Giris

Veri Biliminde ham verilerle farklı deneyler yapmayı ve verilerden bazı iyi içgörüler bulmayı hedefliyoruz. Herhangi bir işi doğru yola yönlendirmek için veri çok önemlidir veya “Veri yakıttır” diyebiliriz. En azından aşağıdakilere yardımcı olabilecek eyleme geçirilebilir bilgiler sağlayabilir:

•Mevcut kampanyaları stratejilendirin,

•Yeni ürünlerin lansmanını kolayca organize edin veya

•Farklı deneyler deneyin.

Yukarıda bahsedilenlerin hepsinde ortak itici bileşen Veridir. Her gün çok fazla Veri ürettiğimiz dijital çağa giriyoruz.

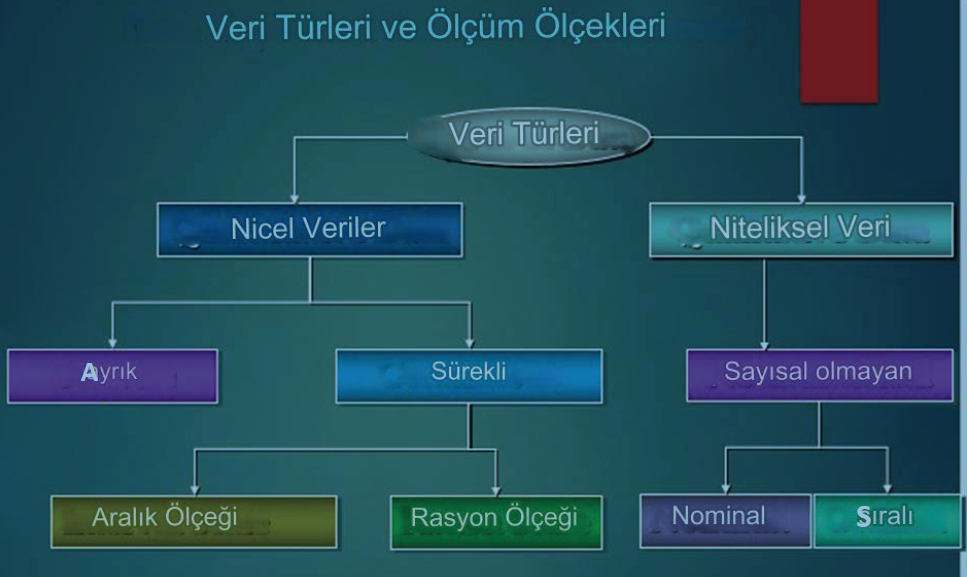
Örneğin, Flipkart gibi bir şirket günlük olarak 2 TB'tan fazla veri üretiyor.

Verinin hayatımızda bu kadar önemli olması nedeniyle, bu verinin doğru şekilde saklanması ve hatasız olarak işlenmesi büyük önem kazanmaktadır. Veri kümeleriyle uğraşırken, veri türü veya verinin kategorisi aşağıdaki soruların cevabını bulmada önemli rol oynar:

• Doğru sonuçları elde etmek amacıyla belirli bir küme için hangi ön işleme stratejisi işe yarar veya

• En iyi sonuçlar için hangi tip istatistiksel analizin uygulanması gerektiği.

Bu nedenle, bu makalede, bir Makine Öğrenimi Projesinin üretim hattındaki en önemli bileşenlerden biri olan Keşifsel Veri Analizini (EDA) uygun şekilde yapmak için bilmeniz gereken istatistiklerdeki farklı veri türlerini tartışacağız.



<https://youtu.be/hZxnzfnt5v8>

İçindekiler

1. İstatistikte Veri Türlerine Giriş ve Önemi

2. Niteliksel ve Niceliksel Veriler

3. Nitel Veriler

•Nominal veri

•Sıra verileri

4. Nicel Veriler

•Ayrık veri

•Sürekli Veri

•Aralık verileri

• Oran Verileri

**İstatistikte Veri Türlerine Giriş**

İstatistikte, verilerle ilgili belirli varsayımları doğru bir şekilde çıkarabilmemiz için istatistiksel ölçümleri verilerinize doğru bir şekilde uygulamak için Veri Türleri çok önemli ve anlaşılması gereken bir rol oynamaktadır.

Benzer şekilde, farklı veri türleri, çeşitli değişken türlerini düzenleme yaklaşımı olarak kabul edildiğinden, doğru algılama tekniğini seçmek için hangi veri analizi ve türü üzerinde çalıştığınızı bilmemiz gerekir.

Genel bir veri bilimi projesinde Keşifsel Veri Analizi (EDA) yaparken, belirli istatistiksel ölçümleri yalnızca belirli veri türleri için kullanabileceğimizden, farklı veri türlerini iyi anlamak çok önemli hale gelir.

Ölçüm Ölçeği olarak da bilinir.

Herhangi bir veri türüyle uğraşırken, hangi görselleştirme yönteminin belirli veri türüne uygun olduğunu da bilmemiz gerekir.

Veri türlerini farklı değişken türlerini kategorize etmenin bir yolu olarak düşünebiliriz.

**Nicel ve Nitel Veriler**

**Nicel Veriler**

1. Bu tür veriler açıklanması en kolay veriler gibi görünüyor. gibi soruların yanıtlarını bulmaya çalışır.

•"kaç tane,

•“ne kadar” ve

•"ne sıklıkta"

2. Sayı olarak ifade edilebildiği için ölçülebilir. Basit bir ifadeyle sayısal değişkenlerle ölçülebilir.

3. Bunlar istatistiksel manipülasyona kolayca açıktır ve çizgi grafikler, çubuk grafikler, dağılım grafikleri vb. gibi çok çeşitli istatistiksel grafik ve çizelgelerle temsil edilebilir.

Niceliksel verilere örnekler:

•Test ve sınav puanları, ör. 74, 67, 98 vb.

• Bir kişinin ağırlığı.

• Bir odadaki sıcaklık.

2 genel niceliksel veri türü vardır:

•**Ayrık veri**

•**Sürekli veri**

**Niteliksel Veri**

1. Niteliksel veri sayı olarak ifade edilemediği için ölçülemez. Esas olarak kelimelerden, resimlerden ve sembollerden oluşur, ancak sayılardan oluşmaz.

2. Bilgiler sayıya göre değil kategoriye göre sıralanabildiğinden Kategorik Veri olarak da bilinir.

3. Bunlar aşağıdaki gibi sorulara cevap verebilir:

•“bu nasıl oldu” veya

•“bu neden oldu”.

Niteliksel verilere örnekler:

•Renkler ör. denizin rengi

•İsviçre, Yeni Zelanda, Güney Afrika vb. gibi popüler tatil yerleri.

• Amerikan Kızılderili, Asyalı vb. gibi etnik köken.

Genel olarak 2 tür niteliksel veri vardır:

**•Nominal veri**

**•Sıra verileri**

**Niteliksel Veri**

**Nominal veri**

1. Bu veri türü herhangi bir niceliksel değere sahip olmaksızın yalnızca değişkenleri etiketlemek için kullanılır. Burada 'nominal' terimi Latince 'isim' anlamına gelen 'nomen' kelimesinden gelmektedir.

2. Herhangi bir özel sıraya başvurmadan sadece bir şeyin adını verir. Nominal verilere bazen “**etiketler**” adı verilir.

Nominal Veri Örnekleri:

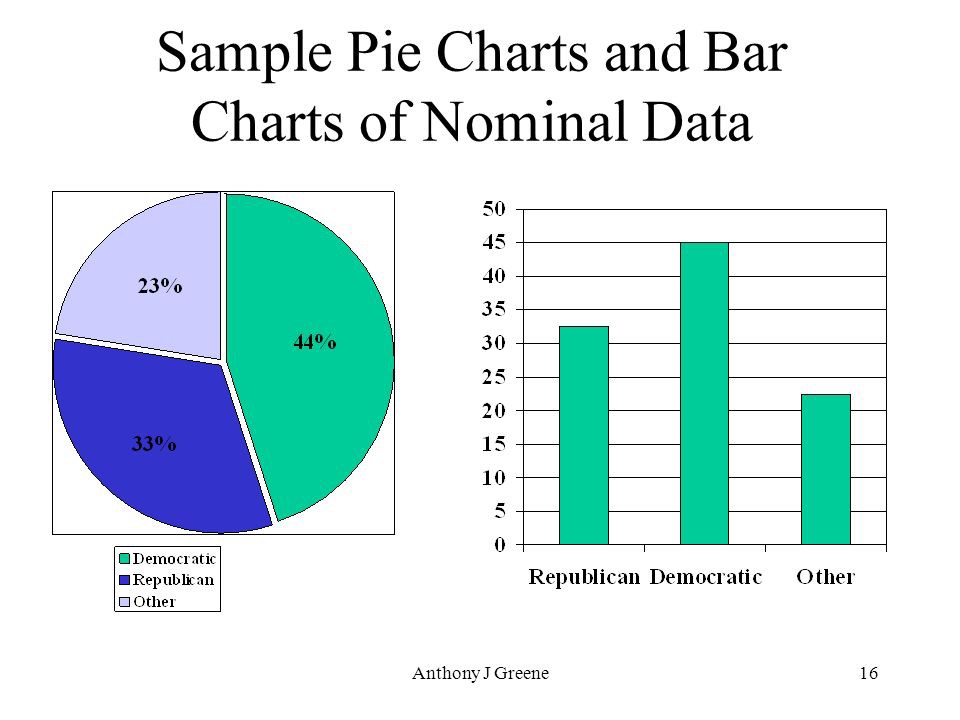
•Cinsiyet (Kadın, Erkek)

•Saç rengi (Sarışın, Kahverengi, Esmer, Kızıl vb.)

• Medeni durum (Evli, Bekar, Dul)

Örneklerden de görebileceğiniz gibi değişkenlerin içsel bir sıralaması yoktur.

Göz rengi, Mavi, Yeşil, Kahverengi vb. gibi birkaç seviyeye veya kategoriye sahip nominal bir değişkendir ve bu kategorileri sıralamaya göre, yani en yüksekten en düşüğe veya tam tersi şekilde sıralamanın mümkün bir yolu yoktur.



**Sıra verileri**

1. Nominal veri türlerinden en önemli fark, Sıralı Verilerin bir sayının nerede bulunduğunu belirli bir sırayla göstermesidir.

2. Bu tür veriler bir ölçekteki konumlarına göre bir tür sıralamaya tabi tutulur. Sıralı veriler üstünlüğü gösterebilir.

3. Sıralı verilerle aritmetik işlemler yapamayız çünkü bunlar yalnızca diziyi gösterir.

4. Sıralı değişkenler niteliksel ve niceliksel değişkenlerin “arasında” olarak kabul edilir.

5. Basit bir ifadeyle sıralı verileri, değerlerin sıralandığı niteliksel veriler olarak anlayabiliriz.

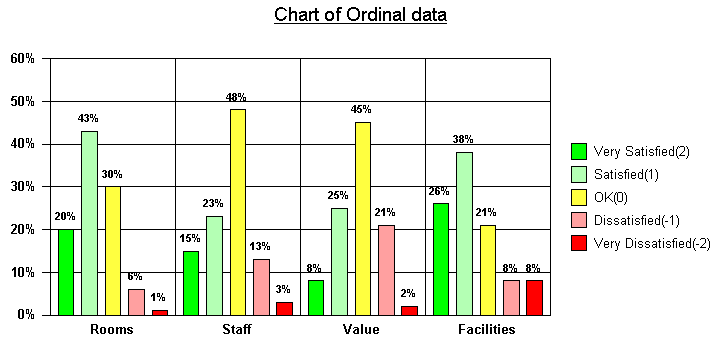
6. Nominal verilerle karşılaştırıldığında ikincisi, değerlerin bir sıraya konulamadığı niteliksel verilerdir.

7. Göreceli konuma bağlı olarak sıralı verilere sayılar da atayabiliriz. Ama bu sayılarla matematik yapamayız. Örneğin, “birinci, ikinci, üçüncü… vb.”

Sıralı Veri Örnekleri:

•Bir yarışmadaki kullanıcıların sıralaması: Birinci, ikinci ve üçüncü vb.

•Firmanın bir ürüne 1-10 arası puan vermesi.

•Ekonomik durum: düşük, orta ve yüksek.

**Nicel Veriler**

**Ayrık veri**

1. Yalnızca tamsayıları içeren sayımı gösterir ve ayrık değerleri parçalara ayıramayız.

Örneğin, bir sınıftaki öğrenci sayısı ayrık verilere bir örnektir, çünkü bireyleri tam olarak sayabiliriz ancak 2,5, 3,75, çocukları sayamayız.

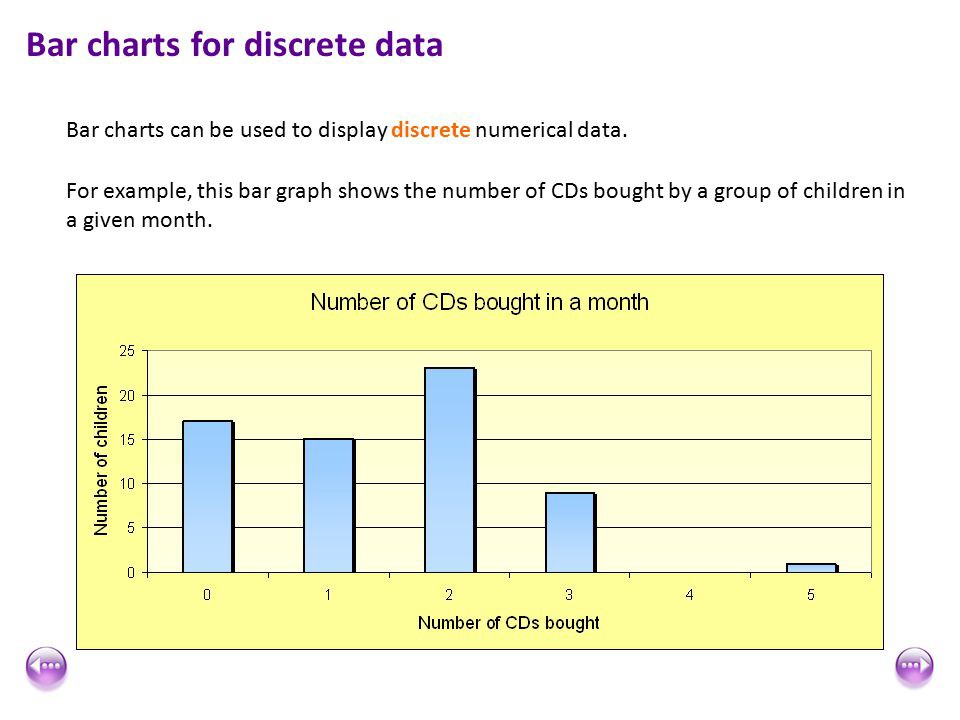
2. Basit bir ifadeyle, ayrık veriler yalnızca belirli değerleri alabilir ve veri değişkenleri daha küçük parçalara bölünemez.

3. Sınırlı sayıda olası değere sahiptir; ayın günleri.

Ayrık verilere örnekler:

• Bir sınıftaki öğrenci sayısı.

• Bir şirketteki işçi sayısı.

•Doğru cevapladığınız test sorularının sayısı.

**Sürekli Veri**

1. Anlamlı bir şekilde daha kücük sayilara bölünebilecek bilgiyi temsil eder. Bir ölçekte veya sürekli ölçülebilir ve hemen hemen her sayısal değere sahip olabilir.

Örneğin boyumuzu metre, santimetre, milimetre gibi farklı birimlerle çok hassas ölçeklerde ölçebiliyoruz.

2. Sürekli ve ayrık veri türleri arasındaki temel fark, birincisinde genişlik, sıcaklık, zaman vb. gibi birçok farklı ölçümde sürekli verileri kaydedebilmemizdir.

3. Sürekli değişkenler iki sayı arasındaki herhangi bir değeri alabilir. Örneğin, 60 ila 82 inç aralığı arasında 62,04762 inç, 79,948376 inç vb. gibi milyonlarca olası yükseklik vardır.

4. Verilerin sürekli mi yoksa ayrık mı olduğunu tanımlamanın iyi bir kuralı şudur: eğer ölçüm noktası yarıya indirilebiliyorsa ve hala anlamlıysa, veriler süreklidir.

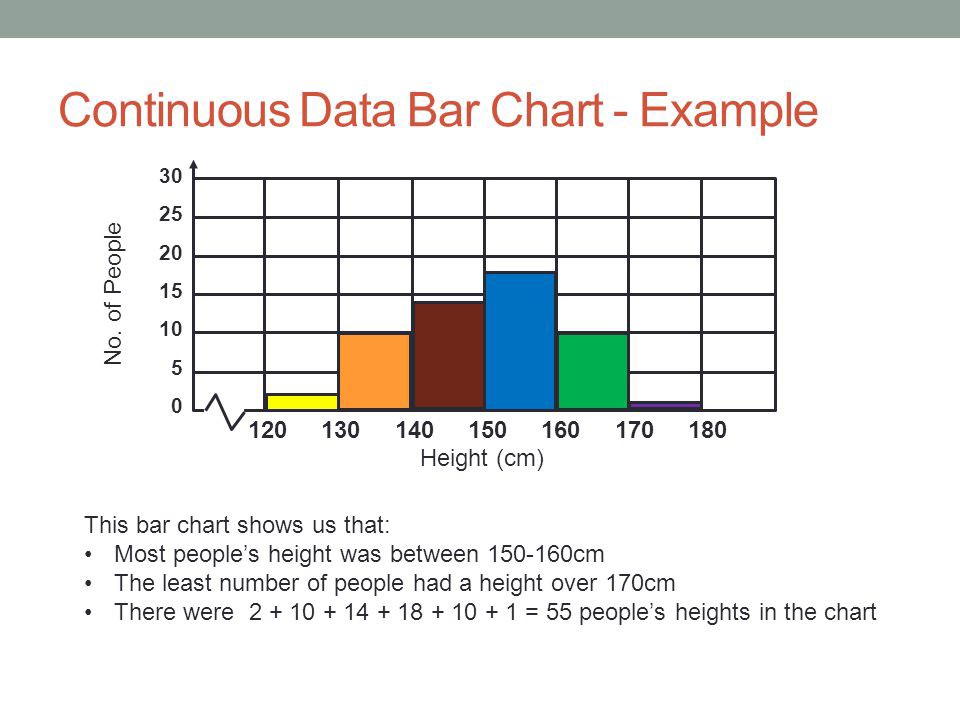
5. Sürekli veriler

Sürekli veri örnekleri:

•Bir projeyi tamamlamak için gereken süre.

•Çocukların boyu.

• Arabaların hızı.



**Aralık verileri**

1. Bu veri türleri ölçülebilir ve en yakın öğelerle sıralanır ancak anlamlı bir sıfıra sahip değildir.

“Aralık Ölçeği”nin anlamını anlayalım:

Aralık ölçeğinde 'Aralık' terimi aradaki boşluğu ifade eder ve bu, hatırlanması gereken önemli bir şeydir, çünkü aralık ölçekleri bizi yalnızca sıralama konusunda eğitmekle kalmaz, aynı zamanda her öğe arasındaki değer hakkında da bilgi verir.

2. Temel olarak aralık verilerini oran verileriyle aynı şekilde gösterebiliriz ancak dikkat etmemiz gereken şey onların karakterize edilen sıfır noktalarıdır.

3. Dolayısıyla aralık verilerinin yardımıyla verilerin derecelerini kolayca ilişkilendirebilir, değerleri ekleyebilir veya çıkarabiliriz.

4. Aralık verileri için hesaplayabileceğimiz bazı tanımlayıcı istatistikler vardır:

•Merkezi eğilim ölçüleri (ortalama, medyan, mod)

•Aralık (minimum, maksimum)

•Yayılma (yüzdelikler, çeyrekler arası aralık ve standart sapma).

Hesaplanacak istatistiksel şeyler yalnızca bunlar değil, ama daha fazlasını da hesaplayabiliriz.

Aralık verilerine örnekler:

•Sıcaklık (°C veya F, ancak Kelvin değil)

•Tarihler (1055, 1297, 1976, vb.)

• 12 saatlik zaman diliminde Zaman Boşluğu (sabah 6, akşam 6)

**Oran Verileri**

1. Bu veriler de aynı farka sahip sıralı birimlerdedir.

2. Oran değerleri aralık değerleriyle aynıdır ancak tek fark Oran verilerinin mutlak sıfıra sahip olmasıdır. Örneğin boy, kilo, uzunluk vb.

3. Bunlar, anlamlı bir sıfıra sahip eşit mesafeli öğelerle ölçülür ve sıralanır ve asla aralık verileri gibi negatif olmaz.

Bunu çarpıcı bir örnekle anlayalım: Yükseklik ölçümü.

Yükseklik santimetre, inç, metre veya fit gibi birimlerle ölçülebilir ve negatif bir yükseklik değeri elde etmek mümkün değildir.

4. Değişkenlerin sırası, aralarındaki zıtlıklar konusunda bizi aydınlatır ve kesinlikle sıfırdır.

5. Oran verileri temel olarak aralık verileriyle aynıdır; sıfır dışında sıfır anlamına gelir.

6. Oran verileri için hesaplayabileceğimiz tanımlayıcı istatistikler, aşağıdaki gibi aralık verileriyle aynıdır:

•Merkezi eğilim ölçüleri (ortalama, medyan, mod)

•Aralık (minimum, maksimum)

•Yayılma (yüzdelikler, çeyrekler arası aralık ve standart sapma).

Oran verilerine örnek:

• Yaş (0 ile 100+ arası)

• Sıcaklık (Kelvin cinsinden, ancak °C veya F değil)

• Zaman aralığı (kronometre veya benzeri bir aletle ölçülür)

Yukarıdaki oran verileri örneklerinde, kişinin yaşı, mutlak sıfır, belirli bir noktadan veya zamandan hesaplanan mesafe gibi gerçek ve anlamlı bir sıfır noktasının olduğunu görüyoruz. Hepsinin gerçek sıfırları var.

**NOT:**

Eğer terazinin sıfır noktasını subjektif olarak seçersek, bu noktada veriler oran verisi olamaz, aralık verisi olmalıdır.

<https://www.analyticsvidhya.com/blog/2021/06/complete-guide-to-data-types-in-statistics-for-data-science/>

# Tanımlayıcı İstatistikleri Anlamak

<https://youtu.be/txNvZ3Zndak>

İstatistik, verilerin toplanması, düzenlenmesi ve yorumlanmasıyla ilgilenen bir matematik dalıdır.

Başlangıçta verileri elde ettiğimizde süslü algoritmalar uygulamak ve bazı tahminler yapmak yerine, öncelikle istatistiksel teknikleri uygulayarak verileri okumaya ve anlamaya çalışırız. Bunu yaparak ne tür bir dağıtım verisine sahip olduğunu anlayabiliyoruz.

Bu derste su sorulara cevap vermeyi amaçliyoruz:

1. Tanımlayıcı İstatistik Nedir?

2. Tanımlayıcı İstatistik Türleri?

3. Merkezi Eğilim Ölçüsü (Ortalama, Medyan, Mod)

4. Yayılım / Dağılım Ölçüsü (Standart Sapma, Ortalama Sapma, Varyans, Yüzdelik, Çeyrekler, Çeyrekler Arası Aralık)

5. Çarpıklık Nedir?

6. Basıklık nedir?

7. Korelasyon Nedir?

Bugün, tanımlayıcı istatistikleri kesin olarak anlayalım. Hadi başlayalım,

**Tanımlayıcı İstatistik Nedir?**

Tanımlayıcı istatistikler, verilerin kolayca anlaşılabilmesi için özetlenmesini ve düzenlenmesini içerir. Tanımlayıcı istatistikler, çıkarımsal istatistiklerin aksine, verileri tanımlamaya çalışır ancak örneklemden tüm popülasyona yönelik çıkarımlar yapmaya çalışmaz. Burada genellikle verileri bir örnekte tanımlarız. Bu genellikle tanımlayıcı istatistiklerin, çıkarımsal istatistiklerden farklı olarak olasılık teorisi temelinde geliştirilmediği anlamına gelir.

**Tanımlayıcı İstatistik Türleri?**

Tanımlayıcı istatistikler iki kategoriye ayrılır. Merkezi eğilim ölçüleri ve değişkenlik ölçüleri (yayılma).

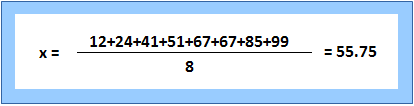
**Merkezi Eğilimin Ölçüsü**

Merkezi eğilim, tüm ölçüm kümesini en iyi şekilde özetleyen bir sayının yani bir şekilde kümenin "merkezi" olan bir sayının olduğu fikrini ifade eder.

**Genel ortalama**

Genel Ortalama veya Ortalama, verilerin merkezi bir eğilimidir, yani tüm verinin etrafına yayıldığı bir sayıdır. Bir bakıma tüm veri setinin değerini tahmin edebilen tek bir sayıdır.

8 tam sayı içeren veri setinin ortalamasını hesaplayalım.



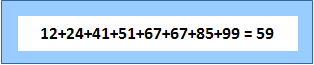
**Medyan**

Medyan, veriyi 2 eşit parçaya bölen değerdir, yani veriler artan veya azalan sırada düzenlendiğinde sağ tarafındaki terim sayısı ile sol tarafındaki terim sayısı aynıdır.

**Not:** Verileri azalan düzende sıralarsanız medyanı etkilemez ancak IQR negatif olur. Bu blogun ilerleyen kısımlarında IQR hakkında konuşacağız.

Terim sayısı tek ise medyan orta terim olacaktır

Eğer terimlerin sayısı çift ise medyan ortadaki 2 terimin ortalaması olacaktır.



Medyan 59'dur ve bir sayı kümesini eşit iki parçaya böler. Kümede çift sayılar olduğundan cevap ortadaki 51 ve 67 sayıların ortalamasıdır.

Not: Değerler aritmetik ilerleme halindeyken (ardışık terimler arasındaki fark sabittir. Burada 2.) medyan her zaman ortalamaya eşittir.



Bu 5 sayının ortalaması 6 ve medyandır.

**Mod**

Mod, veri setinde maksimum süreyi gösteren terim yani en yüksek frekansa sahip olan terimdir.



Bu veri setinde mod 67'dir çünkü diğer değerlerden daha fazlasına sahiptir,

yani iki kez.

Ancak tüm değerler aynı sayıda göründüğü için hiçbir modun olmadığı bir veri kümesi de olabilir. İki değer aynı anda ve geri kalan değerlerden daha fazla görünüyorsa, veri seti iki modludur. Üç değer aynı anda ve geri kalan değerlerden daha fazla görünüyorsa, veri seti üç modludur ve n mod için bu veri seti çok modludur.

**Yayılma / Dağılım Ölçüsü**

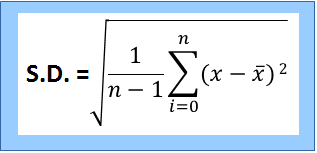
Yayılma Ölçüsü, verileriniz içindeki değişkenlik fikrini ifade eder.

**Standart sapma**

Standart sapma, her büyüklük ile ortalama arasındaki ortalama mesafenin ölçümüdür. Yani veriler ortalamadan nasıl yayılır. Düşük standart sapma, veri noktalarının veri kümesinin ortalamasına yakın olma eğiliminde olduğunu gösterirken, yüksek standart sapma, veri noktalarının daha geniş bir değer aralığına yayıldığını gösterir.

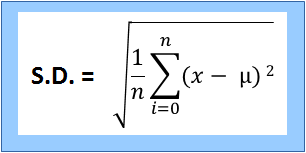
Örneklem veya popülasyon Standart Sapması arasında seçim yapmamız gereken durumlar vardır.

Bir popülasyonun bir kısmının, bir popülasyon segmentinin SD'sini bulmamız istendiğinde; daha sonra örnek Standart Sapmayı kullanırız.



Nerede > bir numunenin ortalamasıdır.

Ancak tüm bir popülasyonla ilgilenmemiz gerektiğinde, popülasyon Standart Sapmasını kullanırız.



µ bir popülasyonun ortalamasıdır.

Örneklem bir popülasyonun parçası olsa da SD formüllerinin aynı olması gerekirdi ama öyle değil. Bu konuda daha fazla bilgi edinmek için bu bağlantıya bakın

Bildiğiniz gibi tanımlayıcı istatistiklerde genellikle bir popülasyonda değil, bir örneklemde mevcut olan verilerle ilgileniriz. Yani önceki veri setini kullanırsak ve örnek formüldeki değerleri yerine koyarsak,



Ve cevap 29,62 olacaktır.

**Varyans**

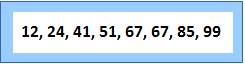
Varyans, her miktar ile ortalama arasındaki ortalama mesafenin karesidir. Yani standart sapmanın karesidir.



Ve cevap 877,34 olacaktır.

**Aralık(Range)**

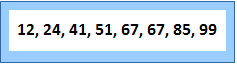
Aralık, tanımlayıcı istatistiklerin en basit tekniklerinden biridir. En düşük ve en yüksek değer arasındaki farktır.



Aralık 99–12 = 87

#### **Yüzdelik (Percentile**)

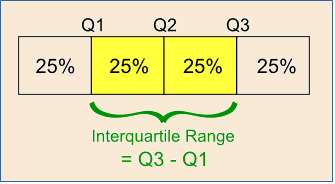
Yüzdelik değer, veri kümesindeki bir değerin konumunu temsil etmenin bir yoludur. Yüzdelik dilimin hesaplanması için veri setindeki değerlerin her zaman artan sırada olması gerekir.



Medyan 59 kendisinden 8 üzerinden 4 değer eksiktir. Şu şekilde de söylenebilir: Veri setinde 59 50. yüzdeliktir çünkü toplam terimlerin %50'si 59'dan küçüktür. Genel olarak k n'inci yüzdelik ise, toplam terimlerin %n'sinin k'den küçük olduğu anlamına gelir.

**Çeyrekler(Quartiles)**

İstatistik ve olasılıkta çeyrekler, verilerin artan düzende(ascending order olarak) sıralanması koşuluyla verilerinizi çeyreklere bölen değerlerdir.



<https://statsmethods.wordpress.com/2013/05/09/iqr/>)

Üç çeyrek değer vardır. İlk çeyrek değeri yüzde 25'te. İkinci çeyrek yüzde 50, üçüncü çeyrek yüzde 75'tir. İkinci çeyrek (Q2) tüm verilerin ortancasıdır. Birinci çeyrek (Q1), verilerin üst yarısının medyanıdır. Üçüncü Çeyrek (Q3) ise verilerin alt yarısının medyanıdır.

Yani burada, benzetme yoluyla,

Q2 = 67: tüm verinin yüzde 50'sidir ve medyandır. Q1 = 41: Verinin yüzde 25'idir.

Q3 = 85: tarihin yüzde 75'idir.

Çeyrekler arası aralık (IQR)= Q3 - Q1 = 85 - 41 = 44

**Not:** Verileri azalan düzende sıralarsanız IQR -44 olacaktır. Büyüklük aynı olacak, sadece işaret farklı olacaktır. Verileriniz azalan sıradaysa negatif IQR iyidir. Sadece daha büyük değerlerden daha küçük değerleri çıkarıyoruz, artan düzeni tercih ediyoruz (Q3 – Q1).

**Çarpıklık(Skewness)**

Çarpıklık, gerçek değerli bir rastgele değişkenin ortalamaya göre olasılık dağılımının asimetrisinin bir ölçüsüdür. Çarpıklık değeri pozitif, negatif ya da tanımsız olabilir.

Mükemmel bir normal dağılımda, eğrinin her iki tarafındaki kuyruklar birbirinin tam ayna görüntüleridir.

Bir dağılım sola çarpık olduğunda, eğrinin sol tarafındaki kuyruk sağ taraftaki kuyruktan daha uzundur ve ortalama moddan daha küçüktür. Bu duruma negatif çarpıklık da denir.

Bir dağılım sağa çarpık olduğunda, eğrinin sağ tarafındaki kuyruk sol taraftaki kuyruktan daha uzundur ve ortalama moddan daha büyüktür. Bu duruma pozitif çarpıklık da denir.



[https://www.safaribooksonline.com/library/view/clojure-for- data/9781784397180/ch01s13.html](https://www.safaribooksonline.com/library/view/clojure-for-%20data/9781784397180/ch01s13.html))

**Yorumlar (Interpretations)**

• Çarpıklığın yönü işaretle verilir. Sıfır hiçbir çarpıklığın olmadığı anlamına gelir.

• Negatif bir değer, dağılımın negatif olarak çarpık olduğu anlamına gelir. Pozitif bir değer, dağılımın pozitif çarpık olduğu anlamına gelir.

• Katsayı, örneklem dağılımını normal dağılımla karşılaştırır. Değer ne kadar büyük olursa, dağılım da normal dağılımdan o kadar farklı olur.

Örnek soru:

Aşağıdaki özelliklere sahip verilerin çarpıklığını bulmak için Pearson Katsayısı #1 ve #2'yi kullanın:

• Ortalama = 50.

• Medyan = 56.

• Mod = 60.

• Standart sapma = 8,5.

Pearson'un Birinci Çarpıklık Katsayısı: -1,17. Pearson'un İkinci Çarpıklık Katsayısı: -2,117.

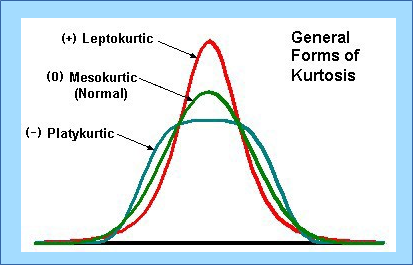
**Not:** Pearson'un ilk çarpıklık katsayısı modu kullanır. Bu nedenle, eğer değerlerin frekansı çok düşükse, bu durum merkezi eğilimin sabit bir ölçüsünü vermeyecektir. Örneğin, her iki veri kümesindeki mod da 9'dur: 1, 2, 3, 4, 4, 5, 6, 7, 8, 9.

İlk veri grubunda mod yalnızca iki kez görünür. Bu nedenle Pearson'un Birinci Çarpıklık Katsayısını kullanmak iyi bir fikir değildir. Ancak ikinci sette 1, 2, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 6, 7, 8, 9, 10, 12, 12, 13.

mod 4 8 kez görünür. Bu nedenle Pearson'un İkinci Çarpıklık Katsayısı size muhtemelen makul bir sonuç verecektir.

**Basıklık (Kurtosis)**

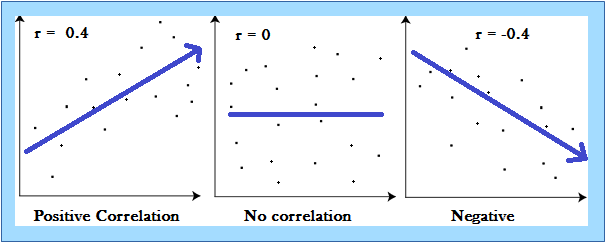
Basıklık ölçüsünün kesin yorumu eskiden tartışmalıydı ama artık çözüldü. Bu, aykırı değerlerin varlığıyla ilgilidir. Basıklık, normal dağılıma göre verilerin ağır kuyruklu (aykırı değerlerin çokluğu) veya hafif kuyruklu (aykırı değerlerin olmaması) olup olmadığının bir ölçüsüdür.



<https://mvpprograms.com/help/mvpstats/distributions/SkewnessKurtosis>

**Korelasyon**

Korelasyon, değişken çiftlerinin ilişkili olup olmadığını ve ne kadar güçlü bir şekilde ilişkili olduğunu gösterebilen istatistiksel bir tekniktir.



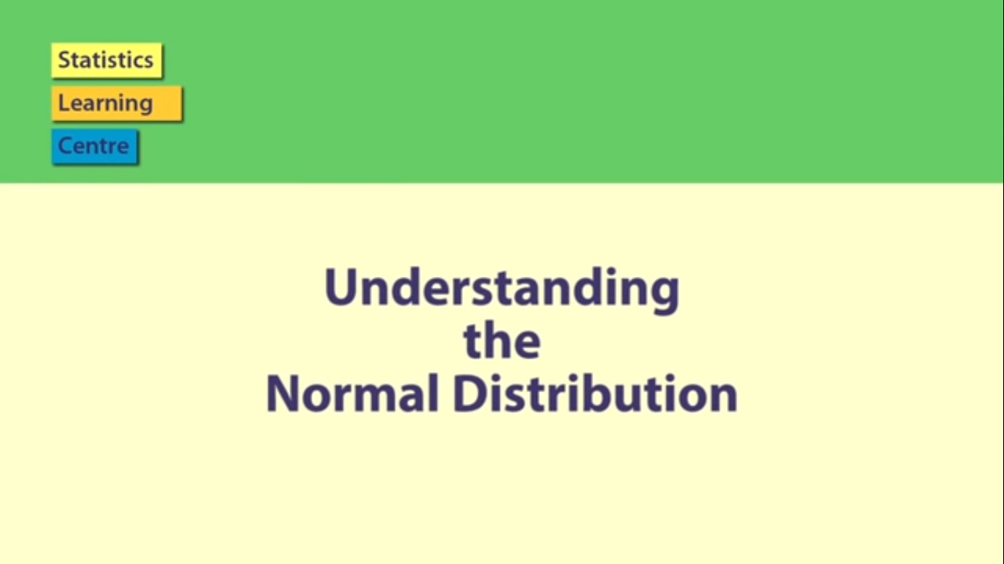
<http://www.statisticshowto.com/what-is-correlation/>

Bir korelasyonun ana sonucuna korelasyon katsayısı (veya “r”) denir. -1.0 ile +1.0 arasında değişir. r, +1 veya -1'e ne kadar yakınsa, iki değişken o kadar yakından ilişkilidir.

r'nin 0'a yakın olması değişkenler arasında ilişki olmadığı anlamına gelir. Eğer r pozitifse, değişkenlerden biri büyüdükçe diğerinin de büyüdüğü anlamına gelir. Eğer r negatifse, bu, biri büyüdükçe diğerinin küçüldüğü anlamına gelir (genellikle "ters" korelasyon olarak adlandırılır).

<https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291>

**NORMAL DISTRIBUTION**



<https://youtu.be/https://youtu.be/mtH1fmUVkfE>

Gauss dağılımı olarak da bilinen normal dağılım, istatistiklerde bağımsız, rastgele değişkenler için en önemli olasılık dağılımıdır. Çoğu kişi istatistiksel raporlarda tanıdık çan şeklindeki eğriyi tanır.

Normal dağılım, ortalaması etrafında simetrik olan, gözlemlerin çoğu merkezi zirve etrafında kümelenen ve ortalamadan daha uzaktaki değerlerin olasılıkları her iki yönde de eşit şekilde azalan sürekli bir olasılık dağılımıdır. Dağılımın her iki kuyruğundaki uç değerler de benzer şekilde olası değildir. Normal dağılım simetrik olsa da tüm simetrik dağılımlar normal değildir. Örneğin, Öğrencinin t, Cauchy ve lojistik dağılımları simetriktir.

Herhangi bir olasılık dağılımında olduğu gibi normal dağılım da bir değişkenin değerlerinin nasıl dağıldığını açıklar. Birçok doğa olayının değerlerinin dağılımını doğru bir şekilde tanımladığı için istatistikteki en önemli olasılık dağılımıdır. Birçok bağımsız sürecin toplamı olan özellikler sıklıkla normal dağılımları takip eder. Örneğin boy, tansiyon, ölçüm hatası ve IQ puanları normal dağılıma uymaktadır.

<https://statisticsbyjim.com/basics/normal-distribution/>

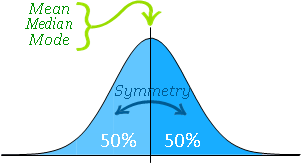
Sol tarafa daha fazla yayılabilir. Veya daha fazlası sağda. Ya da her şey karışabilir



Ancak verilerin sola veya sağa sapma olmadan merkezi bir değer etrafında olma eğiliminde olduğu ve bunun gibi bir "Normal Dağılım"a yaklaştığı birçok durum vardır:



Verilerin "normal dağıldığını" söylüyoruz:



Normal Dağılım şunları içerir:

• ortalama = medyan = mod

• merkeze göre simetri

• Değerlerin %50'si ortalamanın altında ve %50'si ortalamanın üzerinde

**Standart sapma**

Standart sapma, sayıların ne kadar yayıldığının bir ölçüsüdür. Standart sapmayı hesapladığımızda genel olarak şunu buluruz:



Standard Deviations

1. 1 1

68%

−3 −2 −1

+1 +2 +3

Değerlerin %68'i ortalamanın

1 standart sapması dahilindedir

95%

−3 −2 −1 +1 +2 +3

Değerlerin %95'i ortalamanın

2 standart sapması dahilindedir

99.7%

−3 −2 −1

+1 +2 +3

Değerlerin %99,7'si ortalamanın

3 standart sapması dahilindedir

**Örnek**:

Okuldaki öğrencilerin %95'inin boyu 1,1 ila 1,7 m arasındadır.

Bu verilerin normal dağıldığını varsayarak ortalama ve standart sapmayı hesaplayabilir misiniz?

**Cözüm**:

Ortalama 1,1m ile 1,7m arasındadır:

Ortalama = (1,1m + 1,7m) / 2 = 1,4m

%95, ortalamanın her iki tarafında 2 standart sapmadır (toplam 4 standart sapma):

1 standart sapma= (1,7m-1,1m) / 4

= 0,6m / 4

=0,15m

95%

0.95 1.11 1.25 1.4 1.55 1.7 1.85

Ve sonuç

Standart sapmayı bilmek iyidir çünkü herhangi bir değerin şöyle olduğunu söyleyebiliriz:

•muhtemelen 1 standart sapma dahilinde olmalıdır (100 üzerinden 68 olmalıdır)

•2 standart sapma dahilinde olması çok muhtemel (100 üzerinden 95 olmalıdır)

•neredeyse kesinlikle 3 standart sapma dahilinde (1000 üzerinden 997 olmalıdır)

**Standart Puanlar (Standard Scores)**



Ortalamadan standart sapmaların sayısına "Standart Puan(Skor)", "sigma" veya "z-puanı(skoru)" da denir. Bu sözlere alışın!



Bir değeri Standart Puana ("z-puanı") dönüştürmek için:

•önce ortalamayı çıkarın,

•sonra Standart Sapmaya bölün. Buna "Standartlaştırma" denir:

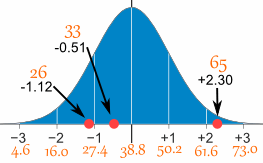
*Standardize*

950 970 990 1030 1050 1070 −3 −2 −1 +1 +2 +3

*Normal Bir Dağılım Standart Normal Dağılım*



Herhangi bir Normal Dağılım'ı alıp Standart Normal Dağılım'a dönüştürebiliriz.



**Örnek**: Seyahat Süresi

Günlük seyahat süresine ilişkin bir ankette şu sonuçlar elde edildi (dakika olarak):

26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34

**Ortalama 38,8** dakika, **Standart Sapma ise 11,4** dakikadır

Değerleri z puanlarına ("standart puanlar") dönüştürün.

**26'yı** dönüştürmek için:

önce ortalamayı çıkarın: 26 − 38,8 = −12,8,

daha sonra Standart Sapmaya bölün: −12,8/11,4 = −1,12 Yani **26**, **−1,12** Ortalamadan **Standart Sapmadır.**

İşte ilk üç dönüşüm

Orijinal Değer Hesaplama Standart Puanı (z-puanı)

**26 (26-38,8) / 11,4 = −1,12**

**33 (33-38,8) / 11,4 = −0,51**

**65 (65-38,8) / 11,4 = +2,30**

Ve işte grafiksel olarak:

Z-puanlarının geri kalanını

kendiniz hesaplayabilirsiniz!

Kullandığımız **z-puanı formülü** şöyledir:

**z = (x-μ)/σ**

•**z** "z-puanı"dır (Standart Puan) •**x** standartlaştırılacak değerdir

•**μ** ('mü") ortalamadır •**σ** ("sigma") standart sapmadır

Ve bu nasıl kullanılacağı:

Örnek: Seyahat Süresi (devam)

"Z-puanı formülünü" kullanan ilk üç dönüşüm şunlardır:

**z = (x-μ)/σ**

•μ = 38,8 , •σ = 11,4 🡪 Daha önce formülü kullanarak yaptığımız hesaplamalar.

|  |  |  |
| --- | --- | --- |
| ***x*** | ***(x − μ)/σ*** | ***z***  ***(z-score)*** |
| *26* | *(26 − 38,8)/11,4* | *= −1,12* |
| *33* | *(33 − 38,8)/11,4* | *= −0,51* |
| *65* | *(65 − 38,8)/11,4* | *= +2,30* |
| *...* | *...* | *...* |

**Neden Standartlaştıralım...?**

Verilerimiz hakkında karar vermemize yardımcı olabilir.

Ayrıca her ortalama ve standart sapma değeri için ayrı ayrı hesaplamalar yapmak yerine yalnızca tek bir tabloya (Standart Normal Dağılım Tablosu) ihtiyacımız olduğundan hayatımızı kolaylaştırır.

örnek: Profesör bir sınava not veriyor. İşte öğrencilerin sonuçları (60 puan üzerinden):

20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17

Çoğu öğrenci 60 üzerinden 30 bile alamadı ve başarısız olacak.

Test gerçekten zor olmuş olmalı, bu yüzden Prof tüm puanları standartlaştırmaya karar verir ve yalnızca ortalamanın 1 standart sapmadan daha düşük olan kişileri başarısız kılar.

Ortalama 23, Standart Sapma 6,6 olup Standart Puanlar şunlardır:

-0,45, -1,21, 0,45, 1,36, -0,76, 0,76, 1,82, -1,36, 0,45, -0,15, -0,91

Artık sadece 2 öğrenci başarısız olacak (−1 standart sapmadan düşük olanlar) Çok daha adil!

Daha Detaylı

Standart sapmanın her yarısı için yüzdeleri ve kümülatif yüzdeleri içeren Standart Normal Dağılım şöyledir:



**Örnek**:

Yakın zamanda yapılan bir testte puanınız ortalamanın 0,5 standart sapma üzerindeydi, kaç kişi sizden daha düşük puan aldı?

•0 ile 0,5 arasında %19,1

•0'dan küçük %50'dir (eğrinin sol yarısı)

Yani sizden daha az olan toplam:

%50 + %19,1 = %69,1

Teorik olarak %69,1'i sizden daha az puan aldı (ancak gerçek verilerle yüzde farklı olabilir)

Pratik Bir Örnek: Firmanız şekeri 1 kg'lık torbalarda paketliyor.

Bir torba örneğini tarttığınızda şu sonuçları alırsınız:

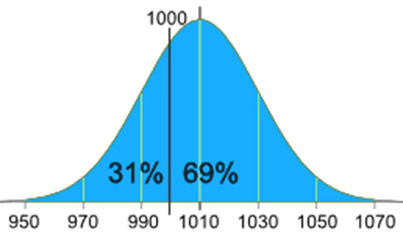
1007g, 1032g, 1002g, 983g, 1004g,... (yüz ölçüm)

•Ortalama = 1010g

•Standart Sapma = 20g

Bazı değerler 1000 gramın altında... bunu düzeltebilir misiniz?

Ölçümlerinizin normal dağılımı şöyle görünür:



Torbaların %31'i 1000 gramın altında, bu da müşteriyi aldatıyor!

Bu rastgele bir şeydir, bu nedenle poşetlerin 1000 gramdan az olmasını engelleyemeyiz, ancak bunu oldukça azaltmaya çalışabiliriz.

Makineyi 1000gr olacak şekilde ayarlayalım:

• −3 standart sapmada:

**Yukarıdaki büyük çan eğrisinden %0,1'in daha az olduğunu görüyoruz. Ama belki bu çok küçüktür.**

• −2,5 standart sapmada:

**3'ün altı %0,1 ve 3 ile 2,5 arası standart sapma %0,5, yani %0,1 + %0,5 = %0,6 (iyi bir seçim bence)**

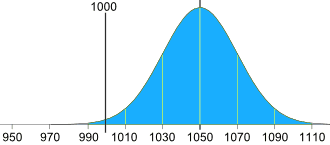
O halde makineyi ortalamadan −2,5 standart sapmayla 1000g olacak şekilde ayarlayalım.

Artık bunu şu şekilde ayarlayabiliriz:

• her torbadaki şeker miktarını artırın (bu, ortalamayı değiştirir) veya

•daha doğru hale getirin (bu da standart sapmayı azaltır) Her ikisini de deneyelim.

Her torbadaki ortalama miktarı ayarlayın

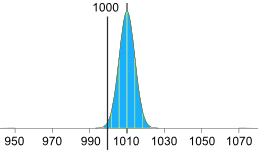


Standart sapma 20g'dir ve bunlardan 2,5'ine ihtiyacımız var:

2,5 × 20g = 50g

Yani makine ortalama olmalı

1050g, şöyle:

Makinenin doğruluğunu ayarlayın

Veya aynı ortalamayı (1010g) koruyabiliriz,

ancak 10g'ye eşit olması için 2,5 standart sapmaya ihtiyacımız var:

10g / 2,5 = 4g

Yani standart sapma şu şekilde 4g olmalıdır: (Umarız makine bu kadar doğrudur!)

Ya da belki daha iyi doğruluk ve biraz daha büyük ortalama boyutun bir kombinasyonuna sahip olabiliriz.

<https://www.mathsisfun.com/data/standard-normal-distribution.html#google_vignette>

Şimdi senin sıran. Aşağıdaki soruları yanıtlayın!

[Question1](https://www.mathopolis.com/questions/q.html?id=2619&amp;t=mif&amp;qs=2619_2620_2621_2622_2623_2624_2625_2626_3844_3845&amp;site=1&amp;ref=2f646174612f7374616e646172642d6e6f726d616c2d646973747269627574696f6e2e68746d6c&amp;title=4e6f726d616c20446973747269627574696f6e)

[Question2](https://www.mathopolis.com/questions/q.html?id=2621&amp;t=mif&amp;qs=2619_2620_2621_2622_2623_2624_2625_2626_3844_3845&amp;site=1&amp;ref=2f646174612f7374616e646172642d6e6f726d616c2d646973747269627574696f6e2e68746d6c&amp;title=4e6f726d616c20446973747269627574696f6e)

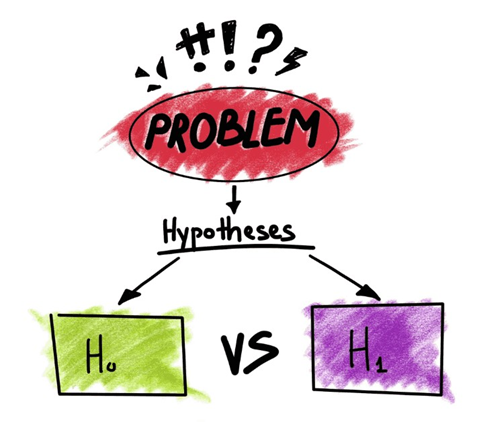
[Question3](https://www.mathopolis.com/questions/q.html?id=2623&amp;t=mif&amp;qs=2619_2620_2621_2622_2623_2624_2625_2626_3844_3845&amp;site=1&amp;ref=2f646174612f7374616e646172642d6e6f726d616c2d646973747269627574696f6e2e68746d6c&amp;title=4e6f726d616c20446973747269627574696f6e)

[Question4](https://www.mathopolis.com/questions/q.html?id=2625&amp;t=mif&amp;qs=2619_2620_2621_2622_2623_2624_2625_2626_3844_3845&amp;site=1&amp;ref=2f646174612f7374616e646172642d6e6f726d616c2d646973747269627574696f6e2e68746d6c&amp;title=4e6f726d616c20446973747269627574696f6e)

[Question5](https://www.mathopolis.com/questions/q.html?id=3844&amp;t=mif&amp;qs=2619_2620_2621_2622_2623_2624_2625_2626_3844_3845&amp;site=1&amp;ref=2f646174612f7374616e646172642d6e6f726d616c2d646973747269627574696f6e2e68746d6c&amp;title=4e6f726d616c20446973747269627574696f6e)

[Normal Distribution Application](https://colab.research.google.com/drive/1Kn6Hv8HC1XQncib4AyZ6G2__TZ2OJYWg#scrollTo%3DJiPByRGIxVeB)

**HİPOTEZ TESTİ**

****

[**https://youtu.be/vRIIUDhCKU8**](https://youtu.be/vRIIUDhCKU8)

Hipotez testi veya anlamlılık testi, bir örneklemde ölçülen verileri kullanarak bir popülasyondaki bir parametre hakkındaki bir iddiayı veya hipotezi test etmek için kullanılan istatistiksel bir yöntemdir.

Hipotez Testi temel olarak popülasyon parametresi hakkında yaptığımız bir varsayımdır.

Örnek: Bir sınıfta öğrencilerin aldığı ortalama not 60'tır.

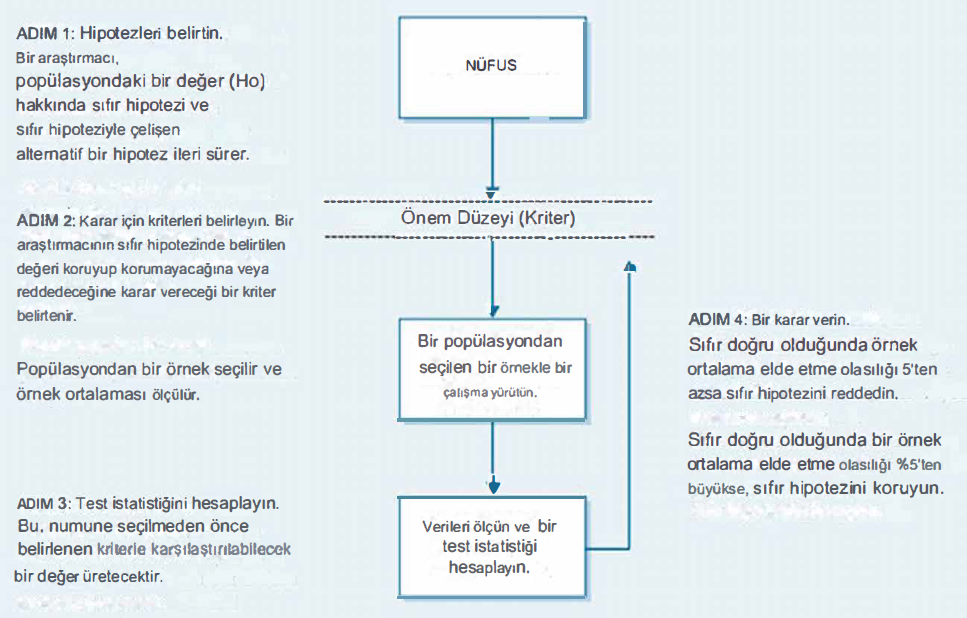
**Hipotez Testini Gerçekleştirme Adımları:**

**Adım 1:** Hipotezleri belirtin.

**Adım 2:** Karar için kriterleri belirleyin.

**Adım 3:** Test istatistiğini hesaplayın.

**Adım 4:** Bir karar verin



**Adım 1: Hipotezleri belirtin:**

Yukarıdaki örnekte Öğretmen bir sınıftaki öğrencilerin ortalama notlarının 60 olduğunu iddia etmiştir.

Şimdi bu Hipotezi bazı öğrenci notları örneklerini dikkate alarak test edeceğiz.

Ortalama not 60 ise Öğretmen beyanını kabul edeceğiz, aksi halde beyanı reddedeceğiz.

Böylece iki parametrenin olduğunu görebiliriz. Bunu Sıfır Hipotezi(H0) ve Alternatif Hipotez(H1) olarak adlandırıyoruz.



**Sıfır Hipotezi(H0):**

Boş Hipotezi, iki değişken arasında hiçbir ilişkinin olmadığını, yani bir değişkenin diğer değişken üzerinde etkisinin olmadığını varsayar.

Basit bir deyişle Sıfır Hipotezi, alan veya problem bilgisine dayalı olarak yapılan temel bir varsayımdır.

Örnek: H0: Ortalama puan=60.

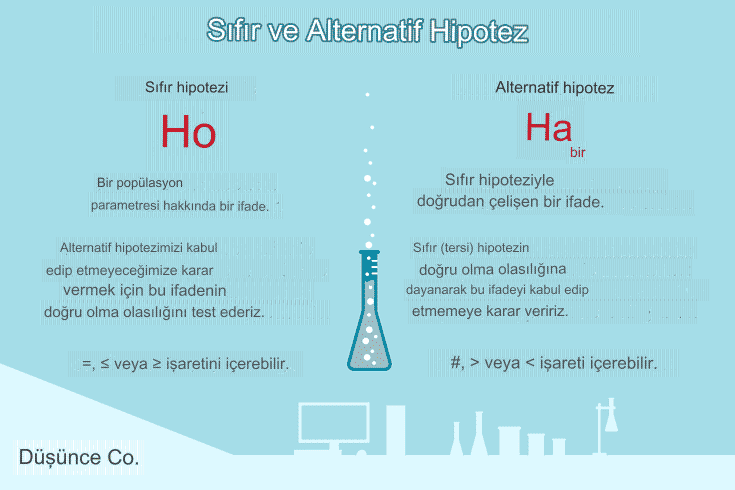


Hipotez Testi, Makine öğrenimine yeni başlayanların çoğunun şaşıracağı yaygın bir kelimedir ve ben de bu Testin nasıl yapılabileceğini tam olarak anlamak için çok zaman harcadım.

**Alternatif Hipotez(H1):**

Alternatif hipotez, hipotez testinde kullanılan sıfır hipotezine aykırı olan hipotezdir.

Örnek : H1: Ortalama notlar!= 60



**Sıfır Hipotezi vs Alternatif Hipotez**

**Not:**

Sıfır hipotezi reddedilirse alternatif hipotezi kabul ederiz. Sıfır hipotezi reddedilmezse alternatif hipotezi kabul etmiyoruz.

**Adım 2:** Karar için kriterleri belirleyin:

Bir karara ilişkin kriterleri belirlemek için testin önem düzeyini belirtiriz.

**Güven düzeyi:**

Verilen ifadeye ilişkin olarak ne kadar emin olduğumuzla ilgili bir şeydir. Genellikle %100 emin olmayacağız, dolayısıyla %100'den az olacaktır.

Şimdi önem düzeyinin ne olduğunu anlayalım.

**Önem düzeyi:**

Sıfır hipotezini kabul ettiğimiz veya reddettiğimizin önem derecesini ifade eder.

Bir hipotezi kabul etmek veya reddetmek için %100 doğruluk mümkün değildir, bu nedenle genellikle %5 olan bir anlamlılık düzeyi seçiyoruz.

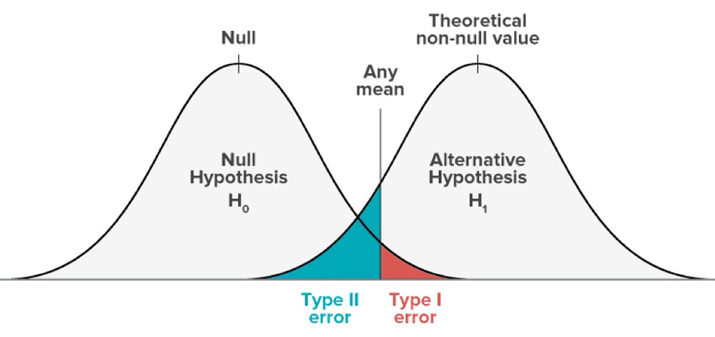
Önem düzeyi α ile gösterilir. α= 1-güven düzeyi

**Not:**

Anlamlılık düzeyi genellikle %5 olarak ayarlanırsa. Sıfır hipotezi doğruysa, örnek ortalama elde etme olasılığı %5'ten az olduğunda, sıfır hipotezinde belirtilen değeri reddederiz.

**Tip I hata:** Hipotez doğru olmasına rağmen sıfır hipotezini reddettiğimizde.

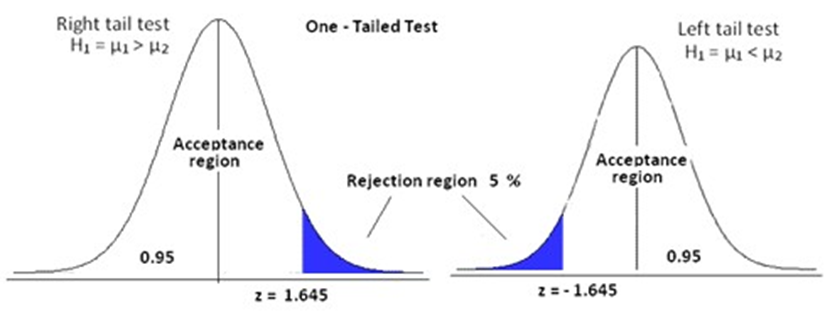
**Tip II hatalar:** Sıfır hipotezini kabul ettiğimizde ancak yanlış olduğu zaman.



**Tek Kuyruk Testi:**

Reddedilen bölgenin örnekleme dağılımının yalnızca bir tarafında olduğu istatistiksel hipotez testine tek kuyruklu test denir.

Örnek : Öğrencilerin ortalama notları<=40

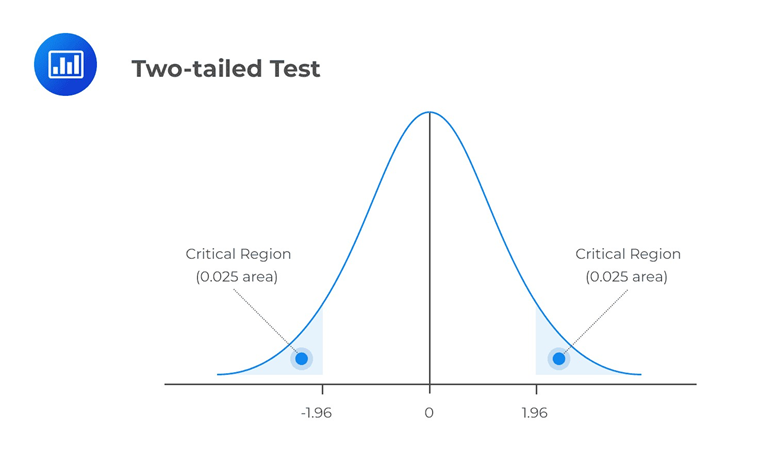


**İki kuyruklu test:**

İki kuyruklu test, bir dağılımın kritik alanının iki taraflı olduğu ve bir numunenin belirli bir değer aralığından büyük veya küçük olup olmadığını test eden istatistiksel bir testtir.

Test edilen numune kritik alanlardan herhangi birine giriyorsa sıfır hipotezi yerine alternatif hipotez kabul edilir.

Örnek: Öğrencilerin ortalama notları!= 40 öğrenci



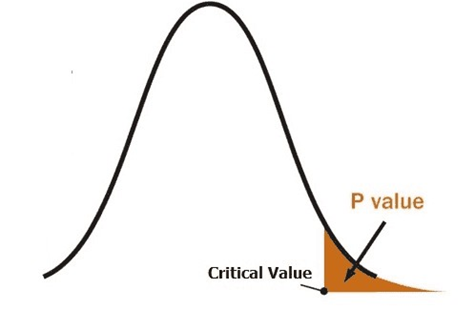
**P-değeri:**

P değeri veya hesaplanan olasılık, gözleneni bulma olasılığıdır.

Sıfır hipotezini desteklemenize veya reddetmenize yardımcı olmak için hipotez testinde p değeri kullanılır.

P değeri sıfır hipotezine karşı kanıttır.

P değeri ne kadar küçük olursa, sıfır hipotezini reddetmeniz gerektiğine dair kanıt o kadar güçlü olur.



**Not:**

P değeriniz seçilen anlamlılık seviyesinden düşükse boş hipotezi reddedersiniz, yani örneğinizin alternatif hipotezi destekleyecek makul kanıtlar sağladığını kabul edersiniz.

**Adım 3:** Test istatistiğini hesaplayın:

Bunlar Hipotez Testi için en yaygın kullanılan testlerdir.

• Z Testi

• T Testi

• ANOVA Testi

• Ki-kare testi

**Z Testi:**

Z testi, z istatistiğinin normal bir dağılım izlediği bir hipotez testidir.

Z testi en iyi şekilde 30'dan büyük numuneler için kullanılır, çünkü merkezi limit teoremine göre numune sayısı arttıkça numunelerin yaklaşık olarak normal dağıldığı kabul edilir.

• Z testi, varyanslar bilindiğinde ve örneklem büyüklüğü büyük olduğunda iki popülasyon ortalamasının farklı olup olmadığını belirlemek için yapılan istatistiksel bir testtir.

• Z testinin normal dağılım izlediği hipotezleri test etmek için kullanılabilir.

• Z istatistiği veya z puanı, z testinin sonucunu temsil eden bir sayıdır.

• Standart sapma bilindiğinde Z testi yapılır.



[**https://youtu.be/I10q6fjPxJ0**](https://youtu.be/vRIIUDhCKU8)

İstatistik, analistlerin verilen verilerden çıkarım yapmak için kullandıkları istatistiksel testlerin düzenlenmesidir. Bu testler verilerden gözlemlenen modele dayanarak karar vermemizi sağlar. Çok çeşitli istatistiksel testler vardır. Hangi istatistiksel testin kullanılacağı verinin yapısına, dağılımına ve değişken türüne bağlıdır. İstatistikte t-testi, Z-testi, ki-kare testi, anova testi gibi birçok farklı test türü bulunmaktadır. Binom testi, tek örnek medyan testi vb.

**İstatistiksel bir test seçmek**

Veriler normal dağılıyorsa parametrik testler kullanılır. Parametrik istatistiksel test, popülasyon parametreleri ve popülasyonun oluşturduğu dağılımlar hakkında bir varsayımda bulunur.

gelen veriler. Bu tür testler, verilerin normal dağılımdan olduğunu varsayan t testleri, z testleri ve anova testlerini içerir.

**Z testi -** Z testi, varyanslar bilindiğinde ve örneklem büyüklüğü büyük olduğunda iki popülasyon ortalamasının farklı olup olmadığını belirlemek için kullanılan istatistiksel bir testtir. Z-testinde popülasyonun ortalaması karşılaştırılır. Kullanılan parametreler popülasyon ortalaması ve popülasyon standart sapmasıdır. Z testi, alınan örneğin aynı popülasyona ait olduğu hipotezini doğrulamak için kullanılır.

Ho: Örneklem ortalaması popülasyon ortalaması ile aynıdır (Boş hipotez)

Ha: Örneklem ortalaması popülasyon ortalaması ile aynı değil (Alternatif hipotez)

**z = (x — μ) / (σ / √n),**

burada , x=örnek ortalama, u=nüfus ortalaması, σ / √n = popülasyon standart sapması. Z değeri kritik değerden küçükse boş hipotezi kabul edin, aksi halde sıfır hipotezini reddedin.

**T-testi-**T-testinde verilen iki örneğin ortalaması karşılaştırılır. Popülasyon parametreleri (ortalama ve standart sapma) bilinmediğinde t testi kullanılır.

Aynı popülasyondaki iki değişken arasındaki farka yönelik Eşleştirilmiş T Testi Testleri (ön test ve son test puanı).

Örneğin - Bir eğitim programında stajyerin programın tamamlanmasından önceki ve sonraki performans puanı.

**Bağımsız T-testi-** İki örnek t-testi veya öğrenci t-testi olarak da adlandırılan bağımsız t-testi, ilgisiz iki gruptaki ortalamalar arasında istatistiksel olarak anlamlı bir fark olup olmadığını belirleyen istatistiksel bir testtir.

Örneğin - bir popülasyondaki kız ve erkek çocukların karşılaştırılması.

**Tek örnek t-testi:** Tek bir grubun ortalaması belirli bir ortalamayla karşılaştırılır. Örneğin ortalama satışlar verilmişse satışlardaki artış ve azalışları kontrol etmek.

**t = (x1 — x2) / (σ / √n1 + σ / √n2),**

burada x1 ve x2 sırasıyla numune 1 ve numune 2'nin ortalamasıdır.

**ANOVA Testi -** Varyans analizi (ANOVA), iki veya daha fazla grubun ortalamalarının birbirinden önemli ölçüde farklı olup olmadığını kontrol etmek için kullanılan istatistiksel bir tekniktir.

ANOVA, farklı numunelerin ortalamalarını karşılaştırarak bir veya daha fazla faktörün etkisini kontrol eder. ANOVA testi yerine t testi kullanırsak örnek sayısı ikiden fazla olduğu için güvenilir olmayacak ve sonuçta hata verecektir.

ANOVA'da test edilen hipotez şu şekildedir:

**Ho:** Tüm örnek çiftleri aynıdır, yani tüm örnek ortalamaları eşittir Ha: En az bir örnek çifti önemli ölçüde farklıdır

Anova testinde F değerini hesaplayıp kritik değerle karşılaştırıyoruz.

**F= ((SSE1 — SSE2)/m)/ SSE2/n-k,**

SSE = kalan kareler toplamı

m = kısıtlama sayısı

k = bağımsız değişkenlerin sayısı

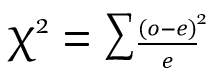
**Parametrik olmayan istatistiksel test:** Veriler normal şekilde dağılmadığında parametrik olmayan testler kullanılır. Parametrik olmayan testler ki-kare testini içerir.

**Ki-kare testi(χ2 testi)-**ki-kare testi iki kategorik değişkeni karşılaştırmak için kullanılır. Ki-Kare istatistik değerinin hesaplanması ve bunun Ki-Kare dağılımından alınan kritik bir değerle karşılaştırılması, gözlemlenen frekansın beklenen frekanstan önemli ölçüde farklı olup olmadığının değerlendirilmesine olanak tanır.

Ki-kare için test edilen hipotez şudur:

Ho: Değişken x ve Değişken y bağımsızdır

Ha: Değişken x ve Değişken y bağımsız değildir.

Ki-kare formülü  burada o=gözlenen, e=beklenen.

Ekstralar:

[https://medium.com/@anushka.da3/types-of-statistical-tests-b8ceb90e13b3](https://medium.com/%40anushka.da3/types-of-statistical-tests-b8ceb90e13b3)

Linkte konu anlatımı ve kodları yer almaktadır. Bu bağlantıdaki ödev ve alıştırmaları yapın.

[https://colab.research.google.com/drive/1Kn6Hv8HC1XQncib4AyZ6G2 TZ2OJYWg?](https://colab.research.google.com/drive/1Kn6Hv8HC1XQncib4AyZ6G2__TZ2OJYWg?hl=tr&amp;scrollTo=3gv-obSa8nVw) [hl=tr#scrollTo=3gv-obSa8nVw](https://colab.research.google.com/drive/1Kn6Hv8HC1XQncib4AyZ6G2__TZ2OJYWg?hl=tr&amp;scrollTo=3gv-obSa8nVw)