**Veri Bilimi İçin İstatistik**

-> örnek teorisi: ana kitlenin alt kümesi. Genelde örneklemler üzerinden çalışırız. py da örneklem değerlendirme vs görücez

-> betimsel istatistikler: merkezi eğilim ve dağılım ölçüleri olarak değinmiştik. kovaryans ve korelasyon kavramlarını da görücez

-> güven aralığı hesaplama

-> olasılık dağılımları: elimizdeki rastgele değişkenlerin dağılımlarına göre olasılık hesaplama

-> hipotez testleri: ab testlerini? öğrenicez

-> varyans analizi: 2 den fazla grup üzerinde ortalamaya ilişkin test yapma işlemleri

-> korelasyon analizi: çok değişkenli yöntemlerin girişi

* **örneklem dağılımı:** birden fazla örneklemin dağılımı
* **merkezi limit teoremi:** bağımsız ve aynı dağılıma sahip rassal d.lerin toplamı ya da aritmetik ortalaması yaklaşık olarak normal dağılmaktadır.

**Betimsel İstatistikler**

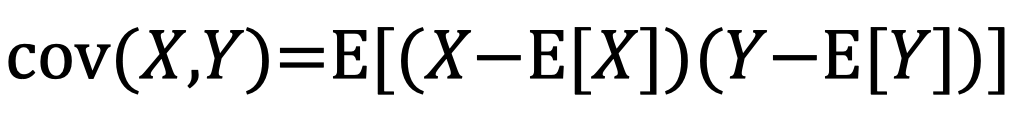
gördüklerimiz

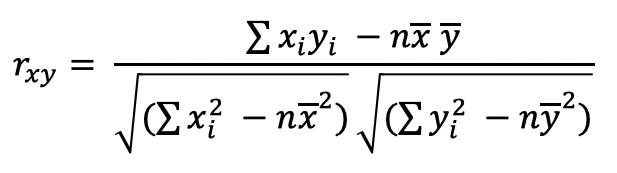
* ortalama
* medyan
* mod
* kartiller
* değişim aralığı
* standart sapma
* değişim aralığı

bu bölüm göreceklerimiz

* kovaryans
* korelasyon

varyans: ortalama etrafındaki dağılımın ölçüsü, standart sapmanın karesi

**Kovaryans:** iki değişken arasındaki ilişkinin değişkenlik ölçüsüdür. İki rastgele değişkenin kendi ortalamalarından olan sapmalarının çarpımlarının beklenen değeridir. Böylece 2 değişkenin birlikte ortaya çıkardığı değişim incelenmiş olur

**Korelasyon:** iki değişken arasındaki ilişkiyi, ilişkinin anlamlı olup olmadığını, ilişkinin şiddetini ve yönünü ifade eden istatistiksel bir terimdir

**Farkları:** korelasyon iki değişken arasındaki ilişki ile ilgileniyor, kovaryans iki d.arasındaki ilişkinin değişkenliği ile ilgileniyor.

**Güven Aralığı:** anakütle parametresinin(örnek istatistiğinin iki sayı tarafından bir aralıkça ifade edilmesi olarak da düşünebiliriz) tahmini değerini kapsayabilecek iki sayıdan oluşan bir aralık bulunmasıdır. Ölçümün hassasiyetinin bir göstergesidir. Yaptığımız tahminlerin ne kadar güvenilir olduğuyla ilgili bir değer bize sunar.

Önceki bölümde ana kitle içerisinden bir örneklem çektik ve örneğin bir ilçenin yaş ortalamasını tahmin ettik. 10k kişi yerine 100 kişi üzerinden bir tahminde bulunduk ve sonucu 39 bulduk. Sağında solunca yüzdeliklerle ilgili bir yorum da yapsak? Gibi bir amaçla ortaya çıkmış bir istatistiksel araçtır güven aralığı.

Web sitesinde geçirilen sürenin güven aralığı nedir?

Ortalama: 180 saniye

Standart sapma: 40 saniye

İstatistiksel olarak %95 güvenilirlikle web sitemizde geçirilen ortalama süre 172 sn ile 188 sn arasındadır.

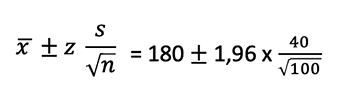
Güven aralığı nasıl hesaplanır?

1. n, ortalama ve standart sapmayı bul.

n=100, ortalama=180, standart sapma=40

1. Güven aralığına karar ver: 95 mi 99 mu? (güven sınırları geniş mi olucak dar mı, genelde default %95)

Z tablo değerini hesapla(1,96 – 2,57) (%95 lik sınırın tablo değeri)

1. Yukarıdaki değerleri kullanarak güven aralığını hesapla: 

(ortalamanın etrafına gözlem sayısını, standart sapmayı ve Z tablo değerini de göz önünde bulundurarak aralık hesapla, ortalamanın etrafında + - bir değer konumlanması yapacak ki 180 ortalamamızdı, + - Z tablo değerimiz 1,96 ydı standart sapmamız 40/kök100 )

1. Sonuç: 180 + - 7,84 yani 172 ile 188 arasıdır.

Sitemize gelen biri %95 güvenilirlikle sitemizde 172 sn ile 188 sn arasında zaman geçirecektir.

**Rassal Değişkenler ve Olasılık Dağılımları**

**Rassal Değişken:** değerlerini bir deneyin sonuçlarından alan değişkenlere rassal değişken denir.

**Olasılık Dağılımı:** bir rassal olaya ait değerler ve bu değerlerin gerçekleşme olasılıklarının bir arada ifade edilmesine olasılık dağılımı denir.

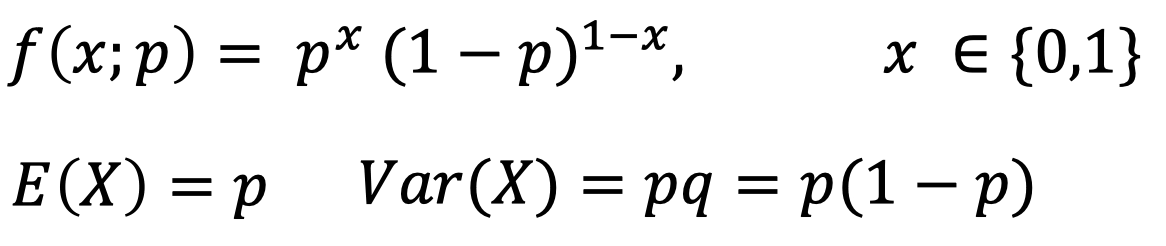
**Olasılık Fonksiyonu:** bir değişkenin herhangi bir değer alması, olasılığını hesaplamaya yarayan fonksiyondur. Olasılık hesaplamaya yarayan fonk.

* Kesikli Olasılık Dağılımları
  + Bernoulli
  + Binom
  + Poisson
* Sürekli Olasılık Dağılımları
  + Normal Dağılım
  + Üniform Dağılım
  + Üssel Dağılım

Kesiklidekileri ve normal dağılımı görücez bu kursta

Veri bilimciler olarak amacımız belirsizlikler altında karar vermek. Bazı tekniklerle belirsizliği azaltmaya çalışıyoruz. Bu belirsizliği azaltmaya çalıştığımız yaklaşımlardan biri de olasılık dağılımları.

**Bernoulli:** Başarılı-başarısız, olumlu-olumsuz şeklindeki iki sonuçlu olaylar ile ilgilenildiğinde kullanılan kesikli olasılık dağılımıdır. (burada hepsini sayılar cinsinden ifade ettiğimiz için kesikli veya sürekli olarak ifade edicez, bernoulli kesikli rassal değişkenlerden biridir, 1 0 başarılı başarısız gibi belirli kategorik d.leri ifade edenler kesikli, 0,6 0,7 falan sürekli)

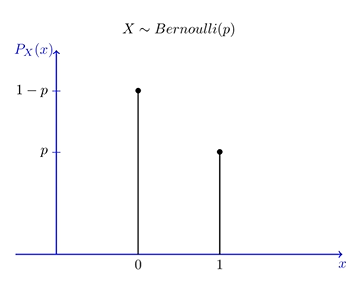


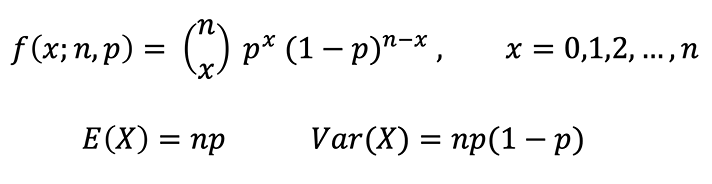
p: olasılık, yazı tura deneyinde yazı gelme olasılığı 0.5 old için p=0.5

x: kesikli değişkenimizin(rassal d.) alacağı değer, bernoulli dağılımı kapsamında 0 yada 1 olabilir, mesela yazı için x=1, tura için x=0

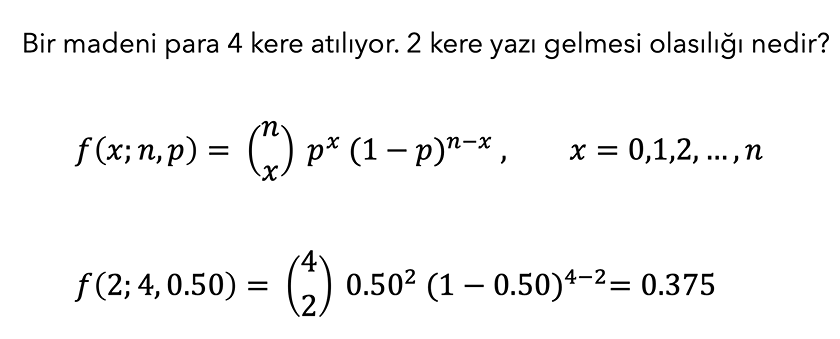
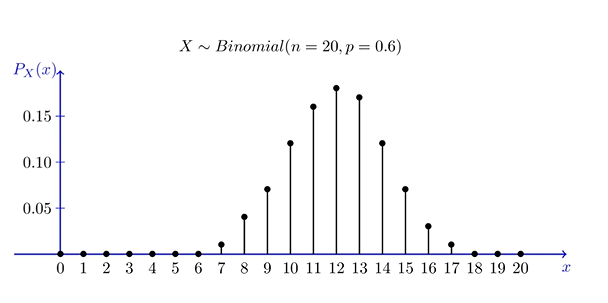
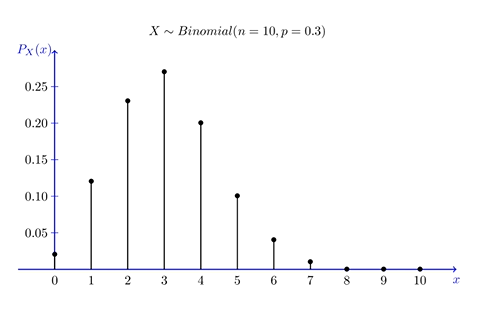
E(x) = p, beklenen değer yani ortalama(yani merkezi eğilime karşılık geliyor)

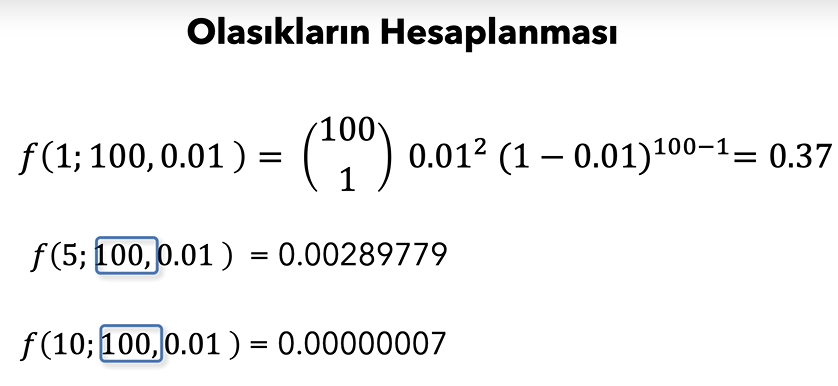
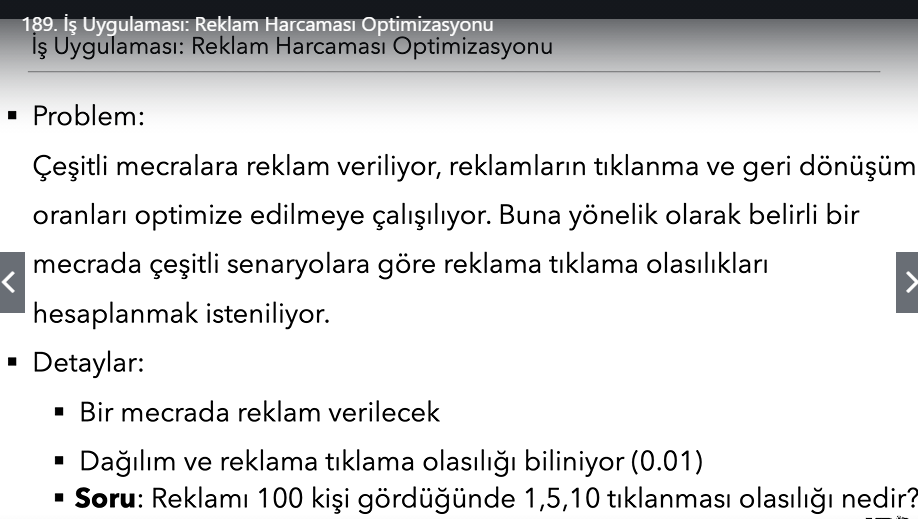
Var(x) varyans(dağılım, yayılım), ilgili olayın gerçekleşmesi olasılığı(p)\*gerçekleşmemesi olasılığı(1-p)

E(x) ve Var(x) varsa bunlar dağılımın merkezini ve o merkez etrafındaki dağılımını ifade ediyor olacak

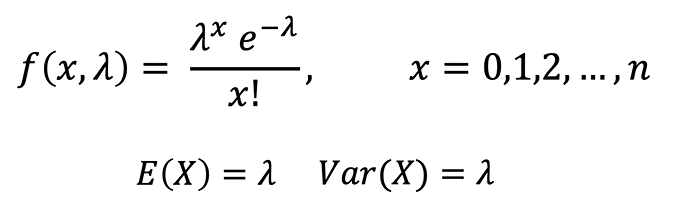
**Binom Dağılımı:** Binom dağılımı, bağımsız n deneme sonucu k başarılı olma olasılığı ile ilgilenildiğinde kullanılan dağılımdır. Bernoulli nin n defa gerçekleştirilmiş versiyonu

n: deneme sayısı





**Poisson Dağılımı:** Belirli bir zaman aralığında belirli bir alanda nadiren rastlanan olayların olasılığını hesaplamak için kullanılır. Gözlem sayısı çok gerçekleşme olasılığı az durumlarında.



Lambda: daha önce bulduğumuz veya çalışmadan bulacağımız değerler olacak

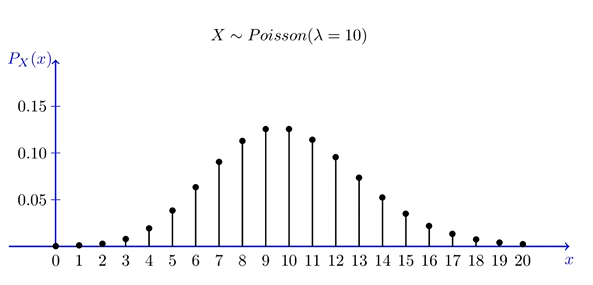
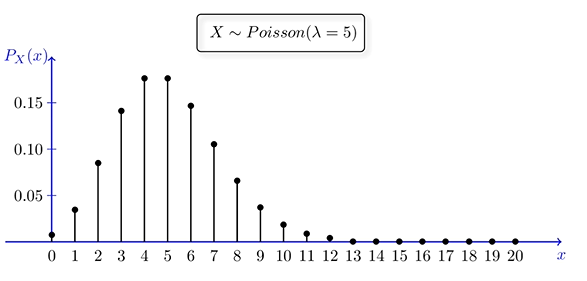
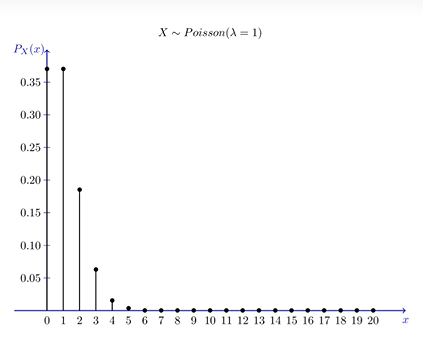
Poisson dağılımında ortalama ve varyans 1-1ine eşittir

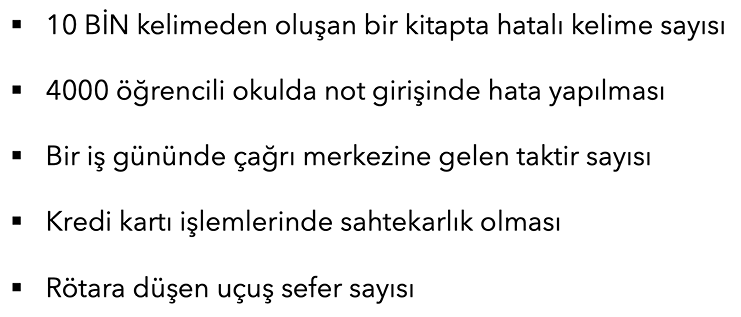
Lambda: beklenen sonucun ortalama gerçekleşme sayısı.

e: euler

x: olayın ortaya çıkma sayısı

lambda arttıkça merkezin etrafında farklı değerler gözlenmesi söz konusu olur, grafik yayvanlaşır. Çünkü lambda hep ortalamayı hem varyansı gösterir.





Gözlem sayısı çok gerçekleşme olasılığı az durumlarında.

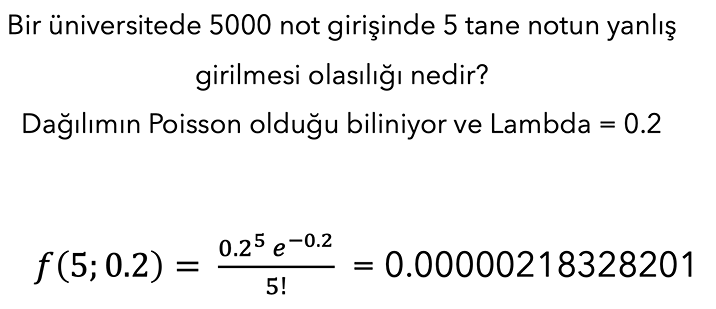
n: büyük

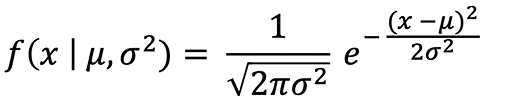
p: küçük

nadir: n > 50 , n\*p < 5

- Rassal denemeler iki sonuçlu olmalıdır.

- aynı koşullar altında gerçekleşmelidir.

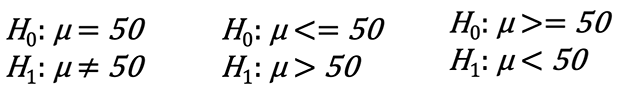
- Rassal denemeler 1-1inden bağımsız olmalıdır.**Normal Dağılım:** Normal dağıldığı bilinen sürekli rassal d.lerde olasılık hesaplamak için kullanılır. ÖNEMLİ! Sürekli d.adı altında yani aralık anlamında bazı olasılık değerleri hesaplıycaksak integral kullanarak alan hesabı yapmamız lazım. Diğerleri kesikliydi değerler ve bu değerlere karşılık gelen olasılıklar vardı. Normal dağılımın beklenen değeri nü



nü: ortalama , x: rassal d.in değeri , sigma kare: varyans , e: euler

**Hipotez Testi Nedir?**

**Hipotez Testi:** Bir inanışı(bir savı, bir tahmini) test etmek için kullanılan istatistiksel terim. Şansa yer vermeyecek şekilde bilimsel dayanaklı ispat. Hipotezler H0 ve H1 hipotezi olarak 2 ye ayrılır.

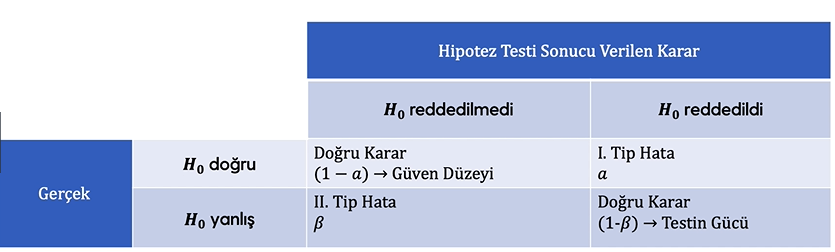


H0: parametrelere belirli değerler girilerek oluşturulan hipotezlere H0 null hipotezi denir.

H1: alternatif hipotez

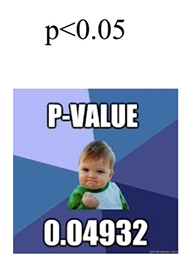
2 ve 3 yönlü hipotezler. Sınayacak olduğumuz, kendisine dayanacak olduğumuz, sabit gibi kabul ettiğimiz H0 hipotezi olacak.

**Hata Tipleri:** Kurduğumuz hipotezleri gerçekleştirdiğimizde ortaya çıkan sonuçlarla ilgili değerlendirmelerdir hatalar.

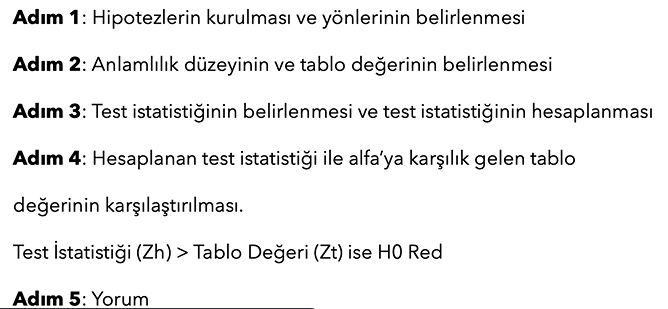
H0 hipotezini kabul etmeyiz. Reddederiz veya reddetmeyiz. diye yorum yaparız !!! istatistikçilerin çok kızdığı bir konu. H0 ı kabul edelim diye bir yorum doğru değildir. Çünkü H0 gerçekte doğru iken , H0 doğru olduğu halde onu reddettiğimde yapacağım hatayı biliyorken(alfa); H0 ı kabul ettiğimde yapacak olduğum hatayı(beta) bilmiyorum.

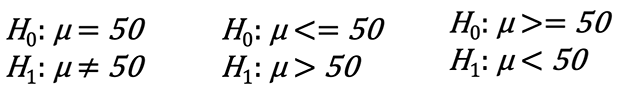
**p-value:** hipotez testlerinin sonuçlarını değerlendirmek üzere programlar tarafından p-value değeri verilir. Bu değer üzerinden kolayca yorum yapabiliriz.

Mesela p-value değerine bakıp 0.05 den küçükse ilgili H0 hipotezi reddettiğimiz sonucuna varıcaz. Eğer p-value değeri alfadan (çalışmamızın başında belirlediğimiz varsayılan kabul edilebilir hata değeri) küçükse H0 hipotezini reddedip seviniyoruz. a(alfa)=0.05 istatistikte kabul gören ön tanımlı değer. p-value 0.05 den her küçük olduğunda sevinemeyebiliyoruz. Mesela dağılıma uygunluk testleri dediğimiz(hipotez:elimizdeki dağılımla teorik dağılımın dağılımı 1-1ine eşittir) dağılım testlerinde H0 reddedilmek istenilmez. Çünkü H0 “örnek dağılımı ile teorik dağılım arasında fark yoktur” der. H0 ı reddetmek istemeyiz, bu yüzden H0 ın 0.05 den küçük olmasına sevinemeyiz.



**Hipotez Testi Adımları:**

****

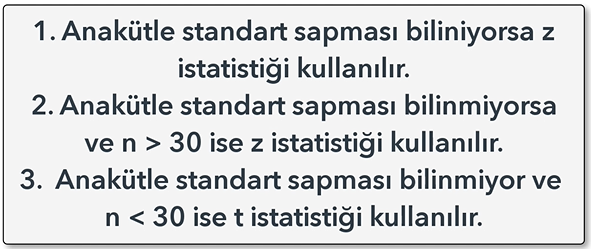
1: 

2: alfa, genelde 0.05 kabul edicez. Bunu güven aralığı bölümünde de ifade ettik. Dolayısıyla alfa anlamlılık düzeyine karşılık gelen tablo değerinde z ile çalışıyorsak z tablosundan, t ile çalışıyorsak t tablosundan, ki kare ile çalışıyorsak ki kareden bakarak bulmuş olucaz.

3: 2.adımdan gelenle karşılaştırmak için farklı bir test istatistiği belirleyip hesaplama

**Tek Örneklem T Testi:** popülasyon ortalaması ile varsayımsal bir değer arasında istatistiksel olarak anlamlı bir farklılık olmadığını test etmek için kullanılan parametrik bir testtir. Örnek ortalamasına ilişkin test yapmak için kullanılır. Elimizde tek bir örnekleme ilişkin tek bir örneklemin ortalamasına ilişkin test yapma ihtiyacı olduğunda kullanılan testtir.

10k kişilik ilçeden 100 örnek çektik. Bu çekmiş olduğumuz örneğe tek örnek denir. Eğer bu örnek üzerinden ana kitleye ilişkin bir değerlendirmede bulunmak istiyorsak(bu da popülasyon ortalamasına karşılık gelir), ana kitle ortalamasına ilişkin teorik bir değer ile bir test yapmak istersek buna tek örneklem t testi yapmış oluyoruz. Kod kısmındaki 2 li hipotezlerin ilk sırasındakiler dayanak noktası olarak düşünülmeli. Hipotez değerlendirme, ilgili hipotez testinin test istatistiğini değerlendirme sonrasında varsayımlar ve iş uygulaması yapmış olucaz. Test istatistiğine karar vermek için yapılacak olan seçim aşağıda

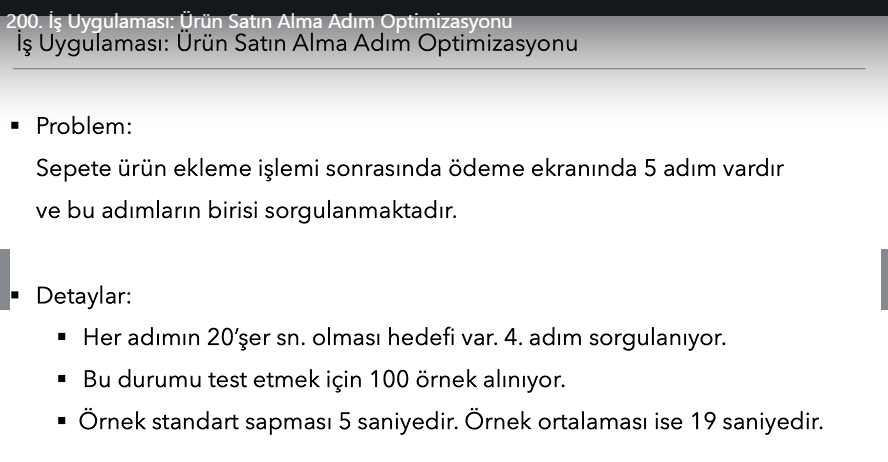
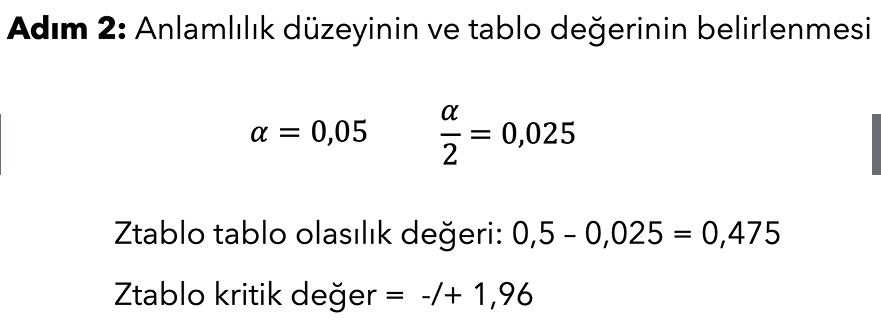
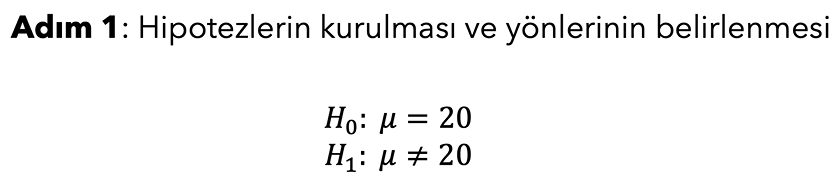


Örnek sayısı arttıkça t dağılımı normal dağılıma yaklaşır. N büyüdükçe t, z’ye yaklaşır. Pratikte R ve Python programlarının içerisinde zaten kullanılan test istatistiği de genelde t oluyor.

Varsayımları test ederiz. Varyasım sağlanıyorsa ilgili hipotez testini kullanıcaz, varsayım sağlanmıyorsa(uygulama kısmında değinecek) num parametrik hipotezlerini değerlendirmiş olucaz.

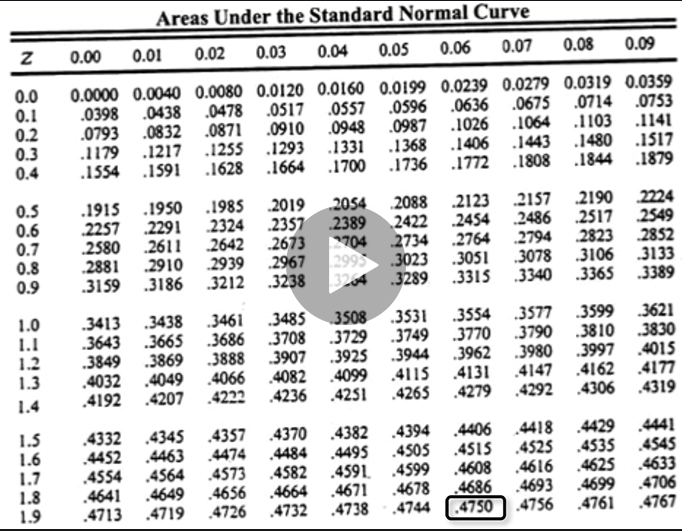
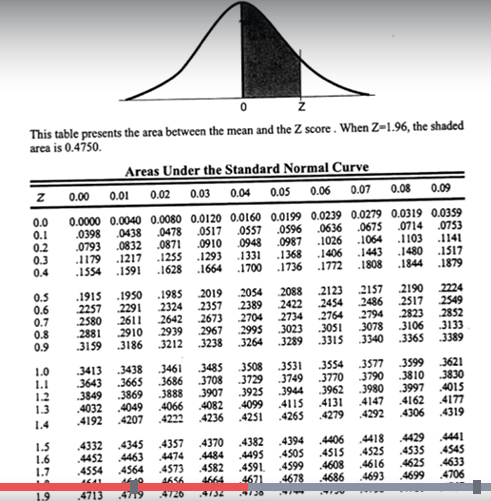
Bir ilçenin yaş ortalamasına ilişkin bir hipotez testi kurmak istediğimizde, web sitemizde insanların geçirdiği süreye ilişkin bir hipotez testi kurmak istediğimizde(180 saniyedir , veya değildir, 180 saniyeden azdır) . elimizde tek bir örneklem tek bir grup tek bir değişken var, tek örneklem t testi kullanılır.

**UYGULAMA: Tek Örneklem T Testi**

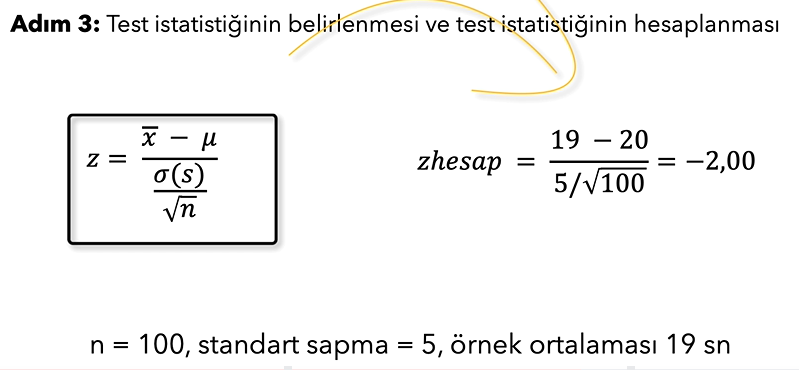
****

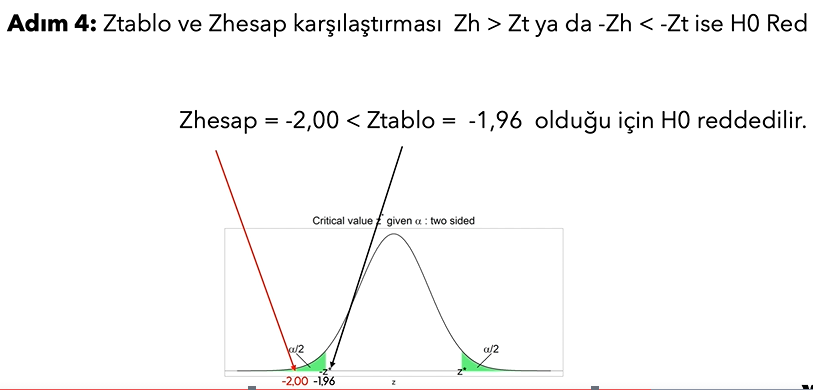
Kurduğumuz hipotez testi 2 yönlü old.için alfa/2 yaptık. > veya < gibi alternatif bir hipotezimiz olsaydı alfayı olduğu gibi kabul edip alfaya karşılık gelen tablo değerine bakacaktık. N>30 old.için z dağılımını kullandık.

Kabul edilebilir hata miktarı olarak belirlediğimiz alfayı 2 ye böldükten sonra elde ettiğimiz değer, bu tabloda eğrinin altında kalan alanları ifade eden bu değerler. Tabloda 0.475 i bulursan z(kritik değeri) sinin 1.96 olduğunu görürsün. Burada kabul edilebilir hata miktarı olasılık değeri üzerinden karşılaştırmak üzere kendimize kritik değer hesapladık, standartlaştırılmış bir tablo üzerinden. Z tablo değerini bulmanın yolu bu tablo. Görülen z değerine gitme adına 0,5-0,075 yaptık. Aşağıdaki şekilde z nin sağı 0,075. 0 ın solu 0.5 sağı 0.5. karalanmış yerin alanını bulduk.



Programlar p tablosunu kullanıyor, p de z ye çok yakın değerler.

p hesap değeri geliyor burdan

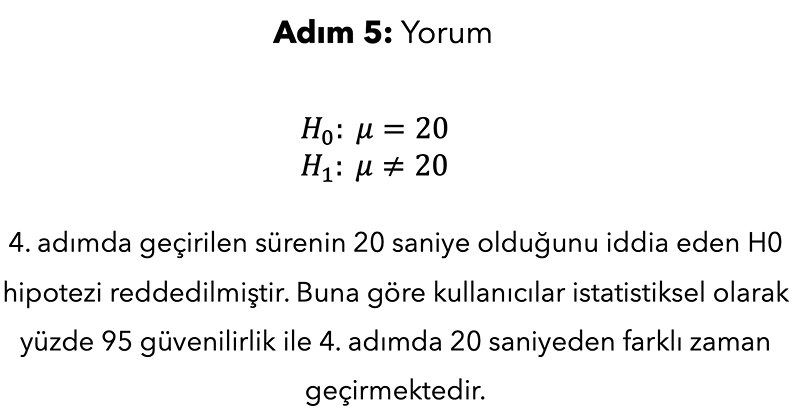


Yeşil yerler: karar noktaları

Zhesap değeri, H0 hipotezinin red bölgesi olan yeşil bölgeye düşüyor. Bu sebeple H0 hipotezi reddedilir.

Elimizde pozitif bir Zhesap değeri olsaydı bu durumda grafiğin sağ tarafındaki yeşil alan ile karşılaştırıp +1.96 ile karşılaştırmış olacaktık.

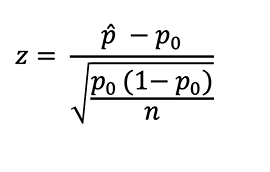
Zhesap hipotez red bölgelerine(yeşil alanlara) düşmemiş olsaydı H0 hipotezini reddetmek için yeterli kanıt bulunamadı diyecektik. İstatistiki olarak anlamlı bir sonuç elde edilemedi diyecektik.



Son cümlede H1 e atıfta bulunarak yorum yaptık. Çünkü H0 ı reddettik.

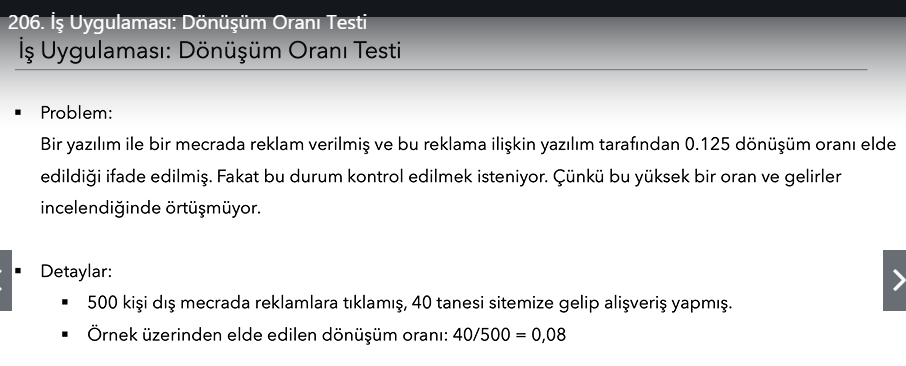
Kolay olması açısından bu örnekte normallik varsayımını incelemedik. Bunun el yordamıyla teorik olarak incelenmesi biraz daha zor. Uygulamalarımızda mutlaka!! Hipotez testinin sonuna gelmeden önce elimizdeki tek örneğin normallik varsayımını sağlayıp sağlamadığı test edilmelidir. Normallik varsayımı sağlanmıyorsa none3+ parametrik peri testi yapılmalıdır.

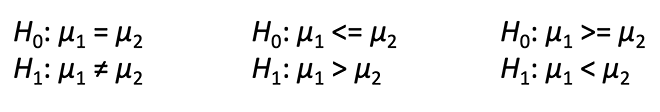
**Normallik varsayımı sağlamaya youtube dan bak**

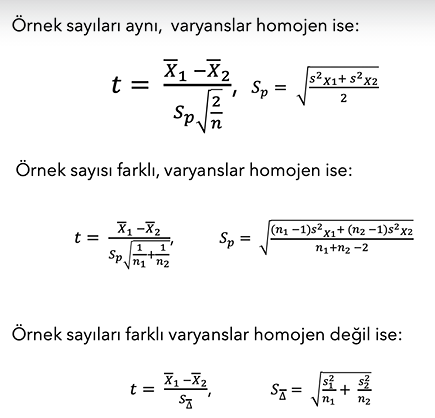
**Tek Örneklem Oran Testi:** Oransal bir ifade test edilmek istendiğinde kullanılır.

Pşapka: örnek üzerinden elde ettiğimiz değer , p0: sınamak üzere odaklandığımız değer

Varsayımımız sadece 1 tanedir: n>30 , toplamış olduğumuz örneklem 30 dan büyük ise bu testi gerçekleştirebiliyoruz.

**Dönüşüm Oranı:** bir reklamı 100 kişi gördü 1 kişi tıkladı, dönüşüm oranı 1/100 dür. Neye odaklanıyorsak ona göre geçişi ifade etmektedir.

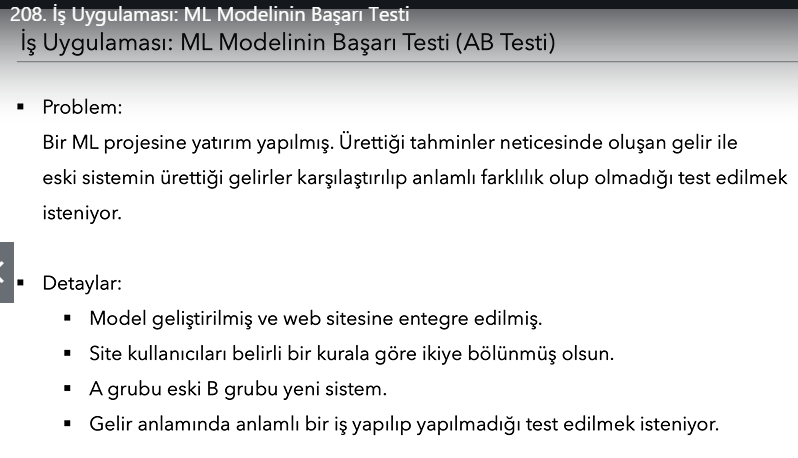
**Bağımsız İki Örneklem T Testi (AB Testi):** iki grup ortalaması arasında karşılaştırma yapılması istenildiğinde kullanılır. Daha karışık tanım: elimizde gerçek değerlerini bilmediğimiz 2 tane ana kitle parametresi var. Ki bunlar 2 farklı ana kitlenin ortalamaları. Bu ort.ların 1-1inden farkını inceliyoruz. Karşılaştırıyoz.



Önemli! Son 2 durumda da örnek sayıları aynı olursa hesaplamalar yapılabilir.

Not: 3.yöntem welch testi olarak da geçer.

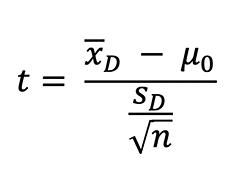
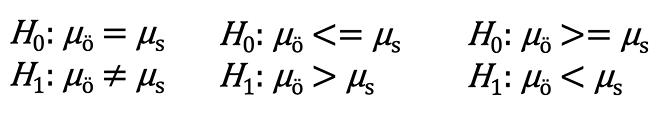
Varsayımlar: Normallik, Varyans Homojenliği(grupların varyanslarının 1-1ine benzer)



8 ay çalışana maaş verdik yatırım yaptık değdi mi?

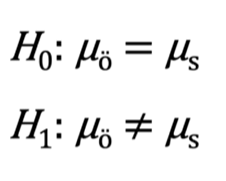
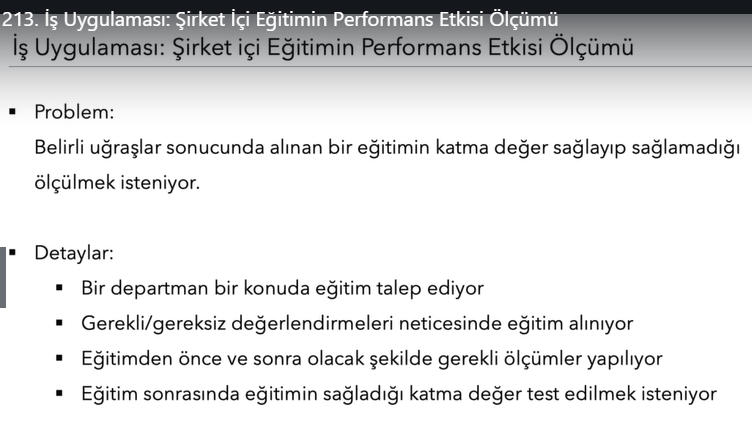
**Bağımlı İki Örneklem T Testi:** Bağımlı iki grup ortalaması arasında karşılaştırma yapılması istenildiğinde kullanılır.

Bağımlı: aynı kitle aynı örneğe 2 farklı uygulama yapıldığında ve bunun sonuçları incelendiğinde bağımlılık durumu söz konusudur. Aynı gruba eğitimin öncesinde ve sonrasında sınav yapılması

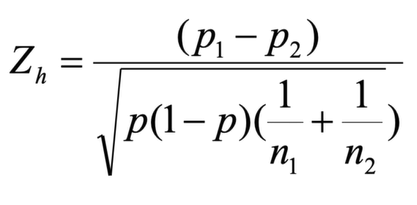
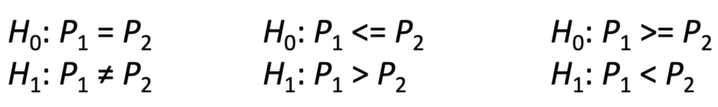
Bağımsız: 2 farklı grup 2 farklı gelir, hepsi 1-1inden bağımsız

Varsayımlar:

* Normallik (yaklaşık olarak normal dağılması da kabul görebilmektedir)
* Varyans Homojenliği

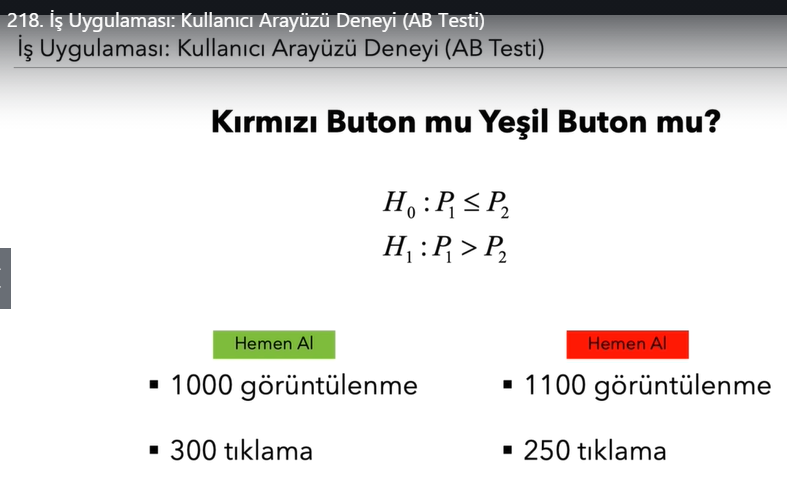


Hipotezimiz: eğitim öncesindeki katılımcıların performanslarıyla sonrasındaki performansları 1-1ine eşittir.

**İki Örneklem Oran Testi:** İki oran arasında karşılaştırma yapmak istenildiğinde kullanılır.p: oran istatistiği sanırım

Varsayımlar:

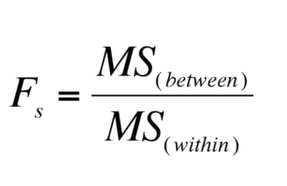
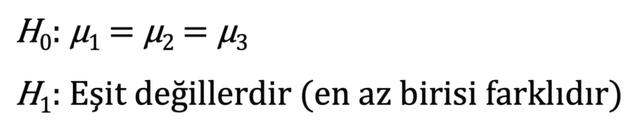
* n1 > 30
* n2 > 30



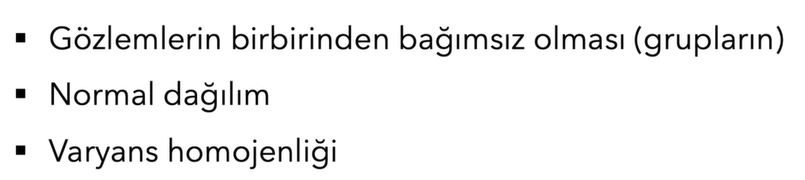
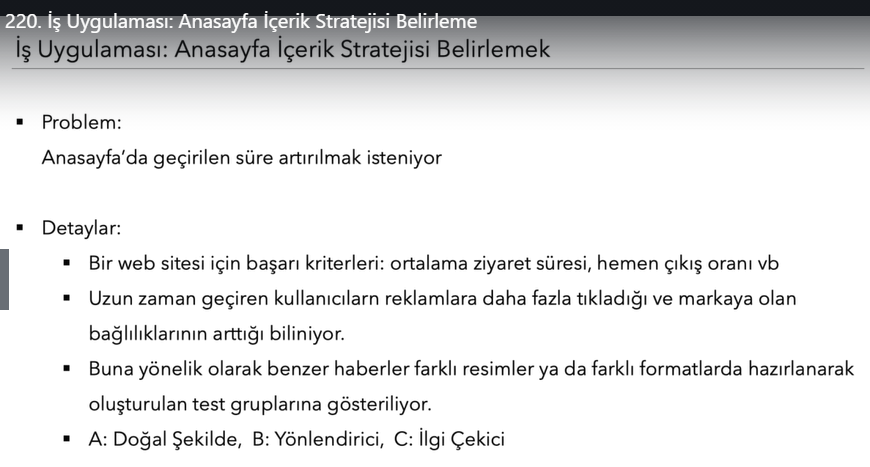
H0: renklerin dönüşüm oranında istatistiki olarak anlamlı bir fark yoktur

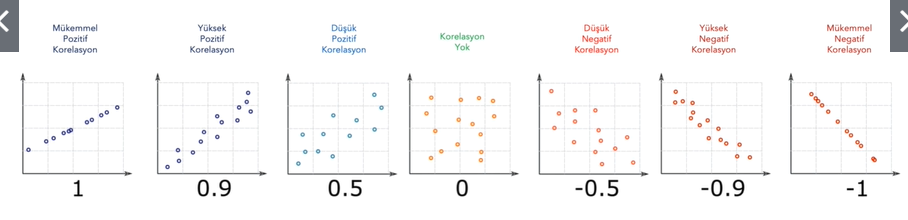
Şirketler bu AB testleri için ayrı departmanlar açıyor neredeyse oyun şirketleri online mağazalar vs

**Varyans Analizi:** 2 veya 2 den fazla grup ortalaması arasında istatistiksel olarak anlamlı bir fark olup olmadığını öğrenmek için kullanılır.

Çok geniş bir konu istatistik bölümlerinde haftalarca okutulu

Varyans analizinde test istatistiğimiz f istatistiği olarak da geçer. Gruplar arası ort.hata (veya değişkenlik de diyebiliriz) / grup içi ort.hata

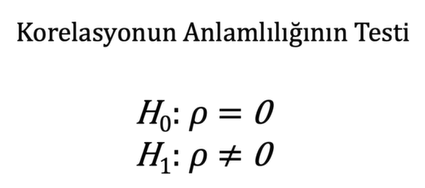
varyans homojenliğini sağlaması önemli

**Korelasyon Analizi:** değişkenler arası ilişki, bu ilişkinin yönü ve şiddeti ile ilgili bilgiler sağlayan istatistiksel bir yöntemdir.

0.9 – 0.5 = düşük pozitif korelasyonla yüksek pozitif korelasyon arasında bir değer alıyor deriz. – lisi de geçerli. 0 a yaklaştıkça ilişkinin şiddeti düşmektedir.

Reklam harcamaları ile ürün satışları arasında pozitif korelasyon var

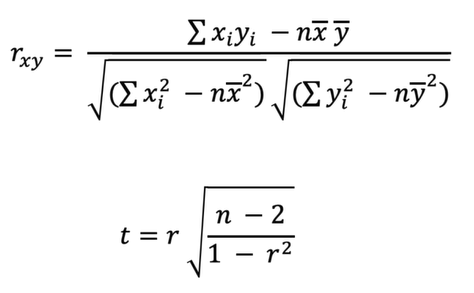
Araç kilometresi ile araç fiyatı arasında negatif korelasyon var

Bundan sonraki derslerde korelasyon katsayısına ilişkin testler de yapıcaz.

Ana kitleye ilişkin korelasyonu göstermek için ro kullanılır.

H0: değişkenler arasında korelasyon yoktur

H1: değişkenler arasında anlamlı bir ilişki vardır

p-value değerinin 0.05 den küçük olup H0 ı reddetmemize sebep olmasını bekleriz

ilk satırda üst taraf soldaki pearson korelasyon katsayısı. r korelasyon katsayısı demek

xi yi ler gözlem birimleri. Çarpılıp toplanıyorlar –(eksi) gözlemSayısı\*(x ort.)\*(y ort.)

bölü kısmı: xi lerin karelerinin toplamı – gözlemSayısı\*(x ortalamanın karesi)

Varsayımlar:

* İki değişken için de normallik varsayımı
* Varsayım sağlanıyorsa Pearson Korelasyon Katsayısı
* Varsayım sağlanmıyorsa Spearman Korelasyon Katsayısı

