İstatistik ikiye ayrılır:

-*Descriptive(betimsel)*;-mean,-median,-mode,-variance,-standart sapma
-*İnferential(Çıkarımsal)*; -veri topla,-örneklem al, -deneyler yap,- sonuç çıkar

-**Population(Ana Kitle)**; mesela tüm Tr
-**Sample(örneklem);** popülasyondan 1000 kişi mesela
-**Mean;** $x = \sum x/n$ örneklem için,,,,,, $\mu = \sum x/n$ popülasyon için
-**Variance**; $\sigma^2 = \Sigma(Xi - \mu)^2/n$ for populasyon —------ $s^2 = \Sigma(Xi - x)^2/n - 1$ for örneklem;
!! n-1 örneklem seçildiği takdirde varyans olması gerektiğinden daha küçük çıkacaktır.
n-1 düzeltmesiyle bu bias (eğilim, yanlılık) düzeltilebilir.

Variable(Değişkenler)

Nicel (Sayısal Değişkenler)

Bu değişkenler sayılarla ifade edilir ve matematiksel işlemler yapılabilir
a)***Continuous Variable***
*b)**Discrete Variable***

Nitel (Kategorik) Değişkenler

Bu değişkenler sayısal değer taşımaz, gruplandırma veya sınıflandırma için kullanılır.
a)***Adlandırılmış (Nominal) [Sıralama yok] Değişkenler;***
—>Bölümler (mat, müh, tıp …)
b)***Sıralı (Ordinal) Değişkenler***
—> Öğrencinin ders notu → (Çok kötü, Kötü, Orta, İyi, Çok iyi)

Random Variable

Bir deneyin olası sonuçlarını sayısal olarak temsil eden bir değişkendir. Bu değişkenin aldığı değerler rastgele belirlenir ve her bir değerin belirli bir olasılığı vardır.

1-Discrete Random Variable

Sayılabilir ve belirli değerler alabilen değişkenlerdir.
Örnekler:
• Bir sınavda doğru yapılan soru **sayısı** (X = {0, 1, 2, …, 50})
• Bir sınıfta bulunan öğrenci **sayısı** (Y = {20, 21, …, 35})

2-Continuous Random Variable

Bir aralıktaki herhangi bir reel sayıyı alabilen değişkenlerdir.
Örnekler:
- Bir öğrencinin boyu (Z = [140 cm, 200 cm])--->176.43544 cm olabilir
- Bir öğretmenin maaşı (Y = [10.000 TL, 20.000 TL])--->11234.02 tl olabilir

Yüzdelik

Yüzdelik, bir veri setinde belirli bir yüzdenin altında kalan değeri ifade eder. Veriyi 100 eşit parçaya böler. Yüzdelikler, bireysel performansı daha büyük bir grupla karşılaştırmak için kullanılır (örneğin, sınav puanları, büyüme eğrileri)

Percentile =$(k/100) * n$

*mesela numbers=(2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)
25. Yüzdelik: Pozisyon = (25 / 100)×11 = 2.75
2.75 tam sayı olmadığı için 3'e yuvarlıyoruz.
25. yüzdelik'e denk gelen 3. sıradaki numara = **6** (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)

aynı hesaplama 100,75,50 filan için de geçerlidir.

Covariance

Iki değişkenin birlikte nasıl değiştiğini gösteren bir ölçümdür. Pozitif veya negatif olabilir:
•Pozitif Kovaryans → Bir değişken artarken diğeri de artıyorsa (veya ikisi de azalıyorsa).
•Negatif Kovaryans → Bir değişken artarken diğeri azalıyorsa (ters yönlü hareket)

$$Covariance_{sample}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{i=1}^{n}(x_i - \overline{\boldsymbol{X}})(y_i - \overline{\boldsymbol{Y}})}{n-1}$$

$$Covariance_{pop}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{i=1}^{n}(x_i - \mu_{\boldsymbol{X}})(y_i - \mu_{\boldsymbol{Y}})}{n}$$

Correlation

Korelasyon, iki değişken arasındaki ilişkinin **gücünü ve yönünü ölçen** bir değerdir. Kovaryanstan farklı olarak, korelasyon her zaman -1 ile 1 arasında bir değerdir, bu da onu daha anlaşılır ve karşılaştırılabilir yapar
- Pozitif Korelasyon (+1'e yakın) → Bir değişken artarken diğeri de artıyorsa.
- Negatif Korelasyon (-1'e yakın) → Bir değişken artarken diğeri azalıyorsa.
- 0'a yakın Korelasyon → Değişkenler arasında belirgin bir ilişki yoksa

$$\boxed{\rho_{XY} = \frac{Covariance(\boldsymbol{X},\boldsymbol{Y})}{\sigma_X \sigma_Y}}$$

Olasılık Kuralları

1-Mutually Exclusive:

Eğer iki olay aynı anda gerçekleşemezse (örneğin bir zar atıldığında hem 2 hem de 5 gelmez), bu olaylara ***karşılıklı dışlayan*** olaylar denir. Böyle durumlarda toplam olasılık, bireysel olasılıkların toplamına eşittir. Formül: P(A∨B) =P(A)+P(B)

2-Non-Mutually Exclusive

Eğer iki olay aynı anda gerçekleşebiliyorsa, toplam olasılığı hesaplarken kesişim (ortak gerçekleşme) ihtimalini çıkarmamız gerekir. Formül: P(A∨B) =P(A)+P(B)−P(A∧B)

Olasılık Dağılımları

1-Olasılık Kütle Fonksiyonu(PMF)

Kesikli (discrete) rastgele değişkenlerin aldığı belirli bir değerin olasılığını gösteren bir fonksiyondur. Biri birini 0,1,2,3 defa filan arayabilir ama 2,5 defa arayamaz tam sayı olmalıdır.

ayrık bir rastgele değişkenin (random variable) alabileceği her bir değerin olasılığını gösteren bir fonksiyondur. Kısaca, "şu değerin gerçekleşme olasılığı nedir?" sorusuna cevap verir.

$$P(X = x) = f(x)$$

2-Olasılık Yoğunluk Fonksiyonu(PDF)

sürekli (continuous) değişkenlerin belirli bir aralıkta bulunma olasılığını tanımlar. Mesela birinin boyu tam 178.00000 olamaz. 178.435353 gibi bir değer olur.

PDF, belirli bir değerin olasılığını değil, bir aralıktaki (örneğin, 1.70 ile 1.80 metre arası) değerlerin gerçekleşme olasılığını hesaplamak için kullanılır. Pdf'de olasılık, bir fonksiyonun altında kalan alana karşılık gelir.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

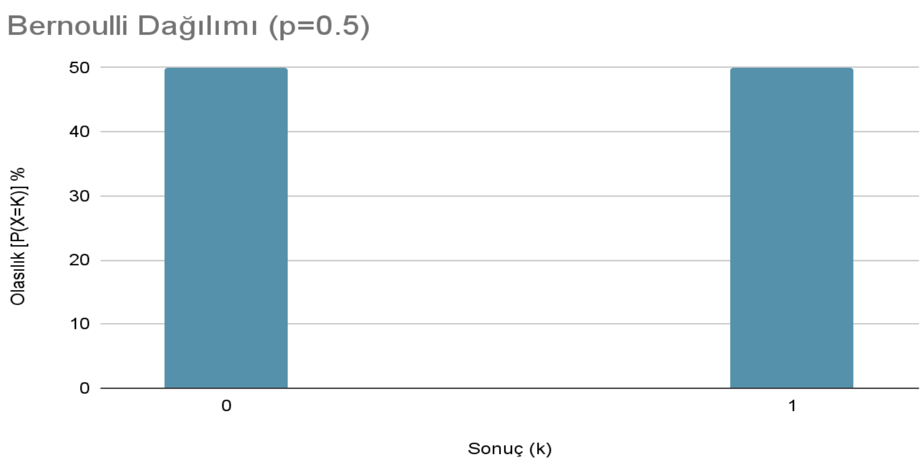
3- Kümülatif Dağılım Fonksiyonu (CDF)

rasgele değişkenin belirli bir değere kadar olan toplam olasılığını gösterir.
bir zar atışında ¸ su olasılıkları hesaplayabiliriz:
- P(X ≤1) = 1/6 - P(X ≤2) = 2/6 - P(X ≤3) = 3/6

Bernoulli Dağılımı

-Başarı(1) veya başarısızlık(0) içeren deneyler için kullanılır.
-Tek bir deneme için geçerlidir(para atmak)
-Başarı olasılığı p ise başarısızlık olasılığı 1-p olarak ifade edilir.

*PMF'i = $P(X = k) = \{p, \quad k = 1\}\{1 - p, \quad k = 0\}$ şeklindedir (k=deney sonucu)



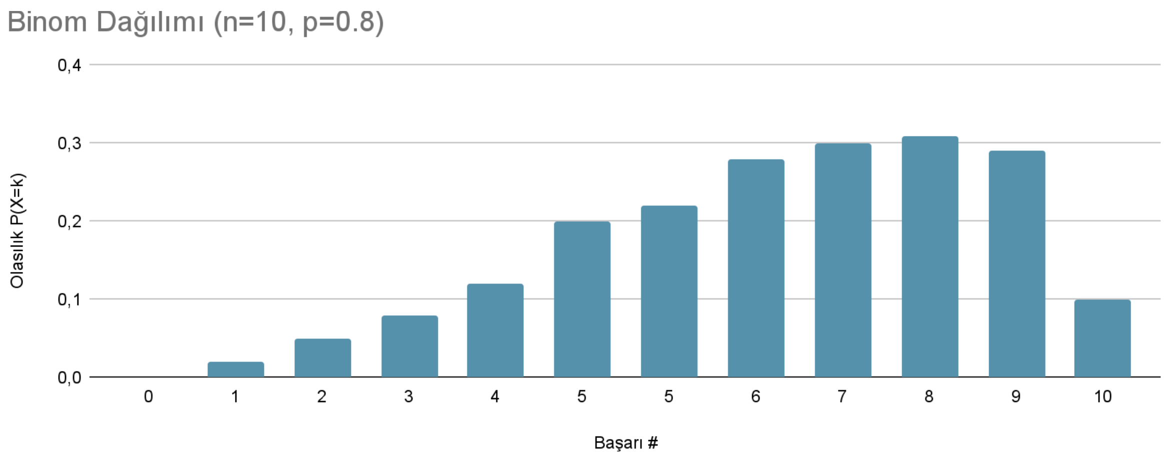
*Başarı olasılığı (p) degistirildiginde dağılım farklılaşır.
*İkili (binary) olayları modellemek için kullanılır.

Binom Dağılımı

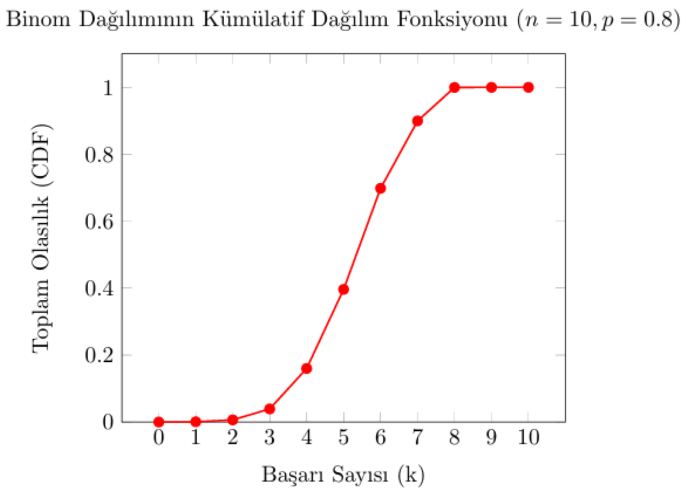
Bağımsız ve aynı olasılıkla gerçekleşen n adet Bernoulli denemesinin başarı sayısını modelleyen bir dağılımdır. Her denemenin başarı olasılığı sabittir. Denemeler bağımsızdır.

*PMF'İ $P(X = k) = C(n,k)p^{k}(1 - p)^{n-k}$ şeklindedir (k=başarı #, n=Deneme #)
*Birbirini takip eden ve birbirinden bağımsız **n** adet Bernoulli denemesinin sonucunda elde edilen toplam "başarı" sayısını modelleyen bir olasılık dağılımıdır

Mesela birinin basket atışında p=0.8 olsun. 10 atış yaparsa 8 tanesini başarılı atma olasılığı;
$P(X = 8) = C(10,8) * (0.8)^{8}(0.2)^{2}$ dir. O da = 0.302.



Bunun için kümülatif dağılım fonksiyonu ise; $F(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$



burdan şunu anlarız;
*P(X ≤6) =0.6983, yani 6 veya daha az başarı elde etme olasılığı yüzde 69.83.
**Bağımsız denemelerde başarı sayısını modellemek için idealdir

Poisson Dağılımı

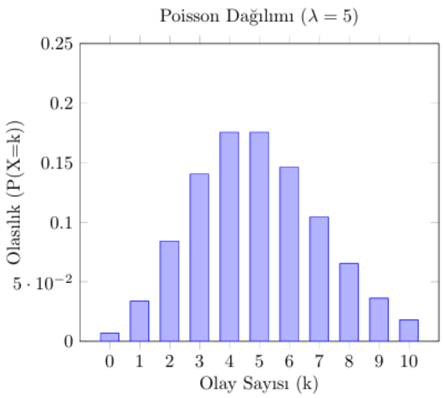
belirli bir zaman veya mekan aralığında meydana gelen nadir olayların olasılığını hesaplamak için kullanılır.
- Olayların bağımsız olması ve belirli bir ortalama hızda gerçekleşmesi gerekir.
- ¨ Örneğin, dakikada gelen müşteri sayısı, bir web sitesine saniyede gelen istekler, telefon santraline gelen çağrı #.

PMF'i $P(X = k) = \gamma^{k} e^{-k}/k!$ (k=gerçekleşen olay #, γ=belli zaman diliminde oluşan olay ort)

mesela,bir çağrı merkezine saatte ort 5 çağrı geliyor.γ=5.3 çağrı gelme olasılığı;
P(x=3)=$5^{3} * e^{-3}/3!$ den %14.04 dür.

•

•



• En yüksek olasılık λ = 5'te gözlemlenir.
Daha düşük ve daha yuksek olay sayıları için olasılıklar azalır.
Poisson dağılımı, olayların **rastgele** ama b**elirli bir ortalama hızda** gerçekleştiği durumları modellemek için idealdir

bunun için kümülatif dağılım fonk ise;

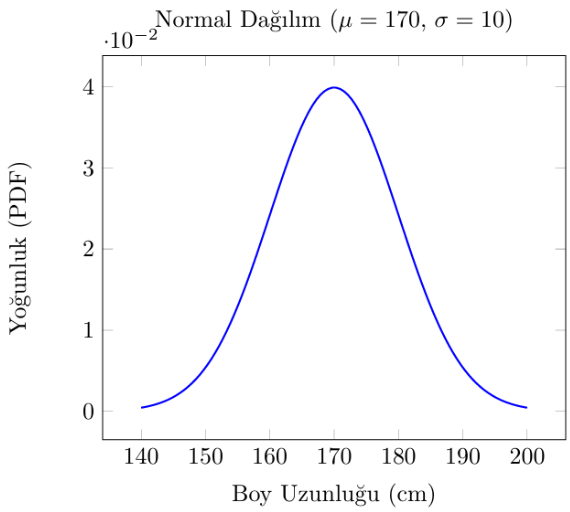
$$F(k) = P(X \leq k) = \sum_{i=0}^{k} \frac{\lambda^i e^{-\lambda}}{i!}$$

***** Ortalama Olay sayısı bilindiğinde, belirli sayıda olayın gerçekleşme olasılığını hesaplamak için kullanılır.***

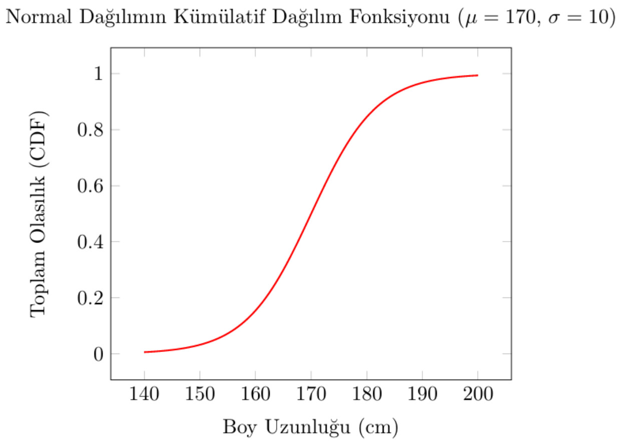
Normal (Gaussian) Dağılım

birçok doğal fenomenin istatistiksel dağılımını tanımlayan en yaygın sürekli olasılık dağılımıdır.Ortalamaya yakın degerler daha olasıdır, uç değerlerin olasılığı düşüktür.Simetriktir.

PDF'i; $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. (μ=dağılım ort, σ=standart sapma,



• Ortalama 170 cm civarında en yüksek yoğunluk vardır.
• Daha uzun veya daha kısa boyluların olasılığı azalır.
• Egri simetrik oldugu icin 160 cm ve 180 cm esit olasılıga sahiptir



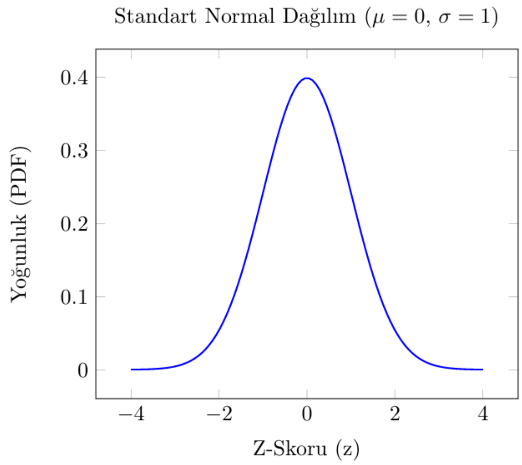
Standart Normal Dağılım

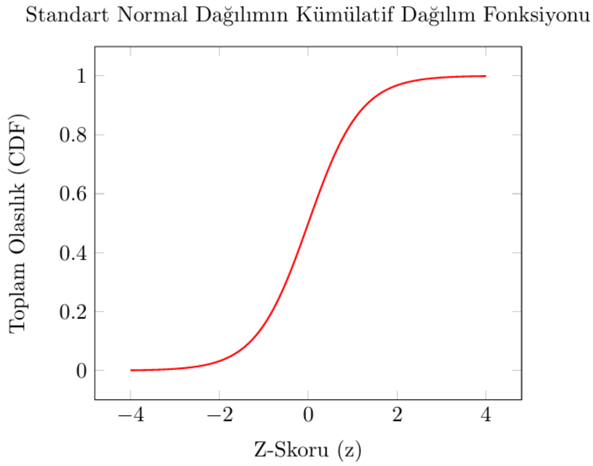
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Burda, ortalama (μ)=0, standart sapma(σ)=1’dir.
Bu dağılım, herhangi bir normal dağılımı ¨**ölçekleyerek analiz etmeyi** kolaylaştırır.

Z-Skoru

Bir veri noktasının ortalamadan kaç standart sapma uzaklıkta olduğunu gösterir.
$Z = (X - \mu)/\sigma$ → X=veri noktası, μ=Dağılım ort.
-Mesela bir sınavın not ort 75, st sapma 10 olsun. Ali 90 almış Zeynep 60 almış.
$Z_{ali} = (90 - 75)/10 --> 1,5$



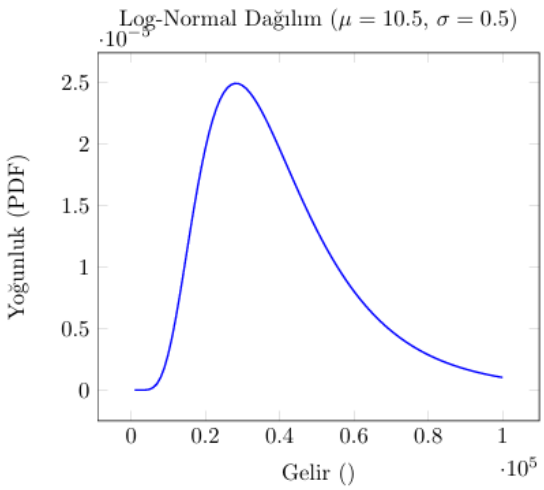


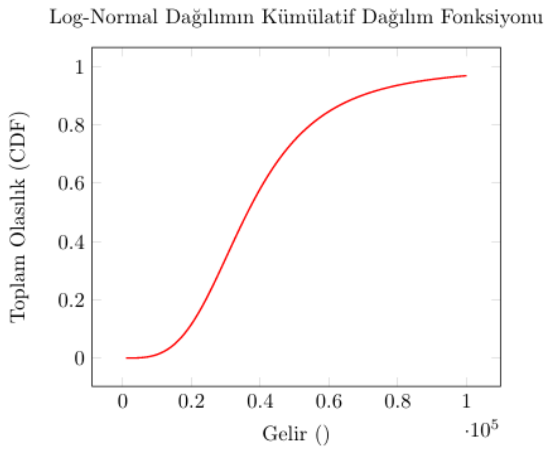
Uniform (düzgün) Dağılım

Belirli bir aralık içindeki tüm değerlerin eşit olasılıkla gerceklestigi bir olasılık dağılımıdır.

Log-Normal Dağılım

bir değişkenin logaritmasının normal dagıldıgı bir sürekli olasılık dağılımıdır.Pozitif buyukluklere sahip verileri modellemek için idealdir (örn: **gelir dağılımı, borsa fiyatları**).
*Cogu ulkede insanların gelir dağılımı log-normal bir dağılıma uyar.





Pareto Dağılımı

bazı sistemlerde az sayıda büyük değer ve çok sayıda küçük değer bulunduğunu gösteren ¸ çarpık bir dağılımdır. Bu dağılım, Guc Kanunu (**Power Law**) ile yakından ilişkilidir. ”***80/20 Kuralı***” olarak bilinen Pareto prensibini temsil eder.
Servet dağılımı, sosyal medya takipçi sayıları¸ şehir nüfusları gibi.

HİPOTEZ TESTLERİ

Güven Aralığı

$$\text{Güven Aralığı} = \bar{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

bir populasyon parametresinin belirli bir guven duzeyinde hangi aralıkta olabileceğini gösterir. (X=örneklem ort., Z∗ → Seçilen güven düzeyine karşılık gelen kritik Z değeri, n → ̈ Örneklem büyüklüğü)

Hata Payı

güven aralığının yarı genişliğidir. $ME = Z^* \frac{\sigma}{\sqrt{n}}$ şeklinde hesaplanır.

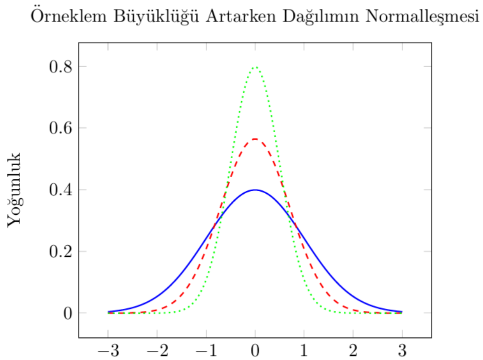
Ex: Bir adayın %52 oy alacağı tahmin ediliyor. Bu tahminin %95 güvenle hangi aralıkta olur.
$\hat{p} = 52,\ n = 1000,\ güven\ düzeyi\ \%95\ yani\ Z\ ^* = 1.96$
-Bir rasgele değişken binom dağılımına sahipse, ***varyansı*** ¸ şu şekilde hesaplanır:
$\alpha^2 = p(1 - p)$ →p=0.52…,örneklem ***standart hata***sı ise $SE = \sqrt{p(1 - p)/n}$ şeklinde hesaplanır

Bu örnekte **SE=0.0158** bulunur. **Hata Payı=Z*xSE'**dir. Hata payı=0.031 bulunur.
$Güven\ Aralğ = \hat{p} \pm Hata\ pay(ME)$ sonuç olarak Güven aralığı (0.489,0.551) yani %95 güvenle %48 ile %55 arasındadır.

Central Limit Theorem

buyuk örneklemelerle çalışıldığında, bağımsız rasgele değişkenlerin toplamı veya ortalamasının **yaklaşık olarak normal dağıldığın**ı soyler.
* *Örneklem büyüklüğü arttıkça*, dağılımın ¸ şekli normal dağılıma yaklaşır
• Ilk deney: 10 zar attık, gelen sayıların ortalaması 3.7 • İkinci deney: 10 zar attık, gelen sayıların ortalaması 3.9 • üçüncü deney: 10 zar attık, gelen sayıların ortalaması 3.5

Bunu binlerce kez tekrar ettiğimizi düşünelim. Sonuc olarak elde edilen ortalamaların histogramını çizdigimizde, bu dağılımın normal dağılıma benzedigini görürüz.



mavi çizgi n=10 için, kırmızı n=30, yeşil ise n=100 için.

Hipotez Testi, bir örneklemden elde edilen verilerin belirli bir iddiayı destekleyip desteklenmediğini belirlemek için kullanılan istatistiksel bir yöntemdir.
H_0=Null Hipotezi → "değişiklik yoktur","etki yoktur" önermesidir.
H_A $veya$ H_1=Alternatif Hp→"farklılık vardır"-"etki vardır" önermesidir.
***Adım adım şöyle hesaplanır:*
1-H_0 ve H_1 belirlenir.
2-Anlamlılık seviyesi seçilir ($\alpha = 1 - güven\ düzeyi$) $genellikle$ %95 $için\ mesela$ 0.05 $olur$
3-Test istatistiği seç: Z-testi, t-testi, chi_sq-testi
4-p değerini hesapla: H_0 $n\ doğru\ olma\ olaslğ$
5- sonuç α'dan küçükse (p< α) H_0 reddedilir. yani H_1desteklenir.

Z-Testi

populasyon standart sapması (σ) bilindiğinde ve örneklem büyüklüğü büyük olduğunda (n > 30) kullanılan bir hipotez testidir.

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

X=örneklem ort μ=popülasyon ort σ →Populasyon st sapma
n → Örneklem büyüklüğü

1-Tek Kuyruklu (One-Tailed) Testler

Hipotez, belirli bir yönde değişiklik olup olmadığını test eder.*¨ Örnek: "Yeni makine eskiye göre daha hızlıdır" hipotezi.

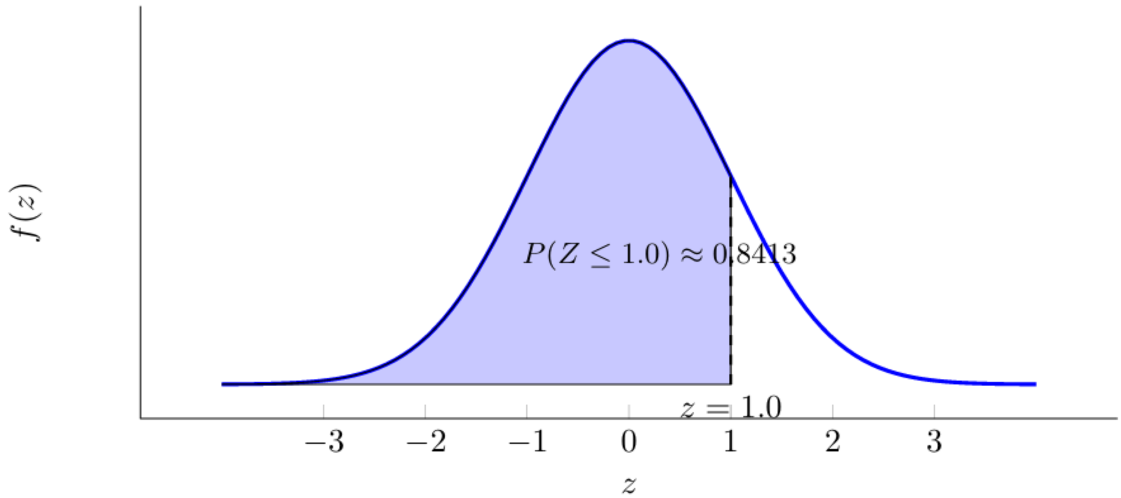
2-Çift Kuyruklu (Two-Tailed) Testler

Hipotez, herhangi bir yönde anlamlı bir ***fark*** olup olmadığını test eder. Örnek: "Yeni makine, eskisine göre farklıdır" hipotezi.

Güven Aralığı ile p Değeri Arasındaki ilişki

anlamlılık düzeyi (α): Hatalı bir şekilde H_0 hipotezini reddetme olasılığıdır.
Güven düzeyi: 1 − α olarak tanımlanır. %95 güven düzeyi ⇒ α = 0.05
*Eğer H_0 hipotezindeki değer, güven aralığı içinde ise → p-değeri > α ise H_0 **reddedilemez.**
Aksi halde ise **reddedilir.**

6 Z Tablosu Alan Hesabı Görseli



Örneğin1; eski makinelerin işlem süresi ort 60 sn. yenilerin ortalamasının **daha kısa** olup olmadığını test edelim.
 $\mu = 60$, Yeni makinelerin ortalama işlem süresi: X=57, $\sigma = 5$, $n = 40$ → Z=-3.80 bulunur.
Tablodan **p =P(Z ≤−3.80) =0.00007** → p < 0.05 olduğundan H0 reddedilir
Örnek2; ¨ogrencilerinin sınav puanlarının önceki yıllardaki ortalamadan farklı olup olmadığını test etmek istiyor.
 $\mu = 70$, Yeni öğrencilerin ortalama puanı: X=74, $\sigma = 8$, n=50 → Z=3.54 bulunur
Tablodan **p =2×P(Z ≥3.54) =2×0.0002 = 0.0004** → p < 0.05 olduğundan H0 reddedilir.

10 Z-Testinin Görselleştirilmesi

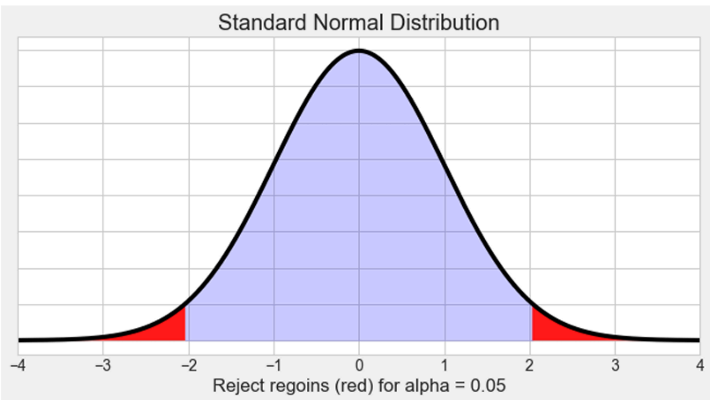


Figure 1: Wikipedia Z Test Visual

Sonuç

- büyük örneklemler (n > 30) için kullanılır.
-Populasyon **standart sapması biliniyorsa** tercih edilir.

T-İstatistiği

*populasyonun standart sapması **bilinmediğinde*** ve örneklem büyüklüğü küçük (n < 30) olduğunda kullanılan bir test istatistiğidir.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

X→örneklem ort , μ →Populasyon ort , s → ¨Örneklem standart sapması , n → Örneklem büyüklüğü

Tek kuyrukluluk ve Çift kuyrukluluk olayı aynı mantıktır.

Ex1:Tek örneklem t-testi

Gene not ort için bakalım. Önceki yılların ort(μ)=70, yeni ort(X)=74, s=8, n=15 olsun.
t=1.94 çıkar.
Serbestlik derecesi: Mesela 5 kişilik bir sınıfın yaş ortalaması biliniyor. Biz 1. ye 2. ye 3. ye ve 4. ye istediğimiz değeri verebiliriz ama sıra 5. ye geldiğinde ortalamaya uymamız gerektiği için buna uygun onu denkleyecek şekilde vermeliyiz. Yani bizim serbestliğimiz sonuncu elemana kadardır. O yüzden **df=n-1** şeklindedir.

örneğimiz için df=15-1=14. **p =P(t ≥1.94) = 0.036** .

Ex2:Bağımsız İki Örneklem T-Testi

İki farklı eğitim modeli var. Karşılaştıracağız. Geleneksel yöntemin ortalaması: X1 =72, s1 = 10, n1 = 20 , Yeni yontemin ortalaması: X2 =78, s2 = 9, n2 = 18

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t = -1.53 , df ≈ min(n1 −1,n2 −1) = min(19,17) = **17** **p=0.14**
p>H_0(0.05) olduğu için reddedilemez.

Ex3: Eşleştirilmiş Örneklem t-testi

Yeni bir diyet programı için önceki diyet ile yeni diyet programı karşılaştırılıyor.
$X_{önce} = 80,\ s_{önce=5}\ X_{sonra} = 76,\ s_{sonra} = 5$, n=12 → t=2.78 bulunur. df=11
p=0.0009 bulunur. H0 reddedilir.

Type-1 Hatası (α-Yanlış pozitif)

Gerçekte doğru olan H_0 hipotezinin hatalı olarak reddedilmesi.Aslında yanlış olmayan bir hipotezin reddedilmesi.
Mesela bir sanığın gerçekten suçsuz olduğunu düşünelim;

-H_0 doğru (sanık suçsuz): mahkeme doğru karar verirse sıkıntı yok.

-H0 yanlış reddedilirse: Suçsuz bir kişi hapse atılır.Yani Type-1 hatası oluşursa!

Type-2 Hatası (β-Yanlış Negatif)

Aslen yanlış olan bir hipotezin reddedilmemesi.
Mesela bir kanser testinde kişinin gerçekten hasta olduğunu düşünelim.
H_0 yanlış (hasta var): Test doğru çalışırsa pozitif sonuç vermeli.
Type-2 hatası oluşursa : Hasta olduğu halde test negatif çıkar.
-Hatalı şekilde H_0 kabul edilir.

Chi-Square Uygunluk Testi

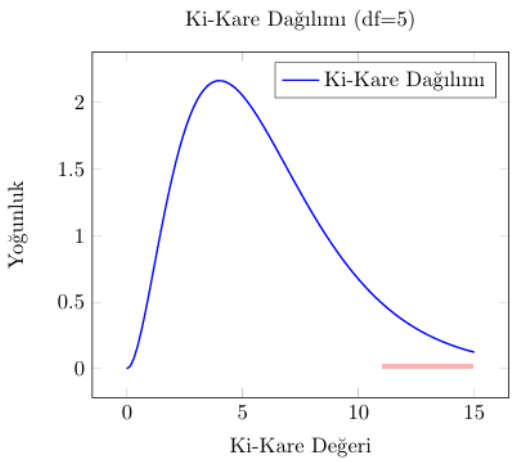
Bir veri setinin beklenen bir dağılıma uyup uymadığını test etmek için kullanılır. *Örneğin bir zarın adil olup olmadığını test etmek.
H_0 : Veriler beklenen dağılıma uygundur.
H1:Veriler beklenen dağılıma uymaz.

$x^2 = \sum(O_i - E_i)^2/E_i$ $\rightarrow O_i$:Gözlenen (observed) değer, E_i: beklenen (expected) değer.

Yüz Değeri	1	2	3	4	5	6
Gözlenen	15	18	22	20	25	20
Beklenen	20	20	20	20	20	20

x^2 hesaplanınca 2.9 bulunur.
df=6-1=5.
Kritik değer $x^2_{0.05} = 11.07$'dir.
$x^2 = 2.9$ olduğu için kritik değerden küçüktür. p>0.05, H_0 reddedilemez.

Zar adildir.



Mavi eğri ki-kare dağılımı gösterir. Kırmızı bölge %5 hata payı içeren kritik bölgeyi temsil eder. Hesaplanan $x^2 = 2.9$ değeri kırmızı bölgeye düşmediği için H0 reddedilemez.

Annova testi

—--pdflerden bak.

Bayes Teoremi

şartlı olasılığı hesaplamak için kullanılan temel bir istatistiksel yöntemdir.

*P(A\B): B olayı gerçekleştiğinde A’nın olma olasılığı.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$