsaydik, ve nufusun gercek varyansi σ^2 biliniyor olsaydi, $P(\cdot)$ icine bu degerleri gececektik, \bar{X} zaten verinin aritmetik ortalamasindan ibarettir, bu bize μ 'nun solunda ve saginda bazi degerler dondurecekti. Bu degerler bizim guven araligimiz olacakti. Mesela veri 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3 seklinde, $n = 10$ cunku 10 nokta var, $\sigma = 1$ olarak verilmiş. Ortalamayi hesapliyoruz, 64.46. $\alpha = 0.05$ icin

$$P\left(64.46 - 1.96 \frac{1}{\sqrt{10}} \leq \mu \leq 64.46 + 1.96 \frac{1}{\sqrt{10}}\right) = 0.95$$

$$P\left(63.84 \leq \mu \leq 65.08\right) = 0.95$$

Yani %95 guven araligi $63.84 \leq \mu \leq 65.08$.

Neler yaptik? CLT bilgisinden hareketle \bar{X} hakkında bir seyler biliyorduk. Fakat \bar{X} 'in kesin hangi normal dagilima yaklastigini bilmek icin nufus paremetreleri μ, σ da bilinmelidir. Diger yandan eger tek bilinmeyen μ ise, teoriyi bu bilinmez

etrafında tamamen tekrar sekillendirip / degistirip CLT'yi bilinmeyen μ etrafında bir guven araligi yaratmak icin kullandik.

Kac Tane n ?

Hatirlarsak guven araligini ustteki sekilde hesaplayabilmemizin sebebi CLT sayesinde \bar{X} 'in normal dagilima yaklasiyor olmasiydi. Ve, teoriyi tekrar dusunursek yaklasma $n \rightarrow \infty$ oldugu zaman oluyordu. Buradan \bar{X} 'in normalliginin “buyukce” n degerleri icin daha gecerli olacagi sonucuna varabiliriz. Peki n ne kadar buyuk olmalı? Literature gore CLT'nin genellikle $n \geq 30$ durumunda gecerli oldugu soylenir. Tabii nufus dagiliminin ne oldugu da onemlidir, eger nufus normal ise, ya da genel olarak simetrik tek tepeli dagilim ise orneklem daha ufak kalsa da bazi sonuclara varabiliriz. Eger nufus dagilimi cok yamuk (skewed), etekleri genis dagilim ise o zaman daha buyuk orneklem daha iyi olur.

Soru

IO 800 yillarında İtalya'da Etrusali (Etruscan) toplumu vardı. Bu toplum geldigi gibi birdenbire ortadan kayboldu. Bilimciler bu toplumun İtalyalılar ile fizyolojik, genetik ve kulturel olarak baglantisi olup olmadigini hep merak etmistir. Bazilari hafa olculerine bakarak sonuclara varmaya ugrasmistir. Arkeolojik kazılarda yapılan olcumlerde 84 Etrusyalinin kafasi olculmustur. Ayrica bugunku İtalyanların kafa olcumlerinin normal dagilimında $\mu = 132.4\text{mm}$, $\sigma = 6.0\text{mm}$ oldugu bilinmektedir. O zaman, veriye bakarak kafa olcumu ortalamasi icin bir %95 guvenlik araligi olusturuz, ve eger bugunku İtalyanların olcusu o araliga dussuyorsa, bugunku İtalyanların Etrusyalılarla baglantisinin olmadigi iddia edilebilir.

```
import pandas as pd
df = pd.read_csv('etrus.csv')
print float(df.mean() - 1.96 * (6.0/np.sqrt(84)))
print float(df.mean() + 1.96 * (6.0/np.sqrt(84)))

142.524107721
145.09035011
```

Bugunku İtalyanların kafa ortalamasi $\mu = 132.4$ bu araliga dussuyor. Diger bir deyisle, 84 tane orneklemde gelen orneklem ortalamasi 143.8 buyuk bir ihtimalle $\mu - 132.4$, $\sigma = 6.0$ boyutlarındaki bir normal dagilimdan gelmemistir. Buna gore, buyuk bir ihtimalle Etrusyalılar İtalyanların atası degildir.

Bilinmeyen σ

Eger σ bilinmiyorsa, bu durumda σ yerine orneklem varyansi S kullanilabilir,

$$s^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

ki ustteki degerin karekoku S olacaktir. σ yerine S kullanmanın buyuk n degerlerinde CLT'yi etkilemedigi ispat edilmistir [5].

Binom Dagilimler ve Normal Yaklasiklik

Binom ile Bernoulli dagilimi arasindaki baglantiyi biliyoruz. Diyelim ki X_1, \dots, X_n birbirinden bagimsiz ve ayni Bernoulli olarak dagilmis, Bernoulli dagilimini temsil eden Y tanımlayalım, o zaman

$$Y = \sum_{i=1}^n X_i$$

Simdi orneklem ortalamasini hatirlayalım,

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

O zaman

$$Y = n\bar{X}$$

Merkezi Limit Teorisinden \bar{X} 'in nufus beklentisi ve sapmasini iceren $N(\mu, \sigma)$ olarak dagilacagini biliyoruz. Nufus parametreleri nedir? Her X_i 'in ayni olan μ, σ 'si ile alakli, durumda Bernoulli parametrelerini alip $N(\cdot)$ icinde direk kullanabiliriz,

$$E(X_i) = p, \text{Var}(X_i) = p(1 - p)$$

o zaman

$$\bar{X} \sim N(\mu, \sigma), \mu = p, \sigma = \sqrt{\frac{p(1 - p)}{n}}$$

Y ile \bar{X} baglantisi: Bir genel teoriye gore eger \bar{X} normal ise $n\bar{X}$ 'in de normal oldugu bilinir, ve bu dagilim $N(n\mu, \sqrt{n}\sigma)$ olarak gosterilir. Bu teorinin ispatini simdilik vermeyecegiz. O zaman $Y = n\bar{X}$ is ve normal olarak dagilmis ise, o zaman

$$Y \sim N\left(np, \sqrt{np(1 - p)}\right)$$

demek dogru olacaktir. Standardize etmek gayet basit,

$$Z = \frac{Y - np}{\sqrt{np(1 - p)}}$$

ya da, bolum ve boleni n ile bolerseniz,

$$Z = \frac{Y/n - p}{\sqrt{p(1 - p)/n}}$$

$$Z = \frac{Y/n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Soru

Amerikalıların yüzde 12'sinin zenci olduğunu biliyoruz. Eğer 1500 kişiyi içeren bir örneklem alsaydık, bu örneklemde 170'den daha az zenci olmasının olasılığı nedir?

Cevap

%12 nüfus parametresidir, yani $p = 0.12$. Örneklem $n = 1500$. Normal yaklaşım ile

```
from scipy.stats import norm
n = 1500
p = 0.12
mu = n*p
std = np.sqrt(n*p*(1-p))
print mu, std
print 'olasilik', norm.cdf(170, loc=mu, scale=std)
180.0 12.585706178
olasilik 0.213437028747
```

Yani $N(180, 12.58)$ dağılımını elde ettik ve hesapları onun üzerinden yaptık. Sonuç diyor ki verilen örneklem ve nüfus p değeri ile 170 altında zenci sayısı elde etmek oldukça düşük bir ihtimalde.

Binom Güven Aralığı

Eğer p bilinmiyor ise onun için maksimum olurluk tahmin edicisi (maximum likelihood estimator) Y/n 'dir [ispatı simdilik vermiyoruz].

$$Z = \frac{X/n - p}{\sqrt{\frac{(X/n)(1-(X/n))}{n}}}$$

Bu durumda Z üzerinden, aynen daha önce yaptığımız gibi, bir güvenlik aralığı tanımlayabiliriz. Başlamak için

$$P\left(-z_{\alpha/2} \leq \frac{X/n - p}{\sqrt{\frac{(X/n)(1-(X/n))}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

ve yine daha önceki benzer cebirsel işlemler sonrası, ve Binom deneydeki başarı sayısı olarak X yerine k kullanalım, $P()$ ifadesini çıkartalım, çünkü zaten o ifadenin içinde oluşacak sayılarla ilgileniyoruz,

$$\left(\frac{k}{n} - z_{\alpha/2} \sqrt{\frac{(k/n)(1-(k/n))}{n}}, \frac{k}{n} + z_{\alpha/2} \sqrt{\frac{(k/n)(1-(k/n))}{n}}\right)$$

Ustteki iki sayi bize gerekli guven araligini verecektir.

Soru

Amerika’da 2009 yilinda halkin ne kadarinin arabalarinda yakit tasarrunu destekledigi merak konusuydu. Bir Gallup telefon anketinde bu soru 1012 yetiskine (18 ve ustü yasta) bu soruyu sordu. Cevap 810 kisinin tasarrufu destekledigi yonundaydi. Yani $n = 1012$, $k = 810$. O zaman p icin %95 guven araligini bulun.

Cevap

$$\left(\frac{810}{1012} - 1.96 \frac{(810/1012)(1 - 810/1012)}{1012}, 1.96 \frac{(810/1012)(1 - 810/1012)}{1012} \right)$$
$$= (0.776, 0.825)$$

Hipotez Testleri (Hypothesis Testing)

Istatistik tek ya da araliklar olarak sayisal tahminler uretmenin otesinde, “iki sey arasinda birisini secmek” turunde bir karar baglaminda da kullanilabilir. Bir psikolog bir davaya uzman gorus vermek icin cagrilmistir ve sanik hakkında ‘akli olarak dengersiz ya da dengeli’ arasinda bir secim yapacaktır. Ilac regulasyonu ile ugrasan kurum yeni bir ilac hakkında ‘etkili’ ya da ‘etkisiz’ seklinde bir karara ulasacaktır.

Bir deneyin mumkun sonuclarini belli seceneklere yonlendirip olasilik teorisini kullanarak bunlardan birisini secmeye Istatistik biliminde Hipotez Test Etmek adi verilir.

Birbiriyle yaris halinde olan iki hipotez vardir, bunlar sifir hipotezi (H_0 olarak yaziliyor) ve alternatif hipotezdir (H_1 olarak yaziliyor). H_0 ve H_1 arasinda nasil secim yapacagimiz kavramsal olarak bir davada jurinin yaptigi secime benzer: aynen sanigin, tersi ispatlanana kadar, masum kabul edilmesi gibi eger veri tersi sonuca varmaya yetmezse H_0 da “kabul edilir”, yani sucsuzlugun devam etmesi gibi H_0 gorusu terkedilmemis olur. Statisko devam eder. Bu karari verirken mahkemenin kanitlari incelemesi, hipotez testinde rasgele degiskenlerle verinin uzerinden hesaplar yapmaya benzer.

Bunu bir ornek uzerinden daha iyi anlayabiliriz. Diyelim ki araba ureten bir sirket yakit performansini (gas mileage) arttirmaya ugrasiyor. Benzine katilan yeni bir madde uzerinde deneyler yapiliyorlar, deney icin Boston / Los Angeles arasinda 30 tane araba sefer yapiliyor. Yeni katki maddesi olmadigi durumda (statisko) yakit performansinin ortalama 25.0 mil/galon ve standart sapmanin 2.4 mil/galon oldugu biliniyor. Diyelim ki deney sonrasinda arabalar ortalama olarak $\bar{y}=26.3$ mil/galon performansi gostermisler. Katki maddesi etkili mi, etkili degil mi?

Arastirmacilar 25.0’dan 26.3’e olan degisikligi daha once bahsettigimiz mahkeme

ornegindeki gibi bir cercevede incelerler. Tipik olarak sifir hipotezi statukoyu temsil eder, yani degismesi icin “ezici sekilde aksi yonde veri olmasi gereken sey” budur. Oyle degil mi? Eger etkisiz bir katkı maddesine evet dersek, ve ileride oyle olmadigi belli olursa bunun sirket icin cok negatif etkileri olacaktır, aynen masum bir kisiyi yanlışlıkla hapse atmış olmak gibi. O yüzden kalmak istedigimiz güvenli konum H_0 ’i temsil etmelidir.

Bu noktada problemi rasgele degiskenlerin terminolojisi uzerinden tekrar tanımlamak faydalı olur. Diyelim ki test sırasında 30 tane aldığımız ölçüm y_1, \dots, y_n , her y_i normal olarak dağılmış ve bu dağılımların μ ’su aynı, ve μ ’u birazdan “eski” ölçümlerin ortalaması olarak alacağız, çünkü curutmek istedigimiz hipotez bu. Ayrıca daha önceki tecrübelerimiz gösteriyor ki $\sigma = 2.4$. Yani,

$$f_Y(y; \mu) = \frac{1}{\sqrt{2\pi}(2.4)} e^{-\frac{1}{2}(\frac{y-\mu}{2.4})^2}, -\infty < y < \infty$$

Hipotezleri şöyle tanımlayalım,

$H_0: \mu = 25.0$ (Katkı maddesi etkili *degildir*)

$H_0: \mu > 25.0$ (Katkı maddesi etkilidir)

Şimdi yeni dağılımı standardize edip, bir hayali ortalama eşik değeri uzerinden bir sonuç çıkartalım, standardize etmek için kullandığımız $\mu = 25.0$ çünkü eski ortalama bu. Şimdi diyelim ki test ettiğimiz eşik değeri 25.25 (esas amaç 26.3 ama oraya geleceğiz), aradığımız olasılık,

$$P(\bar{Y} \geq 25.25)$$

Üstteki ifade “eğer örneklem eski dağılımdan geliyor olsaydı, 25.25 eşik değerini geçmesi ne kadar mümkün olabilirdi” diye bir soru soruyor. \bar{Y} ’yi standardize edelim, o sırada eşitsizliğin sağ tarafı da değişir,

$$P\left(\frac{\bar{Y} - 25.0}{2.4/\sqrt{30}} \geq \frac{25.25 - 25.0}{2.4/\sqrt{30}}\right)$$

$$P(Z \geq 0.57)$$

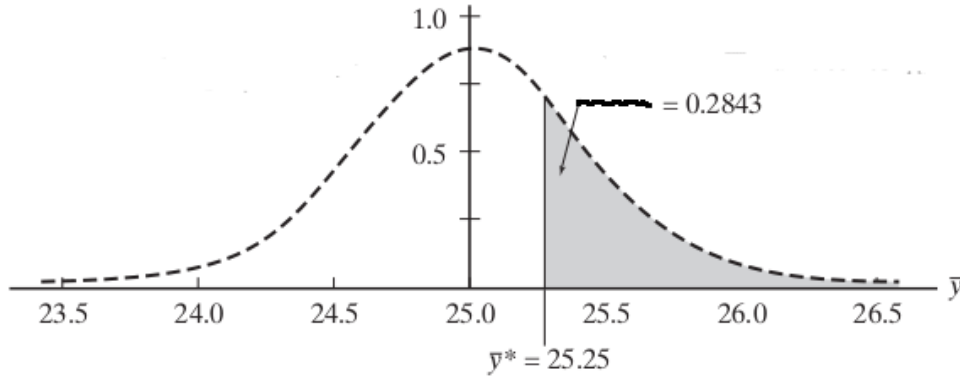
z-Skoru tablosunu kullanarak bu hesabi yapmak için

$$1 - P(Z < 0.57)$$

0.57’nin z-skoru (satır 0.5 kolon .07) 0.7157 olarak gösterilmiş, o zaman $1 - 0.7157 = 0.2843$. Demek ki

$$P(Z \geq 0.57) = 0.2843$$

Demek ki yeni deney sonuclarinin, eski dagilima gore, esik degerinden fazla gelmesi hala az da muhtemel, demek ki eski hipotezi tam curutemedik. Sectigimiz esik degeri bize kesin bir sonuc saglamadi, sezgisel olarak bu olasiligin buyuk oldugunu goruyoruz. Mahkeme durumunda sucsuz olmasi cok muhtemeldir diyemiyoruz. Ya da araba orneginde (ve pozitif baglamda) yeni yakit kesinlikle farklidir / fazladir diyemiyoruz. Bize daha kesin noktalar lazim, aklimizda bize “acaba?” dedittirecek esik degerler istemiyoruz.



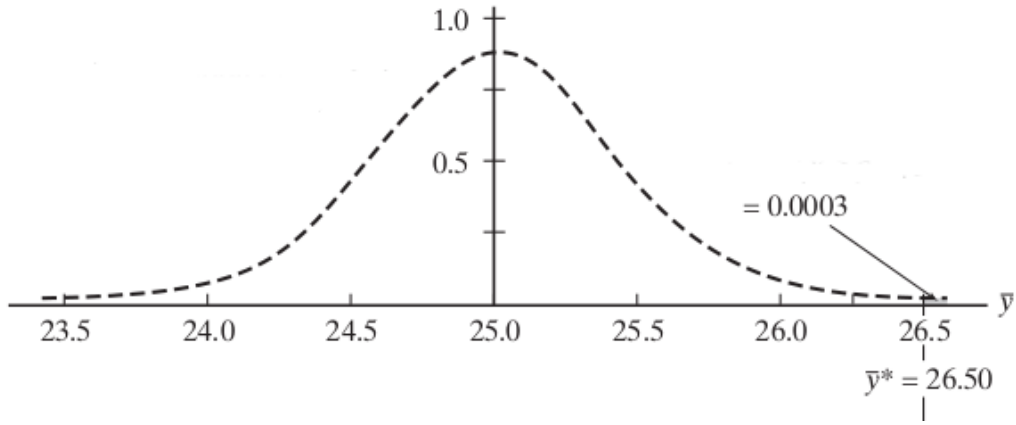
Hayali esik noktası \bar{y}^* 'nin daha buyuk yapsak (ki o zaman ona bagli olan sagdaki olasilik kuculecek). Bu olur mu? Eger $\bar{y}^* = 26.50$ olsaydi?

$$P\left(\frac{\bar{Y} - 25.0}{2.4/\sqrt{30}} \geq \frac{26.50 - 25.0}{2.4/\sqrt{30}}\right)$$

$$P(Z \geq 3.42)$$

$$= 0.0003$$

Bu olasilik ise cok kucuk, yani esik degeri cok buyuk! Citayi cok fazla kaldirdik, mahkeme durumunda sanki diyoruz ki sucun 1000 tane tanigi lazim, sanik su-cunu itiraf etmis olmal, hersey apacik olmal, bir de herseyi bizzat ben gor-mus olmaliyim, yoksa kabul etmem. Araba orneginde katkı maddesi arabaya Formula-1 yarisi kazandirmazsa biz bu yakiti daha iyi olarak kabul etmeyiz diy-yoruz.



Peki eger 0.28 çok fazla, 0.0003 çok küçük ise hangi olasılık en iyi esik değerini verir? Bu soruya kesin olarak ve matematiksel bir cevap vermek mümkün değil, fakat hipotez test etme teknigini kullanan araştırmacıların ulaştığı konsensus 0.05 olasılık seviyesinin en iyi sonuçlar verdiğidir. Bu duruma sıfır hipotezinin çok kolayca kenara atılmaması, ya da ona gereğinden fazla bağlı kalınmaması mümkün oluyor.

O zaman 0.05 olasılığını verdirtecek esik değeri hesaplayalım,

$$P\left(\frac{\bar{Y} - 25.0}{2.4/\sqrt{30}} \geq \frac{\bar{y}^* - 25.0}{2.4/\sqrt{30}}\right) = 0.05$$

$$P(Z \geq \frac{\bar{y}^* - 25.0}{2.4/\sqrt{30}}) = 0.05$$

ya da

$$P(Z \leq \frac{\bar{y}^* - 25.0}{2.4/\sqrt{30}}) = 0.95$$

z-Skor tablosuna bakıyoruz, “hangi z değeri 0.95 değeri sonucunu verir”, kordinalardan 1.64 z-skorunu buluyoruz.

$$P(Z \leq 1.64) = 0.95$$

O zaman

$$\frac{\bar{y}^* - 25.0}{2.4/\sqrt{30}} = 1.64$$

ve buradan $\bar{y}^* = 25.178$ sonucu çıkıyor. 26.3 değeri bu değerden yüksektir demek ki sıfır hipotezi curutulmuştur. Yeni yakıt katkısının performansı arttırıyor olması büyük bir olasılıktır.

Not: Bu testi aslında daha basit şekilde $\bar{y}^* = 26.3$ değerlerini vererek elde edilen değeri 0.05'ten küçük olup olmadığına bakarak ta yapabiliydik. Fakat metodu inşa ediyorduk o sebeple daha fazla örnekli anlatmak gerekti.

Binom Testi

Bazen Hipotez testi (bir veriye dayanarak) farzedilen bir parametreyi bir sabit değere karşılaştırmak, ya da iki parametreyi birbiriyle karşılaştırmak için kullanılır.

Diyelim ki elimizde bir Web sitesinin günlük ziyaret, tıklama sayılarını gösteren bir veri seti var (CVR ziyaretçilerin, sitedeki tıklayan müşteriye "cevirme" oranı, -conversion-)

```
import pandas as pd
from scipy import stats
a = pd.DataFrame({'tiklama': [20., 2., 40., 5., 10., 100.],
                  'ziyaret': [100., 10., 300., 400., 30., 800.]})
a['cvr'] = a['tiklama'] / a['ziyaret']
print a
```

	tiklama	ziyaret	cvr
0	20	100	0.200000
1	2	10	0.200000
2	40	300	0.133333
3	5	400	0.012500
4	10	30	0.333333
5	100	800	0.125000

Bu veri seti için cvr'in 0.16, yani yüzde 16 olduğunu önceden biliyoruz. Üstteki başarı oranı binom dağılı ile modellenabilir, ziyaretler "deneylerdir", yani örneklem büyüklüğünü gösterirler. Tıklama ise başarıdır,

```
p_hat = 0.16
btest = lambda x: (x['cvr']-p_hat) / np.sqrt( p_hat*(1-p_hat)/x['ziyaret'])
a['guven'] = a.apply(btest, axis=1)
a['guven'] = np.round(stats.zprob(a['guven'])*100, 2)
print a
```

	tiklama	ziyaret	cvr	guven
0	20	100	0.200000	86.24
1	2	10	0.200000	63.50
2	40	300	0.133333	10.39
3	5	400	0.012500	0.00
4	10	30	0.333333	99.52
5	100	800	0.125000	0.35

Tek Örneklem t Testi (One-sample t test)

Bu test verinin Normal dağılımdan geldiğini farzeder, tek örneklem durumunda elde x_1, \dots, x_n verisi vardır, ve bu veri $N(\mu, \Sigma)$ dağılımından gelmiştir ve test etmek istediğimiz hipotez / karşılaştırma $\mu = \mu_0$.

```
from scipy.stats import ttest_1samp, wilcoxon, ttest_ind
import pandas as pd
```

```

daily_intake = np.array([5260,5470,5640,6180,6390,6515, 6805,7515,7515,8230,8770])
df = pd.DataFrame(daily_intake)
print df.describe()

```

```

              0
count      11.000000
mean      6753.636364
std       1142.123222
min       5260.000000
25%       5910.000000
50%       6515.000000
75%       7515.000000
max       8770.000000

```

```

t_statistic, p_value = ttest_1samp(daily_intake, 7725)
print "one-sample t-test", p_value

```

```

one-sample t-test 0.0181372351761

```

Sonuc p_value 0.05'ten küçük çıktı yani yüzde 5 önemliliğini (significance) baz aldık bu durumda veri hipotezden önemli derecede (significantly) uzakta. Demek ki ortalamanın 7725 olduğu hipotezini reddetmemiz gerekiyor.

Testi iki örneklemli kullanalım, gruplar 0/1 değerleri ile işaretlendi, ve test etmek istediğimiz iki grubun ortalamasının (mean) aynı olduğu hipotezini test etmek. t-test bu arada varyansın aynı olduğunu farzeder.

```

energ = np.array([
[9.21, 0],
[7.53, 1],
[7.48, 1],
[8.08, 1],
[8.09, 1],
[10.15, 1],
[8.40, 1],
[10.88, 1],
[6.13, 1],
[7.90, 1],
[11.51, 0],
[12.79, 0],
[7.05, 1],
[11.85, 0],
[9.97, 0],
[7.48, 1],
[8.79, 0],
[9.69, 0],
[9.68, 0],
[7.58, 1],
[9.19, 0],
[8.11, 1]])
group1 = energ[energ[:, 1] == 0][:, 0]
group2 = energ[energ[:, 1] == 1][:, 0]
t_statistic, p_value = ttest_ind(group1, group2)
print "two-sample t-test", p_value

```

```
two-sample t-test 0.00079899821117
```

$p - \text{value} < 0.05$ yani iki grubun ortalamasi ayni degildir. Ayni oldugu hipotezi reddedildi.

Eslemeli t-Test (Paired t-test)

Eslemeli testler ayni deneysel birimin olcumu alindigi zaman kullanilabilir, yani olcum alinan ayni grupta, deney sonrasi deneyin etki edip etmedigi test edilebilir. Bunun icin ayni olcum deney sonrasi bir daha alinir ve "farklarin ortalamasinin sifir oldugu" hipotezi test edilebilir. Altta bir grup hastanin deney oncesi ve sonrasi ne kadar yiyecek tukettigi listelenmis.

```
intake = np.array([
    [5260, 3910],
    [5470, 4220],
    [5640, 3885],
    [6180, 5160],
    [6390, 5645],
    [6515, 4680],
    [6805, 5265],
    [7515, 5975],
    [7515, 6790],
    [8230, 6900],
    [8770, 7335],
])
pre = intake[:, 0]
post = intake[:, 1]
t_statistic, p_value = ttest_1samp(post - pre, 0)
print "paired t-test", p_value

paired t-test 3.05902094293e-07
```

Wilcoxon isaretli-sirali testi (Wilcoxon signed-rank test)

t Testleri Normal dagilima gore sapmalari yakalamak acisindan, ozellikle buyuk orneklem var ise, oldukca saglamdir. Fakat bazen verinin Normal dagilimdan geldiği faraziyesini yapmak istemeyebiliriz. Bu durumda *dagilimdan bagimsiz metotlar* daha uygundur, bu tur metotlar icin verinin yerine cogunlukla onun sıra istatistiklerini (order statistics) kullanir.

Tek orneklemli Wilcoxon testi icin prosedür μ_0 'i tum veriden cikartmak ve geri kalan (farklari) isaretine bakmadan numerik degerine gore siralamak, ve bu sıra degerini bir kenara yazmak. Daha sonra geri donup bu sefer cikartma islemi sonucunun isaretine bakmak, ve eksi isareti tasiyan sıra degerlerini toplamak, ayni islemi arti isareti icin yapmak, ve eksi toplami arti toplamindan cikartmak. Sonucta elimize bir istatistik W gelecek. Bu test istatistigi aslinda $1..n$ tane sayi icinden herhangi birini $1/2$ olasiligiyla secmek, ve sonuclari toplamaya tekabul etmektedir. Ve bu sonuc yine 0.05 ile karsilastirilir.

```
z_statistic, p_value = wilcoxon(daily_intake - 7725)
print "one-sample wilcoxon-test", p_value
```

```
one-sample wilcoxon-test 0.0279991628713
```

Hipotezi yine reddettik.

Ustte yaptigimiz eslemeli t-testi simdi Wilcoxon testi ile yapalim,

```
z_statistic, p_value = wilcoxon(post - pre)
print "paired wilcoxon-test", p_value

paired wilcoxon-test 0.00463608893545
```

Gaussian Kontrolu

Diyelim ki Gaussian dagilimina sahip oldugunu dusundugumuz $\{x_i\}$ verilerimiz var. Bu verilerin Gaussian dagilimina uyup uymadigini nasil kontrol edecegiz? Normal bir dagilimin her veri noktası için şöyle temsil edebiliriz,

$$y_i = \Phi\left(\frac{x_i - \mu}{\sigma}\right)$$

Burada Φ standart Gaussian'ı temsil ediyor (detaylar için *Istatistik Ders 1*) ve CDF fonksiyonuna tekabül ediyor. CDF fonksiyonunun aynı zamanda ceşregi (quantile) hesapladığı söylenir, aslında CDF son derece detaylı bir olasılık değeri verir fakat evet, dolaylı yoldan noktanın hangi ceşrek içine dustugu de görülecektir.

Simdi bir numara yapalim, iki tarafa ters Gaussian formülünü uygulayalım, yani Φ^{-1} .

$$\Phi^{-1}(y_i) = \Phi^{-1}\left(\Phi\left(\frac{x_i - \mu}{\sigma}\right)\right)$$

$$\Phi^{-1}(y_i) = \frac{x_i - \mu}{\sigma}$$

$$x_i = \Phi^{-1}(y_i)\sigma + \mu$$

Bu demektir ki elimizdeki verileri $\Phi^{-1}(y_i)$ bazında grafiklersek, bu noktalar eğimi σ , başlangıcı (intercept) μ olan bir düz çizgi olmalıdır. Eğer kabaca noktalar düz çizgi oluşturmuyorsa, verimizin Gaussian dağılıma sahip olmadığına karar verebiliriz.

Ustte tarif edilen grafik, olasılık grafiği (probability plot) olarak bilinir.

Ters Gaussian teorik fonksiyonunu burada vermeyeceğiz, Scipy `scipy.stats.invgauss` hesaplar için kullanılabilir. Fakat y_i 'nin kendisi nereden geliyor? Eğer y_i , CDF'in bir sonucu ise, pur veriye bakarak bir CDF değeri de hesaplayabilmemiz gerekir. Bunu yapmak için bir başka numara lazım.

1. Eldeki sayilari artan sekilde siralayin
2. Her veri noktasina bir derece (rank) atayin (siralama sonrasi hangi seviyede oldugu yeterli, 1'den baslayarak).
3. Ceyrek degeri y_i bu sıra / $n + 1$, n eldeki verinin buyuklugu.

Bu teknik niye isliyor? x 'in CDF'i $x_i < x$ sartina uyan x_i 'lerin orani degil midir? Yani bir siralama soz konusu ve ustteki teknik te bu siralamayi biz elle yapmis olduk, ve bu siralamadan gereken bilgiyi aldik.

[1] Introductory Statistics with R

[2] Introduction to Probability and Statistics Using R

[3] <https://gist.github.com/mblondel/1761714>

[4] Applied Statistics and Probability for Engineers

[5] <http://math.stackexchange.com/questions/243348/sample-variance-conver>