

**BAŞKENT ÜNİVERSİTESİ**  
TİCARİ BİLİMLER FAKÜLTESİ  
Yönetim Bilişim Sistemleri Bölümü  
TBF 491

## VERİ MADENCİLİĞİ

Dr. Ali Serhan KOYUNCUGİL

**BAŞKENT ÜNİVERSİTESİ**  
TİCARİ BİLİMLER FAKÜLTESİ  
Yönetim Bilişim Sistemleri Bölümü  
TBF 491

## BÖLÜM 8 VERİ MADENCİLİĞİNİN İSTATİSTİKSEL ALTYAPISI

Dr. Ali Serhan KOYUNCUGİL

### BÖLÜM 8 VERİ MADENCİLİĞİNİN İSTATİSTİKSEL ALTYAPISI

- 8.1. Temel Tanım ve Kavramlar
- 8.2. Betimleyici İstatistikler
- 8.3. Değişkenler ve Özellikleri
- 8.4. Değişkenlerin Ölçme Düzeyleri
- 8.5. Hipotez Testleri
- 8.6. Veri Tipi ve İstatistiksel Analiz Yöntemleri
- 8.7. Normal Dağılım
- 8.8. Normallik Analizi

### 8.1. TEMEL TANIM VE KAVRAMLAR

- **YIĞIN (KİTLE, POPULATION):** Araştırma kapsamına alınan, aynı özelliği gösteren bireylerin ya da birimlerin tümünün oluşturduğu topluluğa “kitle ya da yığın” denir.
- **ÖRNEKLEM (SAMPLE) :** Örneklem yöntemlerinden yararlanılarak bir kitleden seçilen, aynı özellikleri taşıyan ve kitleyi temsil edebilecek nitelikte ve nicelikteki bireylerin oluşturduğu topluluğa “örneklem” denir. Yığından örneklem seçmenin amacı kitle ile ilgili bazı bilgileri tahmin etmek ya da kitlelerle ilgili bilgiler konusunda iddiaların doğru olup olmadığını ortaya çıkarmaktır. Yani örneklemdeki bilgileri kullanarak kitle (yığın) ile ilgili bilgilere ulaşmaya çalışırız.

- **DENEK (VARIANTE, SUBJECT):** Yığın ya da örneklemde yer alan her bireye ya da birime “denek” denir. Yığındaki denek sayısı **N** ile örneklemdeki denek sayısı **n** ile gösterilir.
- **PARAMETRE: Yığını ya da kitleyi tanımlayan** sayısal değerlere “parametre” denir. Kitle birimlerinin tümüne ulaşıldığında parametreler hesaplanabilir Kitle ortalaması  $\mu$  (mü), kitle varyansı  $\sigma^2$  (sigma kare) ve kitleye ilişkin korelasyon katsayısı  $\rho$  (ro) birer parametredir.
- **İSTATİSTİK: Örneklemi tanımlayan** sayısal değerlere istatistik denir. Örneğin örneklem ortalaması , örneklem varyansı birer istatistiktir.

**DEĞİŞKEN (VARIABLE):** Gerçek dünyaya ilişkin değerleri, matematik dünyaya bağlayan fonsiyona değişken denir.

**VERİ (DATA):** Kitlede ya da örneklemde yer alan denek değerine veri denir. Veriler araştırma konusundaki denekler ile ilgili değerlerdir.

**KORELASYON:** Korelasyon iki ya da daha çok sayıda değişken arasındaki ilişkiyi gösterir. İlgili miktarı da korelasyon katsayısı ile belirlenir. Korelasyon çözümlemesinin amacı, değişkenler arasındaki ilişkinin derecesini ve yönünü belirlemektir. Korelasyon katsayısının değişim aralığı  $-1$  ile  $+1$  arasındadır.  $-1$  ve  $+1$  e yakın ise ilişki kuvvetlidir, sıfıra yakın ise ilişki zayıftır denir.

## 8.2. BETİMLEYİCİ İSTATİSTİKLER

### 8.2.1. MERKEZİ EĞİLİM ÖLÇÜLERİ

Merkezi eğilim ölçüleri, veri dağılımının nerede toplandığını gösterir.

İncelenecek olan merkezi eğilim ölçüleri:

- Aritmetik ortalama,
- Ortanca (medyan)
- Tepe değeri (mod)

**ARİTMETİK ORTALAMA:** En yoğun kullanılan merkezi eğilim ölçüsüdür. Aritmetik ortalama  $X_i$  değişkenine ait değerler toplamının, gözlem sayısı  $N$  ' e bölünmesi sonucu elde edilir.

$$\mu = \frac{\sum X_i}{N}$$

**ORTANCA (MEDYAN):** Gözlem sayısı tek ise, gözlemler büyükten küçüğe doğru sıralandıktan sonra ortada kalan değer ortanca olarak ifade edilir.

Eğer gözlem sayısı çift sayı ise veriler küçükten büyüğe doğru sıralandıktan sonra ortada kalan iki değer toplanarak ikiye bölünür ve elde edilen değer ortanca değer olarak ifade edilir.

**TEPE DEĞERİ (MOD):** En çok tekrarlanan değer veya kategoridir.

Aritmetik ortalama ve ortanca nitel veriler için uygun olmadığından tepe değeri uygun bir merkezi eğilim ölçüsüdür.

### 8.3. DEĞİŞKENLER VE ÖZELLİKLERİ

**Nitel- Nicel :** Bir niteliği belirtiyorsa buna nitel değişken denir (cinsiyet, göz rengi). Nicelik belirtiyorsa nicel değişken (yaş, boy uzunluğu, kilo) adını alır.

**Kesikli- Sürekli:** Nicel değişkenler tamsayı değerler alıyorsa kesikli değişken, (sahip olunan çocuk sayısı), kesirli ondalıklı değerler alıyorsa sürekli değişken (boy uzunluğu) olarak nitelendirilir.

### **Bağımlı- Bağımsız**

**Bağımsız değişken** (independent variable) değeri rasgele oluşan, ölçüm hatası olamayan, diğer değişken veya değişkenleri etkileyen değişkenlerdir.

**Bağımlı değişken** (dependent variable) ise değeri başka değişkenlerin değişiminden etkilenen onların değişimine göre değer alan değişkenlerdir.

## **8.4. DEĞİŞKENLERİN ÖLÇME DÜZEYLERİ**

### **1. İSİMSSEL (NOMINAL) ÖLÇME DÜZEYİ:**

Değişkenin aldığı değerler bir büyüklük belirtmeksizin sadece bir niteliği, sıfatı, isim vb. bir özelliği gösterir nitelikteyse isimsel ölçme düzeyine sahiptir.

Cinsiyet, kan grubu, göz rengi isimsel ölçekli değişkenlerdir.

İsimsel ölçekli değişkenlere merkezi eğilim ölçülerinden sadece mod kullanılabilir.

### **2. SIRALAMA (ORDINAL) ÖLÇME DÜZEYİ:**

Sıralama ölçme düzeyi isimsel ölçme düzeyine göre biraz daha fazla bilgi sağlayan ve belirli bir hiyerarşiye sahip ölçme düzeyidir. Sıralama ölçme düzeyinde değişkenin aldığı değerler, birbirlerine göre göreceli büyüklük-küçüklük ilişkisine sahiptir.

Kısa boylu – orta boylu – uzun boylu

Çok soğuk – soğuk – sıcak – çok sıcak

### 3. (EŞİT) ARALIKLI (INTERVAL) ÖLÇME DÜZEYİ:

Değişkenlerin aldığı değerler gerçek sayılara karşılık gelmesine karşın, başlangıç değeri (orijini) sıfır ile başlamıyorsa değişkenler aralıklı ölçme düzeyine sahiptir.

Boy uzunluğu  
Kilo

### 4. ORANSAL (PROPORTIONAL) ÖLÇME DÜZEYİ:

Aralıklı ölçeğe sahip değişkenin başlangıç değeri (orijini) sıfır ise değişken Oransal ölçeğe sahiptir.

Sıcaklık  
Hız  
Sınav notları

## 8.5. HİPOTEZ TESTLERİ

$H_0$  hipotezinin karşıt hipotez  $H_1$ 'e karşı (belirli bir hata payına göre) sınanmasını sağlayan yöntemlere İstatistiksel testler (hipotez testleri) denilmektedir.

Hipotez testleri  $H_1$  'in formülasyonuna bağlı olarak ikiye ayrılır:

- Tek yönlü test,
- İki yönlü test.

Eğer,

$H_0: \mu_1 = \mu_2$  ya da  $H_0: \mu_1 - \mu_2 = 0$  biçiminde kurulmuş bir hipotez

$H_1: \mu_1 \neq \mu_2$  ya da  $H_1: \mu_1 - \mu_2 \neq 0$  biçiminde kurulmuş bir karşıt hipoteze

karşı test ediliyor ise bu tür teste iki yönlü test adı verilir.

Eğer,

$H_0: \mu_1 = \mu_2$  ya da  $H_0: \mu_1 - \mu_2 = 0$  biçiminde kurulmuş bir hipotez

$H_1: \mu_1 < \mu_2$  ya da  $H_1: \mu_1 > \mu_2$  biçiminde kurulmuş bir karşıt hipoteze

karşı test ediliyor ise bu tür teste tek yönlü test adı verilir.

Tek yönlü ve iki yönlü testlerden elde edilen test istatistiklerinin (Z, t, F) önemlilik düzeyleri farklı biçimlerde bulunur. Test istatistiğinin uyduğu teorik dağılımların kritik değerlerine göre  $H_0$  red ya da kabul edilerek kararlara ulaşılır.

Test Değeri < Tablo Değeri ise  $H_0$  KABUL (RED EDİLEMEZ)

Test Değeri > Tablo Değeri ise  $H_0$  RED

## 8.6. VERİ TİPİ VE İSTATİSTİKSEL ANALİZ YÖNTEMLERİ

### İsimsel Ölçekli Verilerde Uygulanabilecek analizler:

Frekans (Sıklık) dağılımları alınır  
Her sınıfta görel frekanslar (yüzdel) bulunur  
Ki-kare Uygunluk testi gerçekleştirilir  
Deneyisel olasılıklara göre uygunluk testi yapılabilir.  
Mod belirlenir.

### Sıralı Ölçekli verilerde Uygulanabilecek analizler:

Frekans (Sıklık) dağılımları alınır  
Her sınıfta görel frekanslar (yüzdel) bulunur  
Kikare Uygunluk Testi  
Medyan testi yapılır.  
Rasgelelik için Diziler testi yapılır.  
Mod belirlenir.  
Spearman Korelasyon analizi yapılır

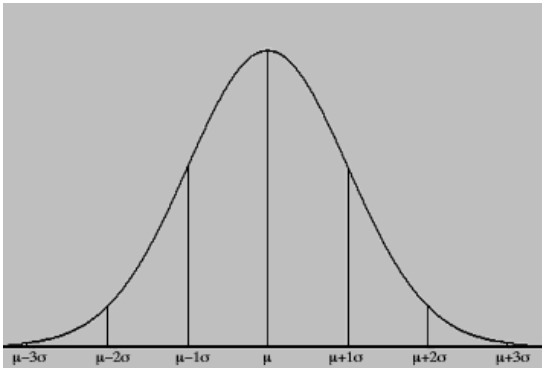
### Aralıklı ve Oransal verilere uygulanabilecek analizler:

Frekans (Sıklık) dağılımları alınır  
Her sınıfta görel frekanslar (yüzdel) bulunur  
Ki-kare Uygunluk testi gerçekleştirilir  
Ortalama, St. Sapma, St. Hata hesaplanır.  
Parametrik testlerle tüm değerlendirmeler yapılır.  
Dağılım varsayımları altında Uygunluk testleri yapılır.  
Tek değişkenli (Univariate) t testi, ANOVA, ANCOVA yapılır.  
Neden-sonuç ilişkileri için doğrusal-eğrisel, Basit/Çoklu Regresyon-Korelasyon Analizleri yapılır.  
MANOVA, MANCOVA ve diğer çok değişkenli analizler yapılabilir.  
(Varsayımları gerçekleştiren veri setlerine)



## 8.7. NORMAL DAĞILIM

- Gauss Dağılımı, Çan Eğrisi gibi isimlerle de anılan Normal Dağılım doğa olaylarının pek çoğunun uyum gösterdiği bir dağılımdır.
- İstatistikte, özellikle de Çok Değişkenli İstatistiksel Yöntemlerde temel varsayımlardan birisi olan Normal Dağılım, pekçok rasgele sürecin dağılımında da gözlenmektedir.
- Pek çok kesikli ve sürekli değişken de bazı varsayımlar altında Normal Dağılıma yakınsamaktadır.



### NORMAL DAĞILIMIN ÖZELLİKLERİ

- Dağılımın parametreleri  $\mu$  ve  $\sigma^2$
- X eksenini ile normal eğri arasında kalan alan 1 birime eşittir.
- Normal dağılım ortalamaya göre simetrik.
- Normal dağılımda tüm birimlerin;
  - %68,26'sı  $\mu \pm 1\sigma$
  - %95,44'ü  $\mu \pm 2\sigma$
  - %99,74'ü  $\mu \pm 3\sigma$aralığında yer alır.
- Aritmetik ortalama, mod ve medyan (ortanca) aynı değerdir.

## 8.8. NORMALLİK ANALİZİ

Elimizdeki verilerin, Normal Dağılım veya başka herhangi bir teorik dağılıma uygunluğunun testi için en yaygın olarak kullanılan testler

- Uyum İyiliği Testleri (Goodness of Fit Tests) olarak ifade edilmektedir.

En yaygın kullanıma sahip Uyum İyiliği Testleri de,

- Kolmogorov-Smirnov Uyum İyiliği Testi,
  - Ki-kare Uyum İyiliği Testi
- olarak ifade edilebilir.