T.C.
AKDENİZ ÜNİVERSİTESİ

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL METOTLAR YARDIMIYLA VERİ ANALİZİNİN DEĞERLENDİRİLMESİ VE BİR UYGULAMA

Esra ŞİMŞEK

FEN BİLİMLERİ ENSTİTÜSÜ

MATEMATİK

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

EYLÜL 2022

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL METOTLAR YARDIMIYLA VERİ ANALİZİNİN DEĞERLENDİRİLMESİ VE BİR UYGULAMA

Esra ŞİMŞEK

FEN BİLİMLERİ ENSTİTÜSÜ

MATEMATİK

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

EYLÜL 2022

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL METOTLAR YARDIMIYLA VERİ ANALİZİNİN DEĞERLENDİRİLMESİ VE BİR UYGULAMA

Esra ŞİMŞEK

MATEMATİK

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

Bu tez T.C. Akdeniz Üniversitesi Bilimsel Araştırma Projeleri (BAP) Koordinasyon Birimi tarafından FYL-2020-5404 nolu proje ile desteklenmiştir.

EYLÜL 2022

T.C.
AKDENİZ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL METOTLAR YARDIMIYLA VERİ ANALİZİNİN DEĞERLENDİRİLMESİ VE BİR UYGULAMA

Esra ŞİMŞEK

MATEMATİK

ANABİLİM DALI

YÜKSEK LİSANS TEZİ

Bu tez 22/09/2022 tarihinde jüri tarafından Oybirliği / ~~Oyçokluğu~~ ile kabul edilmiştir.

Dr. Öğretim Üyesi Füsun YALÇIN(Danışman)
Prof. Dr. Özkan ÖCALAN
Prof. Dr. Murat Alper BAŞARAN

ÖZET

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL METOTLAR YARDIMIYLA VERİ ANALİZİNİN DEĞERLENDİRİLMESİ VE BİR UYGULAMA

Esra ŞİMŞEK

Yüksek Lisans Tezi, Matematik Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Füsun YALÇIN

Eylül 2022; 118 sayfa

Çok değişkenli istatistik yöntemleri birçok disiplinlerde sıkça kullanılmaktadır. Güncel hayatta karşılaşılan problemler ve bilimsel araştırmanın temelleri çoğunlukla çok değişkenlidir. Durum böyle olunca araştırmalar çok değişkenli fonksiyonlar üzerinde yoğunlaşmaktadır. Bu araştırmanın temel amacı, çok değişkenli fonksiyonlar üzerindeki istatistik yöntemlerinin incelenmesidir. Çok değişkenli istatistik metotları olarak bilinen bu metotlar adım adım incelenmiş ve araştırmanın ikinci kısmında örnek bir veri üzerinde uygulama yapılmıştır. Araştırmanın birinci kısmında çok değişkenli istatistiklerin neler olduğundan bahsedilmiştir. Araştırmanın ikinci kısmı uygulama da kullanılan veriler Elazığ'ın Ağın ilçesinde bulunan deri fabrikası etrafından alınmış 32 adet sediman örneklerinin (EADF) kimyasal özelliklerine aittir. Alınan bu sediman örneklerinin verileri Excel programında düzenlenerek SPSS programına aktarılmıştır. Aktarılan bu veriler ile çok değişkenli istatistikler SPSS programında analiz edilmiştir. İlk olarak bu sediman örneklerinin tanımlayıcı istatistiklerine bakılmıştır. Verilere analizlerin uygulanabilmesi için öncelikle Elazığ'ın Ağın ilçesinde bulunan deri fabrikası etrafından alınmış sediman örneklerine ait bu verilerin normal dağılıp dağılmadığı kontrol edildi. Normallik için analizler yapıldıktan sonra bu veriler için çok değişkenli istatistiklere bakılmıştır.

ANAHTAR KELİMELER: Anova, Çok değişkenli istatistik metotlarının varsayımları, Çok değişkenli istatistik metotları, Faktör Analizi

JÜRİ: Dr. Öğr. Üyesi Füsun YALÇIN
Prof. Dr. Özkan ÖCALAN
Prof. Dr. Murat Alper BAŞARAN

ABSTRACT
EVALUATION OF DATA ANALYSIS WITH MULTI VARIABLE STATISTICAL METHODS AND AN APPLICATION

Esra ŞİMŞEK

MSc Thesis in MATHEMATICS

Supervisor: Asst. Prof. Dr. Füsun YALÇIN

September 2022; 118 pages

Multivariate statistical methods are frequently used in many disciplines. The problems encountered in daily life and the foundations of scientific research are mostly multivariate. As such, research focuses on multivariate functions. The main purpose of this research is to examine statistical methods on multivariate functions. These methods, known as multivariate statistical methods, were examined step by step and in the second part of the research, an application was made on a sample data. In the first part of the research, what multivariate statistics are mentioned. The data used in the second part of the research belongs to the chemical properties of 32 sediment samples (EADF) taken from the environment of the leather factory in Ağın district of Elazığ. The data of these sediment samples were arranged in Excel program and transferred to SPSS program. These transferred data and multivariate statistics were analyzed in SPSS program. First, the descriptive statistics of these sediment samples were examined. In order to apply the analysis to the data, firstly, it was checked whether these data belonging to the sediment samples taken from the environment of the leather factory in Ağın district of Elazığ are normally distributed. After normality analysis, multivariate statistics were examined for these data.

KEYWORDS: Anova, Assumptions of multivariate statistical methods, Multivariate statistical methods, Factor analysis

COMMITTEE: Asst. Prof. Dr. Füsun YALÇIN

Prof. Dr. Özkan ÖCALAN

Prof. Dr. Murat Alper BAŞARAN

ÖNSÖZ

Çok değişkenli analiz yöntemi birçok alanda kullanılmaktadır. Çok değişkenli analiz yöntemleri geniş çaplı olmakla birlikte yöntemler farklı farklı alanlardaki verileri inceleyip bu verileri hesaplayarak anlamlılıklarını ortaya çıkarmada kullanılır. Çoğu bilim adamı tarafından kullanarak veriler analiz yapılır.

Bu tez 5 ayrı bölümden oluşmaktadır. Bu bölümler giriş, kaynak taraması, materyal ve metot, bulgular ve tartışma, sonuçlardan oluşmaktadır.

Giriş bölümünde çok değişkenli analiz yönteminin ne demek olduğundan bahsedilmiş ve bazı düşünürlere göre çok değişkenli analiz yönteminin tanımlarına yer verilmiştir. Çok değişkenli analiz yönteminin asıl amacının ne olduğuna yer verilmiştir. Kaynak taraması bölümünde ise tez ile ilgili tanımlara yer verilmiş ve bazı araştırmacıların çok değişkenli analiz yöntemlerini kullanarak yaptığı çalışmalara yer verilmiştir. Materyal ve metot kısmında ise çok değişkenli analiz yöntemleri açık bir şekilde ele alınmıştır. Bulgular kısmında ise bir uygulama yapılmış ve bu uygulamanın bulgu kısımlarına burada yer verilmiştir. Sonuç kısmında ise bulgular kısmındaki analizlerin sonuçları belirlenmiştir.

Çok hevesle başladığım yüksek lisans hikayemin sonuna geldim. Bu süreçte maddi manevi her türlü yanımda olan ve beni her zaman destekleyen anne ve babama sonsuz teşekkürlerimi sunuyorum.

Bu süreçte her zaman destekçim olan tezi yazmam da büyük katkı sağlayan danışman hocam Dr. Öğr. Üyesi Füsun YALÇIN'a sonsuz teşekkürlerimi sunuyorum.

Araştırmada kullanılan verilerin temininde çalışmama yardımcı olan sayın Prof. Dr. Mustafa Gürhan YALÇIN'a sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

AKADEMİK BEYAN

Yüksek Lisans Tezi olarak sunduğum "Çok Değişkenli İstatistiksel Metotlar Yardımıyla Veri Analizinin Değerlendirilmesi ve Bir Uygulama" adlı bu çalışmanın, akademik kurallar ve etik değerlere uygun olarak yazıldığını belirtir, bu tez çalışmasında bana ait olmayan tüm bilgilerin kaynağını gösterdiğimi beyan ederim.

22/09/2022

Esra ŞİMŞEK

SİMGELER VE KISALTMALAR

Simgeler

B	: Gruplar arası kareler toplamları ve çapraz çarpımlar matrisi
H	: hipotez matrisi ve genel kareler ve çapraz çarpım toplam matrisi
Y	: $(n \times 1)$ boyutlu bağımlı değişken vektör
W	: Hata kareler toplamları ve çapraz çarpımlar matrisi
β	: $((p+1) \times 1)$ boyutlu katsayılar vektörü
X	: $(n \times (p+1))$ boyutlu bağımsız değişkenler matrisi
ε	: $(n \times 1)$ boyutlu rasgele hata vektörü
Σ	: $p * p$ boyutlu boyutlu ve p ranklı kovaryans matrisi
x	: $p * 1$ boyutlu gözlem vektörü
μ	: $p * 1$ boyutlu ortalama vektörü

Kısaltmalar

ABA	: Ana bileşenler analizi
CI	: Koşullu endeks sayıları
COV	: Kovaryans
EKK	: En küçük kareler yöntemi
FA	: Faktör analizi
FMANOVA	: Faktöriyel manova
HT	: Hotelling iz kriteri
IYMANOVA	: İki yönlü çok değişkenli varyans analizi
KÇÇT	: Kareler toplamları ve çapraz çarpımlar
TYMANOVA	: Tek yönlü çok değişkenli varyans analizi
VIF	: Varyans artış faktörü

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Gündelik hayatta karşılan problemler çok sayıda etkene maruz kaldığından tek bir değişkenle açıklanamayacak derecede zor ve karmaşıktır, bu yüzden problemleri çözebilmek için problem çok yönlü bir şekilde ele alınmalıdır. Bir problemi etki eden birden fazla faktör bulunmaktadır. Bundan dolayı problem bütün faktörleriyle birlikte incelenir aynı zamanda çözüm önerileri ortaya çıkarılır.

Çok değişkenli istatistik, incelenen olayı ve bu olayın çevresinde bulunan çok sayıda dışsal faktörleri ve içsel faktörleri dikkate alarak problemi asıl oluşumundan yola çıkarak incelemek ve çözüme ulaşabilmek için geliştirilen bir tekniktir (Özdamar 1999). Aynı zamanda çok değişkenli istatistiksel analiz incelenen olayı bütünüyle ele alır ve bu bütünlüğe uygun olan değişkenler için bağımlılık biçimini açıklar. Burada en önemli amaç değişkenlerin arasında bulunan bağımlılık biçiminin analizidir (Tatlıdil, 1996). Diğer bir ifade ile çok değişkenli analiz çok sayıda bağımlı değişken ya da hem bağımlı hem bağımsız değişken fark etmeksizin çok sayıda değişken ile ilgilenildiğinde kullanılan tekniktir (Shin, 1996). Çok değişkenli analiz veri seti üzerinde ikiden fazla değişkeni aynı anda analiz edip inceleyen bir istatistik tekniğidir (Sheth, 1971). Buna istinaden tek değişkenli istatistiksel analiz ise incelenen bu değişkenler üzerinde bulunan dış faktörleri ve iç faktörleri bütün birimler göre sabit veya türdeş kabul eder ve buna göre çözüm oluşturur. Ancak değişkeni incelerken bu değişkenle beraber değişen tüm değişkenleri sabit veya türdeş kabul etmek olanaksızdır. Bu nedenle gerçekçi çözümlere ulaşabilmek için tüm etkilerin çözüme katılması gerekir ve bu amaçla `Çok Değişkenli İstatistiksel Yöntem'lerden yararlanılır.

Çoğu bilim dallarında çok değişkenli hipotezlerin test edilebilmesi için çok değişkenli istatistiksel tekniklerden faydalanılmaktadır. İstatistiğin önemli konularından olan çok değişkenli istatistik yöntemlerin yaygın olarak kullanıldığı bilim alanları; Sağlık Bilimleri, Biyoloji, Ekonomi, Psikoloji, Jeoloji, Kimya, Sosyoloji, Spor, Çevre ve Doğa Bilimleri, Pazarlama, Ziraat, Arkeoloji, Astronomi, Mühendislik, Eğitim vb. şeklinde sıralanabilir. Çok geniş bir alanlara sahip olan çok değişkenli istatistik yöntemler bazı varsayımlara sahiptir. Bu varsayımların içerisinde daha öncelikli olan çoklu normal dağılımdır.

Birkaç düşünüre göre çok değişkenli istatistiksel analiz:

Sheth (1971)'a göre çok değişkenli istatistiksel analiz, bir örnek üstünde ikiden çok değişkeni senkronik yani eş zamanlı olarak çözümleyen bir tekniktir.

Gatty (1966), çok değişkenli analiz tekniklerini değişken kümeleri arasındaki birebir ilişkileri ölçme imkanı ve açıklama imkanı sağlayan bütün istatistik teknikleridir.

Timm (2002)'e göre çok değişkenli istatistiksel teknikler, nesnelerin bağımsız ve bağımlı değişken grupları arasında bir bağlantı kurabilmek için kullanılan tekniklerdir.

Afifi ve Clark (1997)'a göre çok değişkenli istatistiksel analiz, bir birim ya da bir kişi için fazlaca değişkenin edinildiği verilerin analiz kısımlarını açıklamak için kullanılan bir tekniktir.

Shaw (2003)'e göre çok değişkenli istatistiksel analiz, iki ya da ikiden fazla değişkeni birlikte analiz etmeye yarar sağlayan teknikleri tanımlamak için kullanılan terimdir.

2. KAYNAK TARAMASI

Çok değişkenli istatistik farklı disiplinlerde kullanılmaktadır. Araştırmacılar farklı alanlar için farklı çok değişkenli istatistik metotlardan yararlanmıştır. Bazı araştırmacıların inceledikleri konular ve kullandıkları çok değişkenli istatistiksel yöntemler aşağıda verilmiştir.

Ünlükaplan 2008 yılında yaptığı doktora tezinde peyzaj ekolojisini araştırabilmek için çok değişkenli istatistiksel yöntemlerden yararlanmıştır. Burada asıl amaç biyolojik ve fiziksel yapılardan elde edilen verilerin birbiri ile ilişkilendirilmesi ve birimlerin sınıflandırılmasını sağlayan istatistiksel yöntemlerin peyzaj ekolojisiyle bütünleşmesini sağlayabilmektir. Peyzaj ekolojisini araştırılmasında çok değişkenli veri setinin elde edilmesi yöntemi, verilerin güvenirliliği ve birbirleriyle olan ilişkisinin belirlenmesinde Faktör analizi Kümeleme analizi, Kanonik Korelasyon ve Kanonik Uyum analizi tekniği kullanılmıştır. Yapılan araştırmada ayırma yöntemi için kümeleme analizi kullanılmıştır. Atama yöntemi olarak da kanonik uyum analizi, kanonik korelasyon analizi ve faktör analizi kullanılmıştır. Araştırmanın sonucunda peyzaj ekolojisinin araştırmalarında karmaşık veri setlerinin basitleştirilebilmesi ve analizler aracılığıyla karşılıklı olan ilişkilerin belirlenebilmesi için, programlar ve denenen yöntemler etrafında yeterli sonuçlara ulaşılabilmiştir.

Yılmaz (2012) yaptığı çalışmada hastaneler için belirtilen kalite karakteristiklerinin birbirleriyle bağlantılarını olan ilişkilerini göz önünde bulundurarak hastanelerde sunulan işlevin performansını görüntüleyebilmek, başka bir ifade ile sürecin istatistiksel bakımdan kontrol altına alınıp alınmadığını belirlemek eğer kontrol altına alınmamışsa buna sebep olan kalite karakteristiğini belirlemeyi amaçlamıştır. Yapılan bu çalışmada, bir devlet hastanesin bulunan müşterilerin memnuniyeti düzeylerini ölçmek ve yoğun bakım ünitesinin performansını görüntüleyebilmek için çok değişkenli istatistiksel süreç tekniklerinden faydalanılmaktadır. Bundan yola çıkarak, Hotelling T2 kontrol grafiği ve kontrol dışında olduğu bilinen birimlerin sebeplerini belirleyebilmek için Mason-Young-Tracy (MYT) Ayrıştırması tekniği kullanılmıştır. Yapılan çalışmanın sonuçlarına bakıldığında, çok değişkenli Hotelling T2 kontrol grafiğinin hastanenin performans düzeyini belirlemede başarılı bir teknik olduğu belirtilmiş ve MYT Ayrıştırması tekniği yardımıyla da kontrol dışı duruma

sebep olan değişkenler belirlenmiştir.

Çelik (2004) yaptığı bu çalışmada sigarayla ilgili incelediği çok değişkenli yapıyı kümeleme yöntemi ile çözümlemeyi amaçlamıştır. Bu çalışmada sigarayla ilgili birimleri inceleyebilmek için öğretmenler, polisler, öğrenciler, işsizler, sağlık personelleri ve esnaflar olmak üzere toplamda 1133 kişiye ait 64 değişkeni incelemiştir. İncelenen bu değişkenlere ait ilginç gözleme oranlarıyla birlikte küme yapısını ortaya koyan ward kümeleme yöntemini kullanmıştır. Bu altı kümeye ait olan verileri dikkate alarak değişkenler ile ilgili dendogramlar belirlemiştir. Belirlenen dendogramlar aracılığıyla sigara ile ilgili birimlerden bireylerin cinsiyetini, sosyal ve psikolojik durumlarını, değişik alışkanlıklarını belirten kümeler elde etmiştir. Sonuç olarak sigara ile ilgili değişkenler için uygulanan ward kümeleme yönteminin, konuya dair muhtemel olan kümelerden başarılı olduğu ve anlamlı bir sonuç elde etmiştir.

Yılmaz (2009) tarafından yapılan çalışma Türkiye genelinde 19 su toplama havzasından alınmış, EİE İdaresi Genel Müdürlüğü aracılığıyla verimlenen 67 adet Akım gözlem istasyonuna dair 1992-2008 yıllarında olan hesaplanmış Akım ve 11 tane su kalitesi parametresi olacak şekilde toplam 12 tane parametrenin uzun yıllardaki ortalama değerlerine çok değişkenli istatistiksel yöntemler aracılığıyla bakılmıştır. Bu işlem için Kümeleme Analizi, Ana Bileşenler Analizi ve Faktör Analizi kullanılmıştır.

Blinstrub'un (2002) tarafından yapılan bu çalışma için Kaliforniya'da Orange Country Havzasında bulunan 31 tane istasyondan alınmış su kalitesi değerleri üzerine Ana Bileşenler Analizini uygulamış, yaptığı analizin sonunda toplam varyansın %77,6 `sını açıklayan 5 tane ana bileşeni belirlemiş, ortalama Ana Bileşenler Analizi sayılarına göre yaptığı Kümeleme Analizinde ise aynı su kalitesinin görüldüğü 6 adet küme belirlemiştir.

Tekin (2015) tarafından yapılan çalışmada Türkiye' de bulunan illeri temel sağlık göstergeleri bakımından sınıflandırılmıştır. Sınıflandırma için veri madenciliği ve kümeleme analizini tercih etmiştir. Çalışmadaki asıl amaç, illeri temel sağlık göstergeleri bakımından incelenmek ve inceleme sonucunda benzer ve farklı il guruplarının karşılaştırılması ve birlikte gösterilmesidir. Buradan yola çıkarak 2013 yılına dair 16 değişik sağlık göstergesiyle Türkiye' de bulunan 81 ili kümelendirmiş ve

önceden yapılmış olan sosyoekonomik gelişmişlik düzenlemeleriyle ve sağlık gelişmişlik düzeyin de yapılan çalışmalar ile karşılaştırmıştır. Ward kümeleme yöntemi aracılığıyla illeri 21, 7, 5, 13 ve 11 kümede birleştirmiştir. Bu kümelerden en anlamlılarının, 7' li, 5' li ve 11' li kümeler olduğunu belirlemiştir. Sonuç olarak, genele bakıldığında, doğu ve batı illeri arasında oluşan gelişmişlik düzeyi farkı sağlık değişkenleri tarafından belirlenmiştir.

Arslan 2008 yılında yaptığı`Su Kalitesi Verilerinin CBS ile Çok Değişkenli İstatistik Analizi (porsuk çayırı örneği)' adlı makalesindeki asıl amaç su kalitesi verilerinin değerlendirilmesi için CBS tabanlı olan çok değişkenli istatistik analizinin potansiyelini araştırmaktır. CBS teknolojisiyle istatistik yöntemler su kalitesi uygulamasında aktif biçimde kullanılmaktadır. Su kalitesi veri setlerinin karışık bir sisteme sahip olmasından dolayı tek değişkenli istatistik yöntemleri kısıtlayıcı olabileceği için su kalitesinin değerlendirilmesinde yapılan çalışmalarda çok değişkenli istatistik analizi yöntemleri tercih edilmektedir. CBS ortamında çok değişkenli istatistik analiz yönteminin gerçekleştirilmesi için yazılmış makro kodları kullanılarak Ana bileşenler Analizi hazırlanmıştır. Porsuk Çayı üzerinde bulunan gözlem istasyonlarından elde ettiği veriler ile oluşturduğu su kalitesi veri tabanını, belirtilen ara yüz aracılığı ile analiz etmiş, kalite değişiminde etkili olan değişkenleri belirleyip yorumlayarak yöntemin potansiyelini ve etkinliğini göstermiştir. Sonuç olarak bahsedilen çok değişkenli istatistik yöntemin CBS tabanlı uygulamasının su kalitesi analizini geliştirdiği söylenebilir.

Bulut (2014) tarafından yapılan bu çalışmada bölgelerin arasındaki gelişmişlik farklarının azaltılması nedeniyle hizmet veren kalkınma ajansları kapsamında olan illerin sosyo-ekonomik gelişmişliklerinin değerlendirilmesi nedeniyle klasik ve robust temel bileşenler analizi tekniklerini kullanmıştır. Robust yöntemlerinin başlıca kullanılma sebebi, bölgeler arasındaki gelişmişlik farklarının büyük olmasından kaynaklanan aykırı değer problemidir. Veri setinde aykırı değer olup olmadığının belirlenmesi için klasik ve robust istatistiksel yaklaşımlardan yararlanılmıştır. Klasik temel bileşenler analizindeki değişken sayısı gözlem sayısından daha az olması gerekmektedir. Aksi halde kovaryans matrisinin determinantı sıfır olur. Böylece klasik yaklaşımda değişken sayısı en fazla gözlem sayısının bir eksiği kadar olabilir. Robust

temel bileşenler analizi (ROBPCA) tekniği ile değişken sayısı gözlem sayısından fazla olmuş olsa da böyle bir durumda yine temel bileşenler analizi uygulanabilmektedir. Bu uygulamada 26 kalkınma ajans bölgesini ilk önce 19 değişken bakımından klasik ve robust korelasyon matrislerine dayanan temel bileşenler analizi ile değerlendirmiş, sonrasın da ise 46 değişken bakımından ROBPCA tekniği ile değerlendirmiştir.

Demirci'nin (2017) yapmış olduğu bu yüksek lisans tezi yedi bölümden oluşmaktadır. Yedi bölümden oluşan bu yüksek lisans tezinde geometrinin istatistikte kullanımına dair yöntemleri vermiştir.

Toktay 2017'de yaptığı çalışmasında üniversitede okuyan öğrencilerin sorunlarını ya da sorun olarak gördükleri olguları belirlemek amacıyla öğrencilerin demografik bilgilerinin yanı sıra, eğitim memnuniyeti ile ilgili konuları da ele almıştır. Elde ettiği verileri inceleyerek sorunları belirlemek, belirlenen bu sorunlara çözüm yolu ve öneriler geliştirmeyi amaçlamıştır. Bu çalışması için bir likert ölçek geliştirerek bu likert ölçeği çok değişkenli istatistik analiz yöntemlerinden olan diskriminant ve faktör analizine uygulamıştır. Bu çalışmada yapılan faktör analizini 20 likert tipi soru ve 5 faktör şeklinde ifade etmiştir. Öğrencilerin okul memnuniyeti, eğitim, öğretim ders ve şehir hakkındaki görüşlerini barındıran likert tarzı sorular Üniversiteye yerleştikleri puan türü, ulaşım ve barınma tipine göre uygulana diskriminant analizi için zorunlu olan normallik hipotezleri nedeniyle uygulanmıştır.

Pamuk (2005) tarafından yapılan bu makaledeki amaç, bir öğretim üyesinin verimliliğine öğrencilerin görüşlerini dikkate alarak faktörlerin sayısını ve yapısını belirleyebilmektir. Çalışmada kullanılan veriler 2003-2004 eğitim-öğretim yılında İstanbul Üniversitesi İktisat Fakültesinde bulunan 193 lisans öğrencisine dair "Öğretim Üyesi Değerlendirme Anketi"nin sonuçlarını kapsamaktadır. Anket içerisinde bulunan 18 soruyu 1'den 5'e kadar numaralandırmıştır.

Jama Abdı (2017) tarafından yapılan tez, Türkiye'deki üniversitelerde bulunan 18-37 yaş aralığındaki öğrencilere duygusal bakımdan cazip olan, web sitesi tasarımı öğelerini belirleme de Kısmi En Küçük Kareler (PLS) regresyon istatistik tekniklerini, Faktör Analizi (FA) ve Kansei Mühendisliği'ni önermektedir. Web sitesinin dizaynın da kullanıcının memnuniyet bakımından bu çalışmadaki Faktör analizi, Kısmi En Küçük

Kareler regresyon ve Kansei Metodolojisinin önemli olduğunu bu tezde yapılan sonuç ortaya çıkarmaktadır.

Aldemir (2019) yaptığı bu araştırmada, çok değişkenli yöntemlerden olan faktör analizi, çok boyutlu ölçekleme analizi ve kümeleme analizi gibi analizlere teorik açıdan yer vermiştir. Aynı zamanda, TÜİK'den aldığı 2018 yılına ait hayvancılık istatistiklerinden toplam 29 değişkeni bulunduran hayvansal üretim ve hayvan varlığı verilerini çok değişkenli analiz yöntemleri aracılığı ile incelemiştir. Yaptığı bu çalışmasında öncelik olarak kümeleme analizine yer vermiş, en iyi olan küme yapısını belirlemek için hiyerarşik kümeleme yöntemlerinden Ward yöntemini kullanmış ve illeri 2 gruba ayırmıştır. Hiyerarşik olmayan kümeleme analizlerinde en iyi küme sayısı, birtakım parametrelerin yanı sıra küme üyelik kodları aracılığı ile diskriminant (ayırma) analizi yardımıyla belirlemiş olup, fuzzy (bulanık) kümeleme tekniğinin 5 küme yapısında en etkili teknik olduğu neticesine varmıştır. Hatta, çalışmada faktör analizinden elde ettiği faktör yüklerini kullanarak kişi başına düşen hayvan varlığına ve hayvansal üretime ait ham veriler ve standartlaştırılmış veriler aracılığı ile illerin hayvancılık gelişmişlik endekslerini hesaplamıştır. Her iki sıralama içinde ilk sırada yer alan ilin Ardahan son sırada yer alan ilin İstanbul ili olduğunu görmüştür. Son olarak, illerin hayvancılık istatistiklerine göre grafiksel olarak bir değerlendirme elde edilebilmesi amacıyla çok değişkenli istatistik yöntemlerinden metrik çok boyutlu ölçekleme tekniğini kullanmıştır.

Köksal'ın (2019) yaptığı bu çalışmasındaki temel amacı, ülkemizin eğitim sisteminde oldukça büyük bir öneme sahip olan matematik dersine ait üniversite öğrencilerinin kaygı seviyesini ve bu kaygının nedenlerini araştırmak ve bu araştırma ile beraber öğrencilerin matematiksel düşünme seviyelerinin matematik kaygısı üzerindeki etkisine bakarak matematik kaygısının azaltılabilmesi için yol gösterici öneriler sunmaktadır. Öte yandan, öğrencilerin matematiksel düşünme seviyelerini ölçerek öğrencilerin matematiksel düşünme seviyelerinin geliştirilebilmesi, matematiksel düşünmenin önemine dikkat çekilmesini hedeflemiştir.

Hüyüktepe (2018) bu çalışmasında 2017 yılına kadar açıklanmış olan yıllık verilere göre Türkiye de bulanan illerin sosyo-ekonomik gelişmişlik durumlarını çok değişkenli istatistiksel teknikler aracılığıyla incelemiştir.

Akın ve Çörek (2005) Çok değişkenli istatistiksel tekniklerin performans kıstaslarının incelenmesinde uygulanması, müşteri beklentilerini anlamasına yardımcı olduğu ve kuruluşun hizmet sürecini iyileştirmesine yardımcı olmasından dolayı müşteri memnuniyetinde istatistiksel yöntemler adı altında bu çalışmayı yapmışlardır. Bu çalışmayı yürütürken çok değişkenli tekniklerinden olan çok boyutlu ölçekleme analizi faktör analizi, kümeleme analizi, t testi, çoklu belirlilik katsayısının anlamlılık analizi ve regresyon analizini uygulamışlardır. Müşteri isteklerinin ve beklentilerinin sağlanması kurallarına dayanarak müşteri memnuniyetinin ölçülebilmesi zamanla günümüzde oldukça önem kazanmıştır. Müşterinin ürün ve hizmet hakkındaki görüşlerinin belirlenebilmesinde en çok kullanılan yöntem anket çalışmalarıdır. Araştırmacılarda bu çalışmalarında anket çalışmasını uygulayarak gerekli olan bilgileri elde etmişler ve bu bilgileri çok değişkenli analiz yöntemleri ile incelemişler.

Sağır 2020, Tatlı 2015 yılında yaptıkları çalışmalarda çok değişkenli istatistik tekniklerini kullanarak farklı disiplinlerde araştırmalar yapmışlardır.

Tezde ele alınan yöntemlerin işleyişinde kullanılan tanımlar ve formüller aşağıda verilmiştir;

Tanım 2.1 Her bir değişken için elde edilen ortalamalar bir vektörde gösterilir. Bu vektöre ortalama vektör denir. Ortalama vektör şu şekilde gösterilir;

$$\bar{X} = \begin{bmatrix} \overline{X_1} \\ \overline{X_2} \\ \overline{X_3} \\ \vdots \\ \overline{X_n} \end{bmatrix}$$

Tanım 2.2 Bir veri seti içerisinde bulunan değerlerin aritmetik ortalamadan ortalama ne derecede uzaklaştığını belirten merkezi dağılım ölçüsüne varyans adı verilmektedir. Varyansın amacı verilerin dağılımıyla ilgili bilgi vermektir. Her bir değerin aritmetik ortalamadan farkının kareleri toplamı alınarak veri setinin eleman sayısına bölünmesiyle varyans değeri hesaplanır.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Tanım 2.3 Bir veri seti içerisinde merkezi dağılım ölçüleri olan medyan mod ve aritmetik ortalamanın eşit olması durumuna normal dağılım adı verilir. Aşağı yukarı uzaklaştıkça frekansların belirli bir diziliş içerisinde azalması ile ortaya çıkan bir dağılımdır.

Tanım 2.4 1) Dağılımın düzlüğü ya da dikliği yani verilerin tepe noktaları hakkında bilgi sağlayan ölçüte basıklık denir. Sıfıra yakın basıklık normal dağılıma daha yakındır. Basıklık için pozitif değer normal dağılımdan daha dik bir dağılıma, negatif basıklık değeri normal dağılımdan daz düz bir dağılıma işaret eder.

2) Dağılımın ortalama çevresinde simetriden ne derecede saptığını belirten ölçüte çarpıklık denir. Sıfır çarpıklık değeri simetrik bir dağılıma işaret eder. Eğer çarpıklık pozitif ise küçük değerlerin fazla olduğunu, çarpıklık negatif ise büyük değerlerin fazla olduğu görülür. Veri setinde eğer ortalama medyandan büyük ise sağa çarpık dağılım, eğer ortalama medyandan küçük ise sola çarpık bir dağılım vardır.

Tanım 2.5 Parametrenin tahmini için kullanımı gerekli olan bağımsız bilgi parçaları sayısına serbestlik derecesi denir. İstatistikte bir istatistiğin kesin olarak belirlenmesi için kullanılan değerlerin sayısının ne kadar serbestlikte değişmesi gerektiğini sayısal olarak verir. İstatistik parametrelerin tahminleri farklı nicelikte bilgiye ya da veriye dayandırılır.

Tanım 2.6 Bilimsel yöntemlerde olayların arasındaki ilişkileri kurmak ve olayları belirli nedene bağlamak için geliştirilen ve geçerli sayılan önermeye hipotez adı verilir.

Tanım 2.7 K cismi üzerinde bir matris a_{ij} skalalarının dikdörtgen bir tablosu olarak aşağıdaki gibidir.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Matris (a_{ij}), $i = 1,2,\dots,m$, $j = 1,2,\dots,n$ veya kısaca (a_{ij}) biçiminde gösterilir. Yatay m tane $(a_{11}, a_{12}, \dots, a_{1,n})$, $(a_{21}, a_{22}, \dots, a_{2n})$,..., $(a_{m1}, a_{m2}, \dots, a_{mn})$ n'lileri matrisin satırlarıdır. n tane, dikey m'liler

$$\begin{bmatrix} a_{11} \\ a_{21} \\ \dots \\ a_{m1} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \\ \dots \\ a_{m2} \end{bmatrix}, \dots, \begin{bmatrix} a_{1n} \\ a_{2n} \\ \dots \\ a_{mn} \end{bmatrix}$$

matrisin kolonları (sütun) dır. a_{ij} elemanı ij-girdi ij-bileşeni şeklinde adlandırılır ve a_{ij}, i. satır ve j. kolon şeklinde görülür. m satır, n kolon içeren matris $m \times n$ matris biçiminde adlandırılır. (m, n) sayı ikilileri matrisin boyutu ya da tipi olarak adlandırılır. Matrisler genelde A,B,..., şeklinde büyük harfler ile gösterilir. K cisminin elemanları ise a,b, şeklinde küçük harfler ile gösterilmektedir. A ve B matrislerinin karşılıklı elemanları birbirine eşit ise A ve B matrislerine eşittir denir ve $A = B$ biçiminde yazılır (Hacısalihoğlu 1991).

Tanım 2.8 Bir C matrisinin transpozesine eşit $\left(C = C^{\{T\}}\right)$ olmasına simetrik kare matris adı verilir. Elemanları arasında $|c_{ij}| = |c_{ji}|$ ilişkisi bulunan fakat $c_{ij} = c_{ji}$ ilişkisi bulunmayan $\left(c_{ij} \neq c_{ji}\right)$ matrise çarpık simetrik matris denir. Çarpık simetrik matrisin özelliği $A = -A^T$ dir.

Tanım 2.9 Veri matrisi n tane birimden ve p sayıda değişkenden oluşan bir gösterimdir. $X\ nxp$ boyutlu bir veri matrisi olmak üzere aşağıdaki gibi gösterilir.

$$X_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

Veri matrislerindeki her sütun bir değişkeni ifade etmektedir. Vektör elemanları toplamanın eleman sayısına bölünmesi ile bir vektörün ortalaması hesaplanmaktadır. Matrislerde her bir sütunun toplamının sıra sayısına bölünmesi ile ortalama vektörü hesaplanmaktadır. Bir veri matrisinde her bir sütun bir değişkeni belirttiğinden değişkenler için istatistikler hesaplamak mümkün olmaktadır. X veri matrisinde j. değişkenin ortalaması X_j ve varyansı $V\left(X_j\right)$ olmak üzere aşağıdaki gibi hesaplanmaktadır.

$$\overline{X_j} = \frac{\sum_{i=1}^{n} X_{ij}}{n}$$

$$V(X_j) = \left(\frac{\sum_{i=1}^{n} X^2{}_{ij} - (\sum_{i=1}^{n} X_{ij})^2/n}{n-1} \right)$$

Her bir değişken için hesaplanan ortalamalar bir vektör halinde gösterilir ise bu vektöre ortalama vektörü adı verilmektedir. X veri matrisinin ortalama vektörü aşağıdaki gibidir.

$\bar{X} = \begin{bmatrix} \overline{X_1} \\ \overline{X_2} \\ \vdots \\ \overline{X_n} \end{bmatrix}$ veya satır vektörü olarak $\overline{X'} = [\overline{X_1} \quad \overline{X_2} \quad \dots \quad \overline{X_n}]$ biçiminde gösterilir (Özdamar 2018).

Tanım 2.10 Bazı istatistiksel analizlerde değişkenlerin düzeltilmemiş kareler toplamları ve çapraz çarpımlar (KÇÇT) matrisinden faydalanılır. KÇÇT matrisinde bulunan KT (kareler toplamı) ve ÇÇT (çapraz çarpımlar toplamı) ölçütleri aşağıdaki şekilde hesaplanır:

$$KT_{X_iX_i} = S_{ii} = \sum_{i=1}^{n} X_i^2 - \left[\frac{(\sum X_i)^2}{n} \right]$$

$$\text{ÇÇ}T_{X_iX_j} = S_{ij} = \sum_{i,j=1}^{n} X_i X_j - \left[\frac{(\sum X_i)(\sum X_j)}{n} \right]$$

Bulunan değerler Kareler ve Çapraz Çarpımlar Toplamı matrisinde yerine konularak

$$K\text{ÇÇ}T(X) = SSCP(X) = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}$$

şeklinde ifade edilir (Özdamar 2018).

Tanım 2.11 Veri matrisinde bulunan değişkenlerin birlikte değişimlerini ve varyanslarını gösteren matrise kovaryans matrisi denir. Kovaryans matrisinin kaç birimden oluştuğunu belirlemek için $S_n(X)$ gösterimi veya popülasyon kovaryans matrisi ise $S_N(X)$ gösterimi kullanılır. Veri matrislerinden kovaryans matrisini elde

edebilmek için değişkenlerin kareler toplamları, sütun toplamları ve çarpımlar toplamları belirlenmelidir. Elemanlar s_{ij} veya σ_{ij} olarak gösterilir.

$$S_{p\times p} = S_n(X) = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

$$\Sigma = S_N(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

KÇÇT matrisinin elemanlarının $(n-1)$'e bölünmesiyle kovaryans matrisi elde edilir. Bir X veri matrisinin kovaryans matris elemanları aşağıdaki şekilde hesaplanır;

eğer i=j ise $s_{ii} = \frac{s_{ii}}{n-1} = \left(\frac{\sum {X^2}_i - (\sum X_i)^2/n}{n-1}\right)$

eğer i≠j ise $s_{ij} = \frac{s_{ij}}{n-1} = \left(\frac{\sum X_i X_j - (\sum X_i)(\sum X_j)/n}{n-1}\right)$

Tanım 2.12 Korelasyon matrisleri $p * p$ boyutlu simetrik kare matristir. Korelasyon matrisinin ana köşegen elemanları 1 dir. Ancak ana köşegen dışındaki elemanlar -1 ile $+1$ arasında değerler almaktadır $(-1 \le r_{ij} \le +1)$. İki değişken arasındaki ilişki korelasyon katsayısı sıfıra (0) yaklaştıkça azalmakta, ∓ 1'e yaklaştıkça artmaktadır. Korelasyon katsayısı arasındaki ilişki $+1$ çıkarsa değişkenler arasında tam pozitif bir ilişki bulunmaktadır. Eğer korelasyon katsayısı arasındaki ilişki -1 çıkar ise değişkenlerin arasında tam negatif bir ilişki var demektir. Korelasyon matrisi n birimden oluşan verilere göre R_n veya $R_{p\times}$ biçiminde gösterilir. Büyük popülasyonlar için ρ (rho) ile gösterilmektedir. R_n korelasyon matrisi aşağıdaki gibidir.

$$R_n = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & s & \dots & 1 \\ & r_{p2} & & \end{bmatrix}$$

Korelasyon matrisinin köşegen dışı elemanları olan i ve k değişkenlerinin arasındaki korelasyon katsayıları aşağıdaki şekilde hesaplanır.

$$r_{ik} = \left(\frac{\sum_{j=1}^{n}(X_{ij} - \overline{X_i})(X_{ik} - \overline{X_k})}{\sqrt{\left[\sum_{j=1}^{n}(X_{ij} - \overline{X_i})^2\right]\left[\sum_{j=1}^{n}(X_{ik} - \overline{X_k})^2\right]}} \right) = \frac{\text{ÇÇ}T_{X_i X_k}}{\sqrt{KT_{X_i} \times KT_{X_k}}}$$

KÇÇT(X) ve $S_n(X)$ elmanlarından faylanılarakta korelasyon matrisi hesaplanabilmektedir.

KÇÇT matrisi elemanlarından faydalanarak R matrisi aşağıdaki gibi hesaplanır.

$$r_{ik} = \left(\frac{S_{ik}}{\sqrt{S_{ii} \times S_{kk}}} \right)$$

(Özdamar 2018)

Tanım 2.13 Çok değişkenli istatistiksel yöntemlerden bazıları değişkenler veya birimler arasındaki çok boyutlu uzaydaki uzaklıklardan faydalanarak analiz yapabilmektedir. Bu uzaklıkları içeren matrislere uzaklık matrisi denir. Uzaklık matrisinden faydalanılarak; benzerlik farklılık veya yakınlık matrisleri hesaplanabilmektedir. Değişkenler veya birimler arası uzaklıklar öklid uzaklığı veya Mahalanobis uzaklığı olarak hesaplanabilmektedir. p değişkenli $A1$ vektörü $A1 = [x_1, x_2, \ldots, x_p]$; $0 = [0,0,\ldots,0]$ ölçümlerine ait bir noktanın orjine olan uzaklığı aşağıdaki gibi hesaplanır;

$$d(0, A1) = \sqrt{({x_1}^2 + {x_2}^2 + \cdots + {x_p}^2)}$$

Değişkenler arasındaki uzaklığın karesel uzaklık şeklinde hesaplanması sonucunda Karesel Öklid uzaklığı bulunur. Karesel Öklid uzaklığı aşağıdaki gibi hesaplanır.

$$d^2(0, A1) = {x_1}^2 + {x_2}^2 + \cdots + {x_p}^2$$

Çok boyutlu bir uzayda b ve c vektörleri arasındaki uzaklık Pisagor bağıntısı yardımı ile aşağıdaki gibi hesaplanır.

$$b = [x_1, x_2, \ldots, x_p\}]$$

$$c = [y_1, y_2, \ldots, y_p]$$

$$d(b,c) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2)}$$

Bu uzaklıklar değişkenlerin değişim aralığından etkilenmektedir. Bu etkilenmeyi ortadan kaldırabilmek için uzaklıklar değişkenlerin standart sapmalarına göre standardize edilmelidir. Bu standardize işlemi için kovaryans matrisinin ana köşegen elemanlarından faydalanılır.

$A1$ noktasının x_1 ve x_2 ölçümü aşağıdaki gibi standardize edilir.

$$x^*_1 = \frac{x_1}{\sqrt{S_{11}}}$$

$$x^*_2 = \frac{x_2}{\sqrt{S_{22}}}$$

Tanım 2.14 F testi, iki varyans için tercih edilen bir test türüdür. Test istatistiği,

$$F = \frac{S^2_1}{S^2_2} \ , S^2_1 < S^2_2$$

şeklindedir ve hipotezler ise aşağıdaki gibidir;

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Tanım 2.15 X bir rasgele değişken,

a) X kesikli rasgele değişkeni için

$$E(X) = \sum X_i P_i$$

b) X sürekli rasgele değişkeni için

$$E(X) = \int_{-\infty}^{\infty} X_i f(X_i)$$

X'in beklenen değeri denir.

Tanım 2.16 Örneklem ortalamasının standart sapmasına standart hata denir.

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Tanım 2.17 Nominal ölçek araştırmayı konu edinen nitelikleri, benzerlik veya farklılıklarına göre ayırmak için kullanılan ölçek türüdür. Bu ölçeklerin herhangi bir başlangıç noktası ve birimi yoktur. Bundan dolayı matematiksel işlem yapılamaz. Yalnızca saymaya bağlı olan frekans, mod ve yüzde oranları bulunur (Anonim 1).

Tanım 2.18 Veri setinde uç değerlere bakmak için iki önemli neden vardır:

1. Uç değer normal sonuçlar elde etmemizi önleyeceği için, belirlenerek veri setinden çıkartılmalıdır.

2. Bunlar aynı zamanda bir bilgi kaynağıdır. Bu uç değerler belirlenerek nedenleri açıklanmalıdır.

Tanım 2.19

1) İstatistikte, iki değişkenin birlikte ne kadar değiştiğini gösteren ifadeye kovaryans denir. Kovaryans, iki rasgele değişkenin birlikte değişimlerini inceleyen istatistiktir (Anonim 2).

2) İstatistikte, iki rassal değişkenin arasında bulunan doğrusal ilişkinin yönünü ve gücünü belirten ifadeye korelasyon denir. Korelasyon, bağımsızlık durumundan ne derece uzaklaşıldığını belirtir (Anonim 3).

3) Varsayım, araştırmanın en başında, doğruluğu imtihana elverişli olmayan kanıtlanmaya gerek olmayan, doğru biçimde kabul edilmesi gereken, denemesi yapılmayan yargılardır (Anonim 4).

Tanım 2.20 Gözlemlerin ortalamadan ne derece uzaklaştığını ve varyansın pozitif karekökünе eşit olan ifadeye standart sapma denir.

Tanım 2.21 Robust istatistiği, istatistikte yaygın biçimde kullanılan doğrusallık ve

normallik gibi varsayımların tahminleri ile ilgilenen bilim dalıdır (Anonim 5).

Tanım 2.22 Testin gücü, hipotez testlerinde genellikle görmezden gelinen beta hatasında yararlanılarak hesaplanan analizdir. İstatistiksel açıdan karar aşamasındayken verilecek kararın doğruluğu ve güvenirliliğini belirlemede kullanıldığı gibi örneklemin büyüklüğünün belirlenmesinde de kullanılır (Anonim 6).

Tanım 2.23 Hata terimi (ε) modelin stokastik olduğunu belirtir ve modele ait olmayan değişkenleri içerir (Yavuz 2009).

Tanım 2.24

1) Sürekli değişkene ait veri setinden oluşan sınıflarda ki frekansların dağılımını göstermek için kullanılan grafiklere histogram denir.

2) Kategori ya da sınıflara ait olan miktarlar arasında oluşan ilişkiyi göstermek için x eksenine sınıfı, y eksenine mutlak miktarları yerleştirilerek uygulanan grafiklere sütun grafikleri denir.

3) Bir bütünden oluşan parçaların ifade edilmesinde kullanılan grafiklere daire grafiği denir.

4) Verilerin dağılımını göstermenin yanı sıra birçok istatistik ölçüyü belirtmede kullanılan grafiklere kutu grafiği denir. Minimum, maksimum, çeyrek ölçüler, medyan, sapma değerleri gibi ölçüler kutu grafiğinde gösterilir.

3. MATERYAL VE METOT

3.1. Çok Değişkenli İstatistiksel Analiz Yöntemleri

Belirlenmiş veri seti için çok değişkenli tekniklerden hangisinin kullanılabileceğinin belirlenmesi, mevzu bahis olan yöntemler arasındaki yakın ilişkiden dolayı oldukça zordur. Uygun olan çözüm yönteminin belirlenmesi için, araştırmacı analiz yapmak istediği verilerin içindeki değişkenlerin karşılıklı ilişkili veya bağımlı olup olmadığına karar verilir ve eğer bağımlılık var ise bağımlı değişkenin sayısını belirler. Aynı zamanda değişkenlerin hangi metrik veya metrik olmayan ölçek ile ölçüldüğü belirlenmelidir (Hair vd. 1998).

Literatürde yaygın olarak kullanılan çok değişkenli istatistiksel analiz teknikleri aşağıdaki gibi verilebilir:

1- Çok Değişkenli Hipotezlerin Testi (Hotelling T^2 Testi): Çok değişkenli normal dağılım varsayımı için iki ve ikiden fazla değişkenli ($p \geq 2$) tek ve iki örnek hipotezlerinin test edilebilmesi için kullanılan bir tekniktir (Özdamar 1999)

2- Çok Değişkenli Varyans Analizi (MANOVA, Multivariate Analysis Of Variance): İki veya ikiden fazla çok değişkenli ($p \geq 2$) normal dağılım gösteren veri setleri için kurulmuş olan hipotezleri test etmede kullanılan bir tekniktir. İki ve tek yönlü deneme sonuçlarıyla çok faktörlü ve çok değişkenli deneme sonuçlarının analizinde kullanılmaktadır.

3- Çok Değişkenli Kovaryans Analizi (MANCOVA, Multivariate Analysis Of Covariance): İki veya ikiden fazla çok değişkenli ($p \geq 2$) normal dağılım gösteren veri setlerine göre kurulmuş olan hipotezlerin ortak değişkenlerinde bulunduğu veri setinin test edilmesinde kullanılmaktadır.

4- Çok Değişkenli Regresyon Analizi (Multivariate Regression Analysis): İki bağımlı, bir veya daha çok bağımsız değişkenin arasında bulunan neden-sonuç ilişkilerini belirlemek amacıyla kullanılan bir tekniktir.

5- Ana bileşenler Analizi (Temel Bileşenler Analizi): Birbirleri ile ilişkili iki veya ikiden fazla değişkeni kapsayan veri matrislerinden, birbirleriyle bağımsız olan ve çok az sayıda yeni veriler bulmak için kullanılan bir tekniktir.

6- Ayırma Analizi (Discriminant Analysis): Veri setinde birbirleri ile ilişkili iki veya daha fazla yeni alınan değişkenlerin özelliklerine uygun olan gruplara ayırmak için kullanılan bir tekniktir.

7- Faktör Analizi (Factor Analysis): Çok sayıda değişkenden oluşmuş veri setini daha az sayıda anlamlı ve birbirinden bağımsız faktör sistemlerine dönüştürmek, özgün değişkenlere göre açıklanmayan yeni faktör sistemi haline getirmek, öne sürülen gizil yapısal modelleri belirlemek için kullanılan bir tekniktir.

8-Uyum Analizi (Correspondence Analysis): $r * c$ gibi iki boyutlu veya $r * c * m$ gibi çok boyutlu tablolaştırılan değişkenlerin arasında birlikte olan değişimleri, tablolardaki kikare değerleri yardımıyla hesaplanmış olan varyans elemanlarından yararlanılarak grafiksel gösterim yardımıyla yorumlamak, incelemek için kullanılan bir tekniktir.

9-Kümeleme Analizi (Cluster Analysis): Çok değişkenli verilerin arasındaki benzerlik veya farklılıklardan faydalanarak birim veya değişkenlerin veya her ikisinin de oluşturdukları kümeleri ortaya çıkarmak amacıyla kullanılan bir tekniktir.

10-Setlerarası Korelasyon Analizi (Canonical Correlation Analysis): Çok değişkenli iki veya ikiden çok değişkenlerin veri seti arasında bulunan korelasyonları açıklayabilmek ve bulunan bu korelasyonları yorumlayabilmek için kullanılan bir tekniktir.

11-Çok Boyutlu Ölçekleme Analizi (Multidimensional Scaling): Faktör analizine seçenek olarak ortaya çıkmış bir tekniktir. Çok boyutlu ölçekleme analizi gözlenen birimler ve nesnelerin arasında oluşan benzerlikler ile farklılıkları izah etmede araştırmacılara fayda sağlayan, boyutlara yol açan anlamlı yapıları belirlemede kullanılan bir tekniktir.

12. Doğrusal Olasılık Modelleri (Logit Analysis): Bu model ayırma analizi teknikleri ve çok değişkenli regresyonun ve birleşmiş halidir (Hair ve ark. 1998). Burada uygulanmış olan yöntem çok değişkenli regresyona benzemekte birden çok bağımsız değişken, tek bağımlı değişkeni hesaplamada kullanılan bir tekniktir.

13. Probit Regresyon Modeli (Probit Regression Models): Bağımlı değişken evet ya da hayır gibi cevaplardan oluşan ve 0 veya 1 şeklinde ikili olarak kodlanmış kategorik modelleri hesaplamak için kullanılan bir tekniktir. Lojistik regresyona bir seçenek olacak şekilde ortaya çıkmıştır (Albayrak ve ark. 2005).

14. Lojistik Regresyon Analizi (Logistic Regression Analysis): Bağımsız değişkenlerin nominal ve metrik olması halinde çok değişkenli normallik varsayımı, gerçekleşmemektedir. Çok değişkenli normallik varsayımı bozulduğu andan itibaren testlerin sınıflandırılmasını ve testlerin anlamlılığı olumsuz olarak etkilenir. Bağımsız değişkenler için lojistik regresyon analizi, dağılımsal varsayımlar da bulunmamaktadır (Sharma 1996).

3.2. Çok Değişkenli İstatistiksel Analizin Kullanım Amaçları

Çok değişkenli istatistik çok sayıda özelliğin analiziyle ilgilendiğinden dolayı uygulamalar içerisinde farklı amaçlar için kullanılmaktadır. Çok değişkenli analiz bilimsel araştırmalarda başka amaçlarla kullanılmaktadır. Örneğin; (Tatlıdil 1996)

Basitleştirme ve Veri İndirgeme: k kütle olmak şartıyla içerisinde p tane değişken bulunduran veri setinin çeşitliliğini açıklayan aynı zaman da aralarında bir ilişki olmayan az sayıda değişken $(k < p)$ yardımıyla veri yapısını yorumlamayı sağlamak.

Veri Setinin Çok Değişkenli Veri Analizine Uygun Hale Getirmek: Çok değişkenli ve tek değişkenli yöntemlerin veri yapılarına uygulanabilmesi için veri setinin bu iki yöntemin temel varsayımlarına uygun olması gerekmektedir. Mesela çok değişkenli için Normal dağılıma uygunluk, ardışık bağımlılık, bağımsız değişkenler arasında ilişki bulunmaması ve hata terimlerinin farklı varyanslılık içermemesi gibi birkaç varsayımı içermesi gerekmektedir.

Verilerin Kümelenmesi ve Sınıflandırılması: Sınıfsal yapıları belli olmayan değişkenler arasında birbirleri ile benzeyen obje veya değişkenlerden gruplar belirtme faaliyetlerine yardımcı olmak.

Sıralama ve Ölçekleme: Sıralama işlemi yapılırken bazı uygulamalarda birimleri belirli ölçülerine göre sıralamak temel amaç iken, ölçeklemede ise çok sayıda değişkenden yararlanılarak daha az boyutta gösterim temel amaçtır. Karşılaştırmalarda, yakınlık ve uzaklık durumlarında, grafiksel gösterimlerin gözlemlenmesinde kolaylık sağlamaktadır.

Çok Değişkenli Hipotezlerin Test Edilmesi: k tane çok değişkenli ortalamaları vektörünün eşitliğine veya farklılığına dair kurulacak hipotezleri belirlemek.

Çok değişkenli istatistiksel analiz tekniklerinin kullanımı, değişkenlerin arasında bulunan bağlantıya göre bağımlılık ve karşılıklı bağımlılık olacak şekilde ikiye ayrılır.

1. Bağımlılık Analizi: Bir bağımlı değişken veya değişkenler diğer bağımsız değişkenler tarafından tahmin edildiği ya da açıklandığı yöntemlerdir. (Hair vd. 2002).

2. Karşılıklı Bağımlılık Analizi: Değişken ya da değişkenler grubu başkalarına bağlı kalmadığı ya da açıklanamadığı, değişkenler bağımlı veya bağımsız şeklinde belirtilmediğinde tüm değişkenler içerisinde karşılıklı bağıntılar olmaktadır.

3.3. Çok Değişkenli İstatistiksel Analiz Yöntemlerinin Varsayımları

Çok değişkenli istatistik yöntemlerin dayandığı varsayımlar vardır. Çok değişkenli istatistik yöntemler araştırmanlara kolaylık sağlarken aynı zamanda yöntemlerin varsayımlara dayanması araştırmacılara büyük bir zorluk sağlamaktadır. Bu bölümde, varsayımlar ve bunlardan oluşan sapmaların istatistik testler üzerinde oluşan etkileri ele alınmıştır. Aynı, zamanda verilerin varsayımlara uygunluğunu sağlamada kullanılacak dönüşümler ve verilerin varsayımlara uygunluğunun araştırılması gibi konular incelenecektir.

1.Çoklu normal dağılım: Burada kullanılacak veriler çok değişkenli normal dağılım ile uygunluk gösterir.

2.Eşkovaryans: Bütün veri grupları için kovaryans matrisi eşitlik gösterir.

3.Çoklu doğrusal bağlantı: Veri seti içindeki bağımsız değişkenler arasında anlamlı ve doğrusal açıdan tutarlılık yoktur.

4.Doğrusallık: Değişkenler arasındaki tutarlılık doğrusaldır.

5.Bağımsızlık ve otokorelasyonun olmaması durumu: Değişkenlerin birim değerleri birbirleri ile alakasızdır.

3.3.1. İstatistik testlerin gücü ve anlamlılığı

Testin gücü, testin istatistik bakımından anlamlı olması olasılığıdır. Diğer bir ifade ile gerçekte yanlış olan sıfır hipotezinin reddedilmesi olasılığıdır. Araştırmacıların en temel amacı anlamlılığı yüksek olan sonuçlar elde etmektir. Bu testlerin gücü hipotez testleriyle birlikte ele alınmıştır. Genellikle bu hipotezlerde alfa ve beta olmak üzere iki tip hatayla karşılaşılmaktadır. Hipotez testinin ikinci aşaması olan testin anlamlılık düzeyinin belirlenmesi durumunda araştırmacı alfa hatasını belirlerken beta düzeyi de belirlenmiş olur. Alfa ve beta hatalarından yalnızca birinin karar verme aşamasında önemli olması halinde ilgilenilen hatayı küçük tutarak öteki hatanın büyümesi sağlanabilir. Fakat her iki hatanın da önemli olması halinde tek bir çözüm vardır bu çözüm örnek hacmi arttırılarak alfa ve beta hatalarını beraber küçültmektir (Orhunbilge s.150).

Hipotez testlerinde alfa düzeyi araştırmacılar tarafından seçilmektedir. Mesela %5'lik alfa düzeyi için araştırma birkaç kere tekrar edildiğinde bu araştırmaların yaklaşık %5 kadarı sıfır hipotezinin yanlışlıkla reddedilebilceği anlamı taşır. Fakat hipotezlerden sapma olması halinde, yanlış sıfır hipotezinin reddedilmesi oranı alfa düzeyinin üzerinde ya da altında olabilmektedir. Mesela, çok değişkenli ifadelerin normallik varsayımından sapma düzeyine tabi olarak belirlenen nominal alfa düzeyi %5 olsa bile gerçek alfanın düzeyi %25 olabilmektedir. Başka bir ifadeyle hipotez testlerinin gerçek alfa düzeyi ile nominal alfa düzeyi arasındaki eşitlik normallik ve başka varsayımların sağlanabilmesi durumuna bağlıdır. Aksi takdirde gerçek alfa düzeyi ile nominal alfa düzeyi birbiriyle aynı olmamaktadır. Böylelikle istatistik testinin gücü, $1-\beta$ ifadesine eşittir ve gerçekte yanlış kabul edilmiş sıfır hipotezinin reddedilmesi

olasılığıdır. İstatistik testinin gücü düşük ise hesaplanabilen anlamlılık düzeyleri düşüktür. Bundan dolayı, bilimsel araştırmalarda istatistik testlerin gücünün ve alfa düzeyinin varsayımlarda bulunan sapmalardan ne şekilde etkilendiğinin bilinmesi için oldukça önemlidir.

3.3.2. Normallik varsayımı

Çok değişkenli istatistik tekniklerden bazıları modellerin çok değişkenli normal dağılımdan oluşan ana kütleden tahmin etmektedir. Bu tahmin yani varsayım, bazı işlemlerin ve sonuçların yorumlanmasında kolaylık sağlamaktadır (Tatlıdil 1996). Çok değişkenli istatistik tekniklerin esas tahminlerinden birisi olan normallik varsayımından sapmaların hepsi anlamlıysa, t ve F istatistiklerinin hesaplanabilmesin de bu tahmin gerekli olduğundan, ulaşılan testler geçerliliklerini yitirmiştir. Tek değişkenli teknikler tek değişkenli normallik varsayımına, çok değişkenli teknikler ise hem tek değişkenli normallik varsayımına hem de çoklu normallik varsayımına dayanmaktadır.

Araştırmacılar çok değişkenli ve tek değişkenli analizlerde normallik varsayımından sapmaların alfa hatası üzerinde güçlü bir şekilde etkili olmadığını göstermişlerdir (Glass vd.1972; Mardia 1971; Everitt 1979; Olson vd. 1974). Normallik varsayımından sapma istatistik testlerin gücünü ve sınıflandırma oranını etkiler.

Tek değişkenli normal dağılım basit bir şekilde tespit edilebilmektedir. Tek değişkenli istatistik tekniklerinde normallik varsayımına uyum sağlamayan değişkenlerin dağılımını, uygun normal dağılama çevirebilmek için belirli dönüşüm uygulanmaktadır. Çoklu normal dağılım, her değişkenin tek değişkenli normal dağılımları sağladığını ve ilgilenilen değişkenlerin kombinasyonlarının normal olduğunu tahmin etmektedir. Kısaca bir değişken çoklu normal dağılım için uygun ise, tek değişkenli normal dağılım içinde uygundur denilir. Tersi her zaman doğru değildir. Normal dağılımdan sapmaların etkisini büyük örnekler her ne kadar azaltsa da analiz kısmına dahil edilecek bütün değişkenler için normallik varsayımı sağlanır (Hair vd. 1998).

Tek değişkenli normal dağılım 3 basıklık ve sıfır çarpıklık değerlerine sahiptir. Kimi durumlarda basıklık olarak verilen değerden 3 çıkarılır ve bu şekilde basıklık değerimiz normal dağılım gösteren değişkenlerde sıfır olmaktadır. O halde tek

değişkenli normal dağılımların çarpıklık ve basıklık değerleri sıfırdır. Sağa çarpık dağılım pozitif, sola çarpık dağılım negatif çarpıklık değerine sahip olur. Araştırmacılar çoklu normal dağılımın sapmasının bütünüyle çarpıklıktan oluşması halinde istatistik testinin gücüne etki etmediğini göstermiştir (Sharma 1996).

Basıklık ölçüleri dizideki birimlerin dağılımının sivri (leptokurtic) mi, basık (platykurtic) mı yoksa normal (mesokurtic)mi olduğunun araştırılmasında kullanılan bir yöntemdir. Bu ölçülerin asıl hedefi, değişkenlerin ortalamanın etrafında ne şekilde dağılım gösterdiğini ortaya çıkarmaktadır. Basıklığın değeri eğer pozitif değerli ya da 3'ten büyük ise sivri, negatif değerli ya da 3'ten küçük ise basık ve sıfır ya da 3 ise normal kabul edilir. Değişkenlerin normalleştirilmiş basıklık değerleri -3 ile $+3$ aralığın da olduğundan bahsedilen değerler klasik normal dağılımdan gelmektedir. Dağılımın basıklığı istatistik testinin gücü üzerinde etkilidir, fakat bu etki sivri dağılımdan ziyade basık dağılıma göredir.

- **Tek değişkenli normallik testi**

Öncelikle tek değişkenli normal dağılımı inceleyelim. Bu incelemenin birinci amacı çok değişkenli normallik testinin tek değişkenli normallik testine göre karmaşık ve zor olması ve çok değişkenli normallik normallik testini sağlaması tek değişkenli normallik testlerine bağlı olması durumudur. İkinci amacı ise çok azda olsa, tüm aykırı dağılımlar normal dağılım için uygun olsa da çoklu normal dağılımın sağlanamaması durumudur. Tek değişkenli normallik testi varsayımının sağlanıp çoklu normal dağılım testinin sağlanamadığı durumlarda azda olsa bulunmaktadır (Grandesikan 1977). Özetle birimlerin çok değişkenli normal dağılımdan yola çıkarak sapma göstermesi halinde normal dağılımı sağlamayan değişkenler araştırılmalıdır. Tek değişkenli normal dağılım testinin belirlenmesi için grafik ve analitik testlerinden yararlanılmaktadır.

Grafik Testler: Görsel normallik varsayımını belirleyebilmek için gövde-yaprak, Q-Q, histogram, kutu ve P-P vs. grafikleri kullanılır. Burada en çok kullanılan Q-Q grafiğidir (Norusis, 1993). Bu Q-Q grafikleri aşağıda verilen adımlar ile elde edilir (Johnson vd. 1992).

Gözlem sayısı n ise gözlem değerleri $x_1 < x_2 < \cdots < x_n$ biçiminde küçükten büyüğe doğru sıraya konulur. Genellikle sürekli değişkenlerde ki gibi gözlemlerin farklı olması, j kadar gözlem x_j değerinden küçük veya x_j değerine eşit olmalıdır.

x_j değerinden küçük olan gözlemlerin oranları $\frac{(j-0,5)}{n}$ biçiminde hesaplanmaktadır. Gözlemlerden $0,5$ değerinin çıkarılmasının amacı sonsuzluğun düzeltilmesidir. Her j değeri için bu oran, z değerleri ve kümülatif normal dağılım için olasılık düzeyleri, belirlenen normal dağılımlar için muhtemel olasılıkları ortaya çıkarmaktadır.

Teorik z değerleri ile sıralanmış gözlemlerin x_j arasındaki belirtilen grafiğe Q-Q grafiği denir. Doğrusal olan bir grafik normal dağılımı gösterirken doğrusal olmayan bir grafik normal olmayan dağılım göstermektedir.

Tek değişkenli normallik için analitik testler

Kolmogorov-Smirnov (K-S) Z, Ki-kare uygunluk ve Shapiro-Wilks (W istatistiği) testleri normallik varsayımlarını yorumlamada kullanılan analitik testlerdir. Buradaki Ki-kare uygunluk testi rastgele bir örneğin çağrışımından yola çıkarak, bu örneğin belirli kurumsal olasılık dağılımını belirten bir ana kütleden gelip gelmediğini belirlemede kullanılır. Testinlerin uyarlanabilmesi için olasılığa ait frekanslar en az 5 ya da 5'ten daha büyük olmalıdır (Orhunbilge 2000).

Kolmogorov-Smirnov Z testi diğer bir analitik testi olan Ki-kare uygunluk testine seçenek olarak kullanılan bir testtir. Ki-kare testinin uyarlanabilmesi önemli olan şıklara ilişkin frekansların en az 5 ya da 5 den daha büyük olması şartı bulunmamaktadır. Bu testte ulaşılan kümülatif göreceli frekans dağılımı H_0'dan yola çıkarak öne çıkan anakütle teorik olasılık dağılımı ile karşılaştırılmasına dayanır. Kolmogorov-Smirnov Z testi, araştırılan kümülatif göreceli frekans dağılımıyla (f) ile H_0 içinde ki kümülatif teorik frekans dağılımı (f') arasında maksimum mutlak fark aracılığı ile hesaplanır. Yani $K - S\,z = Max|f - f'|$ dır.

1968 yılında Wilks, Shapiro ve Chen aracılığıyla gerçekleştirilen çalışmalardan yola çıkılarak Wilks testinin normallik varsayımını yorumlamada en kuvvetli test

olduğu ispatlanmıştır (Sharma 1996). Veriler normal dağılımlı bir ana kütleden gelmiyor ise dağılımın çarpıklık ve basıklık belirtileri incelenerek daha ayrıntılı yorumlar yapılabilmektedir. Bu basıklık ve çarpıklık istatistikleri aşağıda verildiği gibi hesaplanır.

$$Z_{Çarpıklık} = \frac{Çarpıklık}{\sqrt{\left(\frac{6}{N}\right)}}, \quad Z_{Basıklık} = \frac{Basıklık}{\sqrt{\left(\frac{24}{N}\right)}}$$ (Hair vd. 1995)

Hesaplanan z değeri kritik değeri aşması durumunda dağılımın alakalı özellik için normal dağılım olmadığına karar verilir. Kritik olan z değeri nominal anlamlılık düzeyini göstermektedir ve normal dağılım tablosu aracılığı ile belirlenir. Mesela z değeri $\mp 2{,}58$ değerini geçmesi halinde $\%1$ anlamlılık düzeyinde dağılımda ki normallik varsayımı reddedilmektedir. $\%5$ anlamlılık düzeyi için sıkça yararlanılan kritik değer $\mp 1{,}96$'dır.

- **Çok değişkenli sapan birim değerlerinin incelenmesi**

Çok değişkenli sapan gözlemlerin saptanabilmesi için analiz kısmın da kullanılacak bağımsız değişkenlerin arasında bulunan kareli Mahalanobis uzaklıkları (MD^2) hesaplanmaktadır. Yani analizin içindeki her gözlem için belirlenen kareli Mahalanobis uzaklıkları analizin içindeki değişken sayısına bölünerek ortaya çıkan ölçü (MD^2/df) t dağılımına uyar. Rastgele gözlemlerin sapan değer adı altında yorumlanabilmesi için $\%1$ anlamlılık düzeyinde anlamlı olmalıdır. Şöyle ki ; MD^2/df değeri $5{,}014$' ten büyük olmalıdır(Hair vd.1995). Genellikle, sapan değerlerinin analizden çıkarılması normal dağılıma uyan değişken için önemlidir bu demektir ki dönüştürülmesi gereken değişken daha azdır. Ancak bundan dolayı her sapan değerin analizden çıkarılması düşünülmemelidir.

- **Çok değişkenli norma dağılım testleri**

Çoklu normal dağılımları araştırmada kullandığımız testler tek değişkenli testlerde kullandığımız grafik ve analitik testlerinde ki gibi ikiye ayrılır fakat burada az sayıda test vardır. Bu bölümde tek değişkenli normallik testlerinde belirtilen grafik yöntem ve bu yönteme dayandırılan analitik yöntem açıklanır. Grafik testi, bir diğer test olan Q-Q grafiğine benzemektedir. Analitik testler de basıklık ve çoklu çarpıklık

istatistiklerinin yorumlanmasına dayanır. Çoğu istatistik paket programında çoklu çarpıklık ve basıklık istatistikleri hesaplanamaz. Örneğin; SPSS 10.05, SAS, NCSS 2001 ve Statgraphic Plus 3.0 programlarında çoklu çarpıklık ve basıklık istatistikleri hesaplanamaz. Aynı zamanda istatistikler için matematik dağılımları hesaplanamaz ise bu durumda çok değişkenli normallik testi kullanımı azalmaktadır (Sharma 1996).

Çoklu normallik testine birimler için kareli Mahalanobis uzaklıklarını hesaplayarak başlanır. Eğer ana kütleler normalse ve örnekler yeterince büyük $(n \geq 25)$ olduğu zaman uzaklıklar Ki-kare dağılımına uygundur (Johnson vd.,1992).

Toplam gözlem sayısı n olacak şekilde, öncelikle kareli Mahalanobis uzaklıkları küçükten büyüğe doğru aşağıdaki gibi sıralanır:

$$MD^{12} < MD^{22} < \cdots < MD_n^2$$

İkinci olarak, kareli Mahalanobis uzaklık değeri (MD^2) için, j gözlem değerinin sayısı olmak üzere, $\frac{j-0,5}{n}$ ifadesi ile yüzdelikleri hesaplanmaktadır.

Üçüncü olarak ikinci kademede elde ettiğimiz yüzdeliklere göre, p değişkenin sayısını aynı zaman da serbestlik derecesini belirtmek üzere, ters birikimli ki-kare değerleri hesaplanır.

Son aşamada ise Mahalanobis uzaklık değerleri (MD^2) yardımıyla ile ki-kare (x^2) birimlerinin grafikleri çizilecektir.

Eğer veriler normal dağılım için uygun değilse değişkenlere bazı dönüşümler yapılarak normal dağılıma dönüştürülmesi sağlanır. Çok değişkenli analizler aykırı dağılımı normal olmayan bütün değişkenlere uygun dönüşümler yapılarak çoklu normal dağılıma uygun hale getirilmeye çalışır.

Çizelge 3.1. Normalliği sağlamada kullanılacak dönüşümler (Johnson vd., 1992)

Ölçek Tipi	Dönüşüm
Mutlak Büyüklükler	Karekök, Logaritmik veya Hiperbolik

Çizelge 3.1'in devamı

Oranlar (p)	$Logit(p) = 0{,}5Log\frac{p}{1-p}$ veya arcsin(x)
Korelasyon (r)	$FisherZ = 0{,}5Log\frac{1+r}{1-r}$

Normal dağılıma uymayan değişkenlerin normal dağılıma dönüştürülmesi işlemi dağılımın basıklık ve çarpıklığına göre yapılmaktadır. Çarpık dağılım için logaritmik, karekök veya hiperbolik dönüşümler uygunken, basık dağılımlar için genelde hiperbolik $(1/x)$ dönüşümler en uygundur. Ancak araştırman bu işlemi yaparken her türlü dönüşümleri uygulayarak bu dönüşümler sonucunda en uygun olan, en iyi sonucu veren dönüşümü tercih etmelidir (Hair vd. s.77). Logaritmik dönüşüm genellikle sağa çarpık dağılımları normalleştirmede; karekök dönüşümü ise genellikle sola çarpık dağılımları normalleştirmede en iyi sonucu vermektedir. Normalliğin sağlanması için mutlak büyüklükler de karekök, oranlar da $logit$ ya da $arcsin(x)$, korelasyonlar da ise $Fisher - Z$ dönüşümü en iyi sonucu verir. Bu dönüşümler aracılığı ile normallik varsayımı sağlanmadığı takdirde ise normalliği sağlamak için analitik yaklaşımlar kullanılmaktadır.

3.3.3. Kovaryans matrislerinin eşitliği varsayımı

Farklı varyans probleminin genellikle normallik problemlerinden referans aldığı gibi değişkenlerin dağılım biçiminden de referans alabilmektedir. Farklı varyans halinde regresyon analizi hataları gözlemlendiğinde hataların dağılımları bir koniye benzetilmektedir. Böyle durumlarda eğer koni sola açılmış ise değişkenin karekökü, koni sağa açılmış ise değişkenin tersi alınır. Kimi dönüşümler veri türüne bağlı olmaktadır. Mesela; oranlar için en iyi sonucu $\arcsin\left(x^* = 2arcsin\sqrt{x}\right)$ ya da logaritmik dönüşümler verir. Daima değişkenlere uygulanmış dönüşümlerin çözüm yolu getirip getirmediğinin kontrolü sağlanmalıdır (Hair vd. s. 77).

Tek değişkenli analizler de kovaryans matrisi durağan bir sayı ve bağımlı değişkenin varyansı tüm hücreler açısından eşit olduğunda varsayım sağlanır. Fakat Manova analizi ile diskriminant analizinde kovaryans matrisinin bütün hücreleri eşit

olduğunda varsayım sağlanmaktadır. Örneğin, kovaryans matrisinde üç varyans $(\sigma_1{}^2, \sigma_2{}^2, \sigma_3{}^2)$ ve üç kovaryans $(\sigma_{12}, \sigma_{13}, \sigma_{23})$ mevcut bağımlı değişkeni inceleyelim. Kovaryans matrislerinin eşitliklerinin varsayımının bulunabilmesi için matrislerde bulunan altı unsur eşit olmalıdır. Bundan dolayı diskriminant analizi ve Manova analizinde kovaryans matrisleri eşitliği hipotezinden sapma ihtimali Anova analizinin sapma ihtimalinden daha yüksektir (Sharma vd. s.383).

Alfa ve beta hatalarını kovaryans matrislerinin eşitliğinden sapmalar etkilemektedir. Yapılan simülasyon çalışmalarında etkinin beta hatasından çok alfa hatasını daha çok etkilediği gözlemlenmiştir. Yapılan çalışmalar eşit hücre büyüklükleri halindeki anlamlılık düzeyleri, eşit olmayan kovaryans matrislerinden etkilenmemiştir (Holloway vd. 1967). Bundan dolayı yapılması gereken şey eşit hücre büyüklükleri elde etmektir. İki kümeli analizler için aşağıda verilen çıkarımlar simülasyon faaliyetlerinden elde edilmiştir (Holloway vd. 1967). Eğer küçük grup büyük gruptan çok daha büyük bir varyansa sahip ise test liberal olmaktadır ve aynı zaman da testin gerçek alfa seviyesi nominal alfa seviyesinden büyüktür. Tersi durumda, eğer büyük grup küçük gruptan çok daha büyük varyansa sahip ise test tutucu olmaktadır ve testin gerçek alfa seviyesi nominal alfa seviyesinden daha küçük olur.

- **Kovaryans matris eşitliğinin test edilmesi**

Manova, Diskriminant analizi ve diğer bütün çok değişkenli istatistik teknikler kovaryans matrislerinin eşit olduğu varsayımında bulunmaktadırlar. Bundan dolayı ilk olarak 1949 yılında Box adlı kişi tarafından M istatistiği geliştirilmiştir. Box-M istatistiği eşvaryans testi olan tek değişkenli Barlett-Box F testinin genelleştirilmesi sonucunda belirlenmiştir. Aynı zamanda $p = 1$ için hesaplanıcak olan Box-M istatistiğinin eşit olduğu istatistik Barlett-Box F dir. p sayıda değişken için ölçülmüş g sayıda grup vardır ve bu grupta ki nesne sayısı n_j ile gösterilsin. Bu durumda Box-M, Barlett-Box F istatistiğinin genelleştirilmesi sonucunda şu şekilde ifade hesaplanır (Tacg s.248-255).

$$Box - M = (N - g) \times Ln|S_w| - \sum_{j=1}^{g} (n_j - 1) \times Ln(S_{\{j\}})$$

Buradaki,

$$N = \sum_{j=1}^{g} n_j \qquad S_w = \frac{\sum_{j=1}^{g}(n_j - g)S_j}{N - g}$$

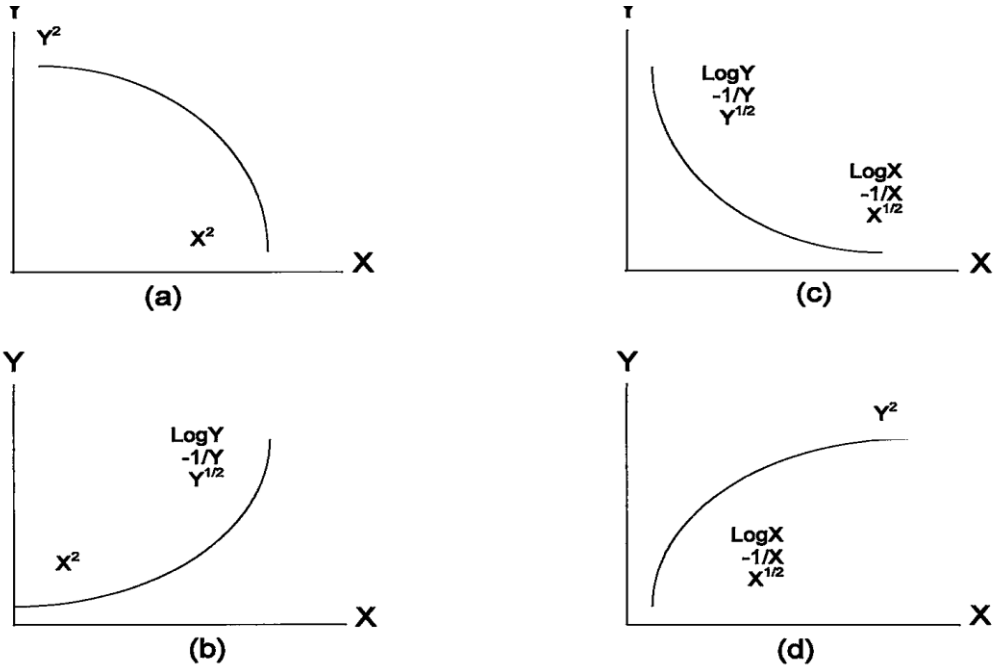
Box-M istatistiğinin anlamlılığı Ki-kare ve F istatistiği ile test edilir.

Kovaryans matrislerinin eşitliğinin tespit edilmesinde yararlanılan bütün istatistikler normallik varsayımına karşı hassastır.

Eğer Box M istatistiği önemli bulunduğu durumda eşit kovaryans matrisin varsayımının sağlanmadığı görülür. Böylelikle elde edilen grupların varyanslarını inceleyerek testin tutucu olup olmadığına ya da liberal olup olmadığına karar verilir. Örneğin küçük grupların varyansı çok daha büyük ise test liberaldir. O halde çok değişkenli testlerin sonuçları tesadüfi olabilmektedir. Bundan dolayı, kovaryans matrisinin eşit olmasını sağlayacak dönüşüm uygulandıktan sonra analiz tekrardan yapılmalıdır. Buna göre gruplarda hangi değişkenin değişik varyansa sahip olduğunu belirtmede tek değişkenli testler uygulanmaktadır. Bundan dolayı Barlett-Box F testi kullanılır.

3.3.4. Doğrusallık varsayımı

Yapısal eşitlik modeli, çoklu regresyon analizi, diskriminant analizi ve faktör analizi gibi korelasyonların katsayılarına dayanan çok değişkenli yöntemlerin gizli olan varsayımlarından biride doğrusallık varsayımıdır. Doğrusal olmayan örnekler için hesaplanan doğrusal korelasyonlar ilişkiyi daima az gösterir.



Şekil 3.1. Doğrusallık varsayımı içi dönüşüm seçimi (Hair vd. s.77)

İki değişkenin arasında bulunan doğrusallığı sağlamak için birçok sayıda dönüşüm uygulanmakta, fakat doğrusal olmayan bağlantılar dört grup da toplanmıştır. Verilen bir grafik kuadratik fonksiyonun bağımlı ve bağımsız değişkenlere uygulanmış dönüşümlerini gösterir. Çoklu dönüşüm alternatifleri doğrusallık olana kadar en üst basamaktan ki dönüşümden başlayarak aşağı doğru hareket ederek dönüşümler gerçekleştirilir. Bu dönüşümler doğrusallığın şartı sağlanıncaya kadar yukarıdaki dönüşümden aşağıdaki dönüşüme doğru olacak şekilde gerçekleştirilir.

3.3.5. Eşvaryans, normallik ve doğrusallık için dönüşümler

İstatistikte veri dönüşümünün üç temel amacı vardır, bunlar durağanlaştırma, normalleştirme ve doğrusallaştırmadır. Bazı dönüşümlerle genelde ilk iki amaç bazı durumlarda da üçüncü amaç gerçekleştirilmektedir. Yaygın olarak kullanılan dönüşümler aşağıda verilmiştir:

- **Logaritmik dönüşüm** $[x^* = \boldsymbol{LG10}(x), \boldsymbol{LN}(x)]$: Logaritmik dönüşüm negatif sayıların logaritmasının alınamamasından dolayı değerleri pozitif olanlara uygulanmaktadır. Fakat negatif değerleri pozitif değere çevirmek için değişken değerlerine bir sabit sayı eklenerek negatif değerler pozitif değerlere çevrilmiş olur. Bu

dönüşüm şunun için kullanılır; x nin dağılımı sağa çarpık ise x'in dağılımını normalleştirmek; x artarken x'nin varyansıda artıyor ise varyansı durağanlaştırmak ve bağımlı değişken bağımsız değişkenlerle sürekli artış olan bir eğimi gösteriyor ise modeli doğrusallaştırma da kullanılır.

- **Logit dönüşüm [$\boldsymbol{Logit(p) = 0,5LG10(\frac{p}{1-p})}$]**: Eğer değişken değerleri oranları gösteriyor ise değişkeni normalleştirmek amacıyla kullanılır. p değişkeni 0 ve 1 aralığında değer alır.

- **Karekök dönüşüm** $\boldsymbol{(x^* = \sqrt{x})}$: Eğer varyans x'nin ortalamasıyla orantılıysa varyansı durağanlaştırmak için kullanılır. Dönüşümün daha olumlu sonuçlar vermesi için bağımlı değişkenin Poisson dağılımına uyması gerekmektedir. Aynı zamanda değişkenin dağılım sola çarpık ise değişkeni normalleştirmede kullanılır.

- **Arcsin dönüşümü ($\boldsymbol{x^* = Arcsin\sqrt{x} = Sin^{-1}\sqrt{x}}$ veya $\boldsymbol{x^* = 2Arcsin\sqrt{x}}$)**:

Değişkenlerin değerleri oran şeklinde verilmişse varyansı durağanlaştırmak amacıyla kullanılır. Dönüşümün yapılabilmesi için x pozitif olmalıdır.

- **Fisher Z dönüşümü [$\boldsymbol{FisherZ = 0,5LG10\left(\frac{1+r}{1-r}\right)}$]**: Eğer değişkenlerin değerleri korelasyonlar ise değişkeni normalleştirmek amacıyla kullanılan bir dönüşümdür.

- **Kare dönüşümü** $\boldsymbol{(x^* = x^2)}$: Bağımlı değişkenlerin hata değerleri sola çarpık ise bağımlı değişkeni normalleştirme de kullanılır.

- **Hiperbolik dönüşüm** $\boldsymbol{\left(x^* = \frac{1}{x}\right)}$: Eğer varyans x 'nin dördüncü derecesiyle orantılıysa varyansı durağanlaştırmak için kullanılır ve dağılımı basık olarak verilmiş değişkenlerin dağılımını normalleştirme de kullanılır. Hiperbolik dönüşüm ise değişkende ki büyük değerlerin etkisini azaltır.

3.3.6. Dönüşüm (transformasyon) için genel kurallar

Değişkenlere dönüşümü uyguladıktan sonra çözüm getirip getirmediğinin kontrolü sağlanır. Sürekli uygulanmış olan dönüşüm kesin olarak çözüm

getirememekte, yalnızca varsayımlardan oluşan sapmaların derecesini azaltmaktadır. Dönüştürme yapılmış değişkenler varsayımlardın sapmalarını azaltmış olsa da analiz sonuçlarının belirlenmesini zorlaştırdığı, karmaşıklık oranını arttırdığından dolayı analizden ayrı tutulması öncelikli tercihtir. Bundan dolayı dönüşüm uygulama yapılmadan önce mümkün ise değişkeni olandan farklı bir ölçü birimine dönüştürme durumunun uygunluğu araştırılmalıdır (Helberg 2002). Dönüşüm yapılırken dikkat edilmesi gereken önemli noktalar aşağıda verilmiştir (Hair vd. s.78):

Dönüşümlerden iyi etki gözlemlemek için değişkenin ortalamasının standart sapmasına oranı 4'ten küçük olmalıdır.

Dönüşüm farklı iki değişkenden birisine uygulanacak ise ortalamasının standart sapmasına oranının çok düşük olduğu değişken tercih edilmelidir.

Birbirinden farklı varyanslılık durumu yok ise sadece bağımsız değişkenler dönüştürülmektedir.

Farklı varyanslılık durumu, bağımlı değişkenler dönüştürülerek düzeltilir. Farklı varyans bağlantısı doğrusal değil ise bağımlı değişkenler gibi bağımsız değişkenlerde dönüştürülmelidir.

Dönüşümler değişkenlerin yorumunu değiştirebilmektedir.

3.3.7. Çoklu doğrusal bağlantı (multicalinearity) varsayımı

Bağımsız değişkenler arasındaki bağlantı çoklu doğrusal bağlantı şeklinde adlandırılır. İki değişkenin arasında verilen ilişki eğer $+1$ oluyorsa aynı yönlü, -1 oluyorsa zıt yönlü tam bağımlılık söz konusu iken sıfıra eşit olması durumunda ise tam bağımsızlık söz konusu olur. Çoklu bağlantı durumunda bağımsız bir değişken ile başka bağımsız değişken ya da değişkenler aracılığıyla eksiksiz bir şekilde hesaplanması durumunda gerçekleşen en uç durumlara tekillik adı verilmektedir.

Çoklu bağlantı, bir bağımsız değişkenin diğer bağımsız değişkenlerle olası ilişkisinin derecesine göre bağımsız değişkenin tahmin hızını azaltmaktadır. Çoklu bağlantı arttıkça buna özgü varyans ise azalmakta ve ortak varyansın yüzdesi artar.

- **Çoklu doğrusal bağlantı probleminin sonuçları**

Çoklu doğrusal bağlantı problemleri genellikle aşağıda verilen sorunlara yol açmıştır (Gujarati 1995).

Çoklu doğrusal bağlantı olması durumunda, katsayıları tanımsızdır ve bu tanımsız olan katsayıların standart yani klasik hataları sonsuz olur.

Çoklu bağlantı durumlarında katsayıların hem kovaryans hem de varyansları artmaktadır.

Çok sayıda bağımsız değişkenden kaynaklanan örneğin çoklu korelasyon katsayısı yani (R^2) fazla fakat bağımsız değişkenlerin çok azı anlamlı çıkmıştır.

- **Çoklu bağlantı probleminin saptanması: koşullu endeksler ve varyans artış faktörü (VIF)**

Bağlantı problemini saptamak için kullanılan bazı yöntemler vardır (Gujarati s.335-339):

Bu yöntemlerden birincisi basit korelasyon matrisinin incelenmesi durumudur. Bağımsız iki değişken arasında verilen basit korelasyon katsayısı eğer anlamlı ise bu durum çoklu bağlantı problemine yol açmıştır. Ancak daima anlamlı korelasyon çoklu bağlantı problemine yol açmamaktadır.

Çoklu bağlantı probleminin saptanabilmesi için kullanılan bir diğer yöntemde, modelimize bağımsız değişkenler eklendikçe R^2 deki değişimi incelemektir. Eğer R^2 katsayısın da önemli gelişme olamaz ise bu bağlantı problemi için zemin oluşturmaktadır.

Çoklu bağlantı probleminin ortaya çıkarılmasında kullanılan bir diğer yaklaşımda, kısmi korelasyon katsayılarının varlığının incelenmesidir. Eğer iki değişken arasında ki basit korelasyon katsayıları anlamlı olması halinde diğer bir yaklaşım olan kısmi korelasyonun katsayıları anlamsız çıkıyor ise, bu çoklu bağlantı probleminin ortaya çıkmasını sağlar. Bundan dolayı, kısmi korelasyon daima etkili yaklaşım olamamaktadır.

Çoklu bağlantı probleminin saptanmasına yardımcı olan bir diğer önemli yaklaşım ise varyans artış faktörüdür (VIF). VIF değerlerini hesaplayabilmek için üç bağımsız değişkenli olan modeli ele alalım:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

VIF değerleri şu şekilde hesaplanır:

İlk adım olarak, X_1 bağımsız değişkenini bağımlı değişken olarak alır ve bu akınan bağımsız ile R^2 hesaplanabilir. Böylelikle X_1 değişkeninin varyans artış faktörü $VIF(X_1) = 1/(1 - {R_1}^2)$ şeklinde hesaplanır.

İkinci adım olarak, X_2 değişkeni bağımlı değişken olarak alınıp X_1 ve X_3 bağımsız değişkenlerinin arasında yer alan R^2 hesaplanır. Böylelikle X_2 değişkeninin varyans artış faktörü, $VIF(X_2) = 1/(1 - {R_2}^2)$ biçiminde hesaplanır.

Üçüncü adım olarak, X_3 değişkeni bağımlı değişken olarak alınıp X_1 ve X_2 değişkenleri arasında yer alan R^2 hesaplanır. Böylelikle X_3 değişkeninin için varyans artış faktörü, $VIF(X_3) = 1/(1 - {R_3}^2)$ biçiminde bulunur.

Bağımlı değişkenle bağımsız değişken arasında bir bağlantı yok ise $(R^2 = 0)$ $VIF = 1$ olmaktadır. Bağımlı değişken ve bağımsız değişken arasında bir bağlantı var ise $(R^2 = 1)$ $VIF = \infty$ olmaktadır. $R^2 = 0{,}9$ olduğunda $VIF = 1/(1 - 0{,}9) = 10$ olur (Webster 1995).

Çoklu bağlantının saptanmasında kullanılacak bir diğer teknikte, değişkenlerin tolerans değerlerini bulmaktır. Bu değer şu şekilde hesaplanır; $T = 1 - R_i^2$. Böylelikle daha az tolerans değeri fazla VIF demektir (Gujarati 1995).

Çoklu bağlantının saptanmasında kullanılacak diğer bir yöntemde yardımcı regresyon eşitliklerinden faydalanarak F değerlerinin hesaplanmasıdır. Örneğin, yukarıda verilen üç bağımsız değişkeni içeren modeli aşağıda tekrar ele alalım:

Bağımlı olarak tanınan bütün bağımsız değişkenle başka bağımsız değişkenlerin arasında $R_{1,23}, R_{2,13}, R_{3.1.2}$ şeklinde ki çoklu korelasyon katsayıları hesaplanmaktadır. Çoklu korelasyon katsayılarından faydalanarak tüm bağımsız değişkenler için F değeri

hesaplanabilir. Örneğin; X_1 bağımsız değişkenin de F aşağıda verilen biçimde hesaplanmaktadır (Gujarati 1995).

$$F_{1.23} = \left(\frac{\frac{R^2_{\{1.23\}}}{(k-1)}}{\frac{\left(1 - R^2_{\{1.23\}}\right)}{(n-k)}} \right)$$

Formüldeki n, örnek birim sayısının toplamını k ise sabit terim ile birlikte tahmin edilecek parametrenin sayısı $(k = 3)$ olarak ifade edilir. F değeri, bir alfa anlamlılık düzeyi ve serbestlik derecesi $df_1 = k - 2$ ve $df_2 = n - k + 1$ olan kritik F değeri ile karşılaştırılır. Eğer belirlenen $F_{1.23}$ değeri kritik olan F değerinden büyükse, X_1 değişkeni ile öteki bağımsız değişkenler olan X_2 ve X_3 değişkenleri arasında belirlenen ilişkinin anlamlı olduğu düzeyine karar verilir.

Çoklu bağlantı probleminin ortaya çıkarılmasında kullanılan bir diğer yöntem, koşullu endeks sayılarını (CI) hesaplamaktır.

$$CI = \sqrt{\frac{V_{\{max\}}}{V_{\{x_i\}}}}$$

Verilen formülde V_{max} maksimumum varyansı; V_{x_i}, i. değişken aracılığıyla belirlenen toplam varyansı gösterir. CI değeri, $10 - 30$ aralığında ise orta düzeyde, 30'u geçerse çok kuvvetli çoklu bağlantı problemi vardır.

- **Çoklu doğrusal bağlantı probleminin çözümü**

Çoklu bağlantı problemi için bazı yöntemler vardır (Gujarati 1995).

Bir ya da birden fazla bağımsız değişkeni modelden çıkartabiliriz. Ancak hangisi çıkartılabilir? Bu şekilde bir yaklaşım modeli yanlış tanımlamaya sürükleyebilir.

Kimi zaman örnek büyütülerek çoklu bağlantı problemi çözülebilmektedir. Ancak örneğe birim eklemek her zaman mümkün olmayabilir.

Birbirleri ile bağlantılı olan iki değişken ve bu değişkenlerin toplamını içeren tek değişken şeklinde modele dahil edilmektedir.

Polinomlu regresyonlarda çoklu bağlantı problemi ile çok fazla karşılaşılmaktadır. Bağlantı probleminin tekrar devam etmesi durumunda ortogonal polinomların yardımıyla bağımlı değişkenler bağımsız hale dönüştürülür. Ortogonal polinomlarda, değişkenlerin gerçek verileri yerine değişkenlerin arasında verilen farklı çeşitteki dereceli polinomları yansıtan kodlanmış biçimde verilen katsayılar bağımsız değişken adı altında kullanılır. Ortogonal değişkenlerin arasında belirtilen korelasyon sıfır olmaktadır.

Birbirlerinden bağımsız bileşenler oluşturan asal bileşenler analizi ya da birbirlerinden bağımsız faktörler oluşturan faktör analizi kullanılabilir.

3.3.8. Hataların bağımsızlığı ve otokorelasyon

Otokorelasyon rastgele bir zaman serisi ya da eşleştirilmiş zaman serisinin değerlerinin arasında bulunan korelasyondur. Çapraz korelasyon, farklı iki zaman serisinin arasında bulunan ilişkidir. Sıra korelasyonu ise hem çapraz korelasyon hem de otokorelasyon anlamına gelmektedir. Otokorelasyon ile sıra korelasyonu eş anlamda kullanılabilir (Gujarati 1995). Anlamlı otokorelasyon sistemin yanlış bir şekilde açıklandığını gösterir. Yani modelde fonksiyonel ilişki yanlış tanımlanmış ya da önemli bir değişken unutulmuş olur. Böyle bir durumda, en küçük kareler yöntemi kullanılıyor ise, regresyon analizinin hesap edilen standart yani klasik hataları düşük hesaplanacak ya da değişkenlerin anlamlılık düzeyleri ile hatalı sonuçlara varılacaktır (Orhunbilge 1996).

- **Otokorelasyon yöntemi-genelleştirilmiş en küçük kareler yöntemi:** En küçük kareler yöntemi anlamlı otokorelasyon olduğu hallerde geçersiz sonuç verebilmektedir. Geçersiz sonuçlar vermesi durumunda genelleştirilmiş en küçük kareler yöntemi kullanılmaktadır. Bu teknikte dönüştürülmüş olan değişkenlere en küçük kareler yöntemi uygulanmakta, böylelikle çok daha etkili ve kabul edilmiş standart hataları, hipotez testleri ve tahmin aralıkları elde edilir. Otokorelasyonun olması durumunda en çok kullanılan en küçük kareler yöntemimiz Cochrane-Orcutt prosedürüdür. Bu yöntem iki aşamalı bir yöntemdir.

(1) $Y_t = b_0 + b_1 X_t + e_t$

(2) $Y_{t-1} = b_0 + b_1 X_{t-1} + e_{t-1}$ [$t-1$ dönemi için]

(3) $\rho Y_{t-1} = \rho b_0 + \rho b_1 X_{t-1} + \rho e_{t-1}$ [mutlak değer ρ ile eşitlik (2) çarpılır.]

(4) $Y_t - \rho Y_{t-1} = (b_0 - \rho b_0) + b_1(X_t - \rho X_{t-1}) + (e_t - \rho e_{t-1})$

$=(1-b)b_0 + b_1(X_t - \rho X_{t-1}) + (e_t - \rho e_{t-1})$ [eşitlik (1) - eşitlik (3)]

(5) $Y_t^* = b_0{}^* + b_1{}^* X_t^* + e_t^*$ [eşitlik (4) tekrar yazılır]

Burada,

$Y_t^* = Y_t - \rho Y_{t-1}, b_0{}^* = (1-\rho)b_0, X_t^* = (X_t - \rho X_{t-1}), e_t^* = (e_t - e_{t-1}) ve\ E(e_t^*) = 0$ olur.

Yukarda verildiği üzere Cochrane-Orcutt işlemi beş adımda gerçekleşmektedir. İlk basamakta regresyon modeli yazılmalıdır. İkinci aşamada aynı modeli bir dönem geriye götürerek yazılır. Üçüncü aşamada ikinci aşamada yazdığımız regresyon doğrusunu otokorelasyonun mutlak değeri ile çarpılır. Dördüncü aşamada birinci aşamada elde ettiğimiz regresyondan üçüncü aşamada bulduğumuz regresyon çıkartılır. Beşinci aşamada ise dördüncü aşamada bulduğumuz otokorelasyon içermeyen regresyon tekrardan yazılır ve bu regresyon en küçük kareler yöntemi (EKK) yardımıyla çözülür.

En küçük kareler yönteminde ki hataların beklenen değeri sıfırdır yani $[E(e) = 0]$ Buradan hareketle 4. aşamadaki hatalar birbirlerinden bağımsızdır. Burada ki eşitliğe EKK yöntemi uygulanmaktadır.

Otokorelasyonu çözmek için yararlanılan diğer teknikler aşağıdaki gibidir:

1.Cochrane-Orcutt İki Aşamalı Prosedür (CO2)

2.Maksimum Olabilirlilik Prosedürü (ML) (Johnston 1984)

3.İteratif Yule-Walker Yöntemi (YWi)

4.Bayes yöntemi

5.Prais-Winsten Yöntemi (PW) (Johnson vd. 1992)

6.Doğrusal Olmayan En Küçük Kareler Yöntemi (NLS)

7.Hildreth-Lu Prosedürü (HL)

8.İteratif Cochrane-Orcutt Prosedürü (COi) (Cochrane 1949)

9.Theil-Negar Yöntemi (TN)

10.Yule-Walker Yöntemi (YW, iki aşamalı tam dönüşüm yöntemi)

11.Durbin İki Aşamalı Prosedürü (Durbin 1960)

3.4. Çok Değişkenli İstatistiksel Analiz Yöntemlerinde Dikkat Edilecek Hususlar

Araştırmalarda kullanılacak çok değişkenli istatistik tekniklerin hepsi belirli varsayım ve hususlar altında gerçekleşir. Bu varsayımları yerine getirmek araştırmacılar için her ne kadar zor olsa da kullanılan yöntemlerin geçerlilikleri bu husus ve varsayımlara dayanmaktadır. Çok boyutlu uzayın ve tahmin değerlerinin realist sonuçlar verebilmeleri için kullanılan veri setleri bu husus ve varsayımlara göre incelenir (Sharma 1996).

Araştırmalarda verilerden geçerli sonuçlar elde etmek istiyorsak verilerin kaliteli ve nitelikli olması gerekir. Çok değişkenli analizi yapmadan önce veri setinde dikkat edilmesi gereken önemli hususlar ve varsayımlar vardır (Mertler ve Vannatta 2017).

Hususlar ise sırasıyla aşağıda verilmiştir;

- **Verilerin hatasızlığı:** Araştırma esnasında kullanılan veri setlerinin analize uygun bir biçimde olmamaları ortaya çıkan sonuç ve yargıları geçerli kılmayacaktır. Bundan dolayı araştırma için seçilecek veri setleri kaliteli veriler olmalıdır (Mertler ve Vannatta 2017).

- **Kayıp veriler:** Veri analizlerinde sürekli karşılaşılan problemlerden biriside kayıp verilerdir. Bu sorunun önemli kısımları kayıp verilerdeki örüntünün ne kadar

büyüklükte olduğu ve sebebinin ne olduğunun bilinmemesidir. Burada önemli olan kısım kayıp verinin miktarından ziyade, kayıp verinin örüntüsüdür. Veri matrislerinde az sorun çıkaran kayıp verilerin tarafsız dağılım gösterdiği ve sorun oluşturmadıkları gözlemlenmiştir. Araştırmalarda geniş veri setleriyle çalışılıyor ise bu veri seti tarafsız biçimde dağılım gösteriyorsa oluşan sorunlar kaygı verici düzeyde olmayıp, kayıp değerleri düzeltmede kullanılan tekniklerle yeniden benzer sonuçları ortaya çıkarmaktadır. Ancak kullanılan orta ve küçük ölçekte örneklem büyüklüğü çok fazla kayıp değer oluşturuyorsa sorunlar ciddi boyutta olabilmektedir. Ne yazık ki veri kayıplarının ne seviyeye kadar örneklem büyüklüğünü tolere edebilecekleri hakkında kesin bir yargı bulunmamaktadır. Tarafsız dağılım gösteren veri kayıplarında ise az sayıda olsa da sonuçları etkilemede çok ciddi sorunlara neden olacaktır (Tabachnick ve Fidell 2015).

- **Uç Değerler:** Çok değişkenli ve tek değişkenli durumlarda, kesin ve sürekli değişkenlerde, bağımlı ve bağımsız değişkenlerde, veri ve sonuçlarda görülmektedir. I.Tip ve II.Tip hatalara neden olabilir ve bu hatanın nedenleri belirlenemeyebilir. Uç değerler ile genelleme yapmak oldukça zordur. Bu tarzda genellemeleri yapmak için örneklemin bütününün uç değerlerden oluşması gerekmektedir. Uç değerlerin oluşmasında dört temel sebep vardır:

1.Verilerin yanlış girilmesi

2.Veri setinde kayıp veri olarak tespit edilen verinin istatistiksel paket programa düzgün tanımlanması

3.Uç değerin alındığı örneklemin bir üyesi olmaması

4.Uç değerin evrenin diğer üyelerinden farklı olması

Tek değişkenli uç değerlerin dağılım sayıları bazı istatistiksel teknikler aracılığıyla standart Z değerlerine dönüştürülebilmektedir. Ayrıca tek değişkenli uç değerleri belirlemede grafiksel yöntemlerde kullanılabilmektedir. Kutu grafiği, Histogram grafiği, normal olasılık ve kısıtlandırılmış normal olasılık grafikleri tek değişkenli uç değerleri bulmada kullanılan uygun grafiksel yöntemlerdir. Çok değişkenli uç değerleri belirlemede Mahalanobis uzaklıkları kullanılmaktadır. Bu

istatistiksel işlemin sayesinde bir deneğin diğer deneklerin merkezlerinden olan uzaklıkları gösterilir. Uç değerlerin belirlenmesinden sonra bu değerlerin örneklemin alındığı evrenin üyesi olmamasından mı, deneğin diğer üyelerden farklı olmasından mı yoksa yanlış veri girişinden mi kaynaklandığı gözlemlenmektedir. Yanlış veri girişinden kaynaklanan uç değerler kontrol edilerek kolaylıkla düzeltilirken, üçüncü ve dördüncü temel nedenlerle bir uç değer belirlenmiş ise bu veriyi değiştirmek veya silmek oldukça zor bir karar olacaktır (Tabachnick ve Fidell 2015).

3.5. Çok Değişkenli İstatistiksel Analiz Yöntemlerinin Seçim Algoritması

Hair et al.'ın 1998 de yaptığı çalışma da çok değişkenli analiz teknikleri, veri setindeki değişkenlerin bir kurama göre iç bağımlı ve bağımlı şeklinde ki gibi bir farka bağlı kalmasına, şayet böyle bir farka bağlı kalıyorsa analiz kısmındaki değişkenlerden kaçının bağımlı değişken adı altında incelendiğini, iç bağımlı ve bağımlı bütün değişkenlerin ne şekilde ölçüldüğüne dair değişkenlerin ölçeğinin belirlenmesi gereklidir.

Çok değişkenli istatistiksel analiz teknikleri belirlenirken aşağıdaki kriterlerin dikkate alınması gerekir:

Çok değişkenli analiz tekniklerinin belirlenmesinde öncelikle değişkenler arasında iç bağımlı ve bağımlı ayrımının olup olmadığı belirlenmelidir. Bundan dolayı seçilen yöntemin bağımlı mı iç bağımlı olduğuna karar verilmiş olacaktır.

Bağımlı sistemlerde, bağımlı değişken ve bağımsız değişkenlerin belirlenmesi temel amaçtır, iç bağımlı sistemlerde iç ilişkilerin belirlenmesi temel amaçtır aynı zamanda iç bağımlı sistemlerde bağımlı veya bağımsız ayrımı yapılmamaktadır.

Çok değişkenli analizde geçerli olan bağımlı yöntemler iki ölçüt şeklinde sınıflandırılmaktadır. Bu ölçütler analizde kullanılan değişkenlerin ölçeği ve bağımlı değişken sayılarıdır. Örneğin çok değişkenli analizde mevzu olan sistem tek metrik yani ölçülebilir bağımlı değişkenle tanımlanıyor ise uyumlu olan teknik konjoint ya da çoklu regresyon analizidir. Diğer taraftan eğer tek bağımlı değişken metrik ölçüm değil ise doğrusal olasılık sistemleri, lojistik regresyon teknikleri ya da ayırma analizi uygun yöntemler olmaktadır.

Sistem eğer birden fazla bağımlı değişken ile açıklanacak ise dört farklı teknik uygulanır. Eğer bağımlı değişkenimiz metrik ölçümse ise bağımsız olan değişkenlere bakılmalıdır. Eğer bağımsız değişkenlerimiz metrik ölçüm değil ise çok değişkenli varyans analizi yani MANOVA, kullanılmalıdır. Sistem bağımlı değişken ya da bağımsız değişken ilişkiler setiyle tanımlanabiliyorsa yapısal eşitlik modeli kullanılmalıdır.

İç bağımlı olan tekniklerde bağımsız değişken ve bağımlı değişken farkı yapılmamaktadır. Eğer değişkenlerimiz için veri setinin esas ölçüleri araştırmanın asıl mevzusu ise faktör analizi en uyumlu yöntem olmaktadır. Eğer amaç değişkenler için veri setinin esas ölçülerini araştırmaksa kümeleme analizi uygun yöntem olmakta, nesneler için veri setinin ölçülerini araştırmaksa metrik çok boyutlu ölçekleme yöntemi en uygun yöntem olmaktadır. Çok boyutlu ölçekleme yardımıyla metrik olan yada metrik olmayan değişkenleri analiz etmektedir.

3.6. Çok Değişkenli Normal Dağılım

Çok değişkenli normal dağılımın, çok değişkenli analizlerdeki rolü büyüktür. Birden fazla çok değişkenli teknik, örneklerin çok değişkenli normal dağılım gösteren popülasyondan alınmış rastgele örnekler olduğunu farz ederek çıkarımlar yapar. Bu teknikler MANOVA, Setler arası Korelasyon Analizi, Hotelling T^2 testi, Ayırma Analizidir.

Çok az veri dışında birçok veri normal dağılım göstermektedir. Normal dağılım göstermeyen veriler bazı dönüştürme yaklaşımlarıyla normal dağılıma dönüştürülebilir. Doğrudan veya dönüştürme aracılığı ile değişkenlerin kurumsal olarak gerçek popülasyon da normal dağılım gösterdiği varsayımı, birçok problemi standartlaştırarak yaklaşımıyla çözümleme kolaylığı sağlamaktadır. Çok değişkenli normal dağılımdan yararlanarak çözümlemeler yapmanın birçok avantajlı yönü vardır. Bu avantajlar şu şekilde sıralanır:

1) Çok değişkenli normal dağılımın, matematiksel açıdan incelenmesi ve çözümler yapılması basit olan bir dağılımdır. Sürekli değişken içeren bazı olaylar normal dağılım gösterirler.

2) Çok değişkenli normal dağılım dışındaki yaklaşımlar matematikte yaygın olarak kullanılmamakta ayrıca çok değişkenli Normal dağılım dışındaki yaklaşımlarla ilgili birçok uygulamacı yeterli bilgiye sahip olmamaktadır.

3) Normal dağılım bazı hallerde popülasyonda ki dağılımların asimptotik yapısına uygun olmaktadır. Normal dağılım çoğu hallerde gerçek toplum modeli olarak görev yapmaktadır.

4) Çok değişkenli istatistiklerin örneklem dağılımları çok değişkenli normal dağılım göstermektedir.

5) Kısaca belirtmek gerekirse dünyadaki birçok problemin araştırılıp incelenmesinde normal dağılım varsayımlarından yararlanılması oldukça uygun bir yaklaşımdır

Normal dağılım dünyada ki birçok olaya ilişkin popülasyon dağılımlarının normal olması açısından bir öneme sahiptir. Normal dağılım çok değişkenli istatistik analizlerinin örneklem dağılımlarının normal olması açısından bir öneme sahiptir.

3.6.1. Çok değişkenli normal dağılım ve özellikleri

Bilindiği üzere tek değişkenli normal dağılımın varyansı (σ^2) ve ortalaması μ olan bir dağılımdır [$X \sim N(\mu, \sigma^2)$] ve olasılık yoğunluk fonksiyonu aşağıdaki gibidir;

$$f(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$-\infty < x < +\infty$ şeklinde yazılır.

Yukarıdaki eşitlikte üs olarak verilen $(\frac{x-\mu}{\sigma})^2$ teriminin matris gösterimi aşağıda verilmiştir.

$$(\frac{x-\mu}{\sigma})^2 = (\mathrm{x}-\mu)\sum^{-1}(\mathrm{x}-\mu)$$

Sabit değer olarak alınan $\frac{1}{\sqrt{2\pi\sigma^2}}$ değeri ise;

$$\frac{1}{\sqrt{2\pi\sigma^2}} = (2\pi)^{-1/2}(\Sigma)^{-1/2}$$

Σ: $p * p$ boyutlu boyutlu ve p ranklı kovaryans matrisi

x: $p * 1$ boyutlu gözlem vektörü

μ: $p * 1$ boyutlu ortalama vektörü

Bu gösterim kullanılarak tek değişkenli yapıdan p değişkenli yapıya genelleme yapılması mümkün olmaktadır.

Tek değişkenli Normal dağılım fonksiyonu; p değişken içeren X rastgele değişken veri matrisi $n * p$ (boyutlu) $\sum$ kovaryans matrisi ($p * p$ boyutlu) ve μ ortalama vektörüne ($p * 1$ boyut) göre çok değişkenli normal dağılım biçiminde aşağıdaki gibi ifade edilir.

$$f(x) = \left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) e^{-\left(\frac{1}{2}\right)(X-\mu)'\Sigma^{-1}(X-\mu)}$$

$-\infty < X_i < +\infty$; $i = 1,2,\ldots,p$

Çok değişkenli normal dağılım fonksiyonunun matris formu ise aşağıdaki gibidir:

$$f(x) = \left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) \exp\left(-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right)$$

Yukarıdaki eşitlikte $|\sum|$ kovaryans matrisinin determinantını belirtmektedir. $\sum p * p$ boyutlu simetrik bir kare matris olmaktadır. X $n * p$ boyutlu veri matrisidir.

Çok değişkenli normal dağılıma ait olan X matrisinin dağılım fonksiyonu ise $X - N_p(\mu, \Sigma)$ biçiminde gösterilir.

Kovaryans matrisi $\sum$ pozitif ve simetrik tanımlı bir kare matristir.

Çok değişkenli normal dağılım gösteren toplumlardan çekilen n birimlik örneklerdeki ($X = [x1, x2, \ldots, xp]$) p değişkeni için parametre vektörlerinin tahminleri aşağıdaki gibi ifade edilir.

$$\hat{\mu} = \bar{X}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})' = \frac{n-1}{n}S = S_n$$

Çok değişkenli istatistiklerin örneklem dağılımları ise;

1- $\bar{X} \sim N_p(\mu, (1/n)\Sigma)$

2- $(n-1)S$, $(n-1)$ serbestlik dereceli WISHART dağılır.

Wishart dağılımı; X ve S bağımsız olmak şartıyla aşağıdaki şekilde tanımlanır.

Çok değişkenli istatistikler için merkezi limit teoremi ise aşağıdaki gibidir.

μ ve S parametreli normal dağılımdan alınan rastgele sütun vektörlerinde gözlemler $(x_1, x_2, \ldots, x_n)$ biçiminde gösterilmektedir.

$n - p$ yeterince büyük olmak şartıyla;

1. $\bar{X}$ yaklaşık $N_p(\mu, \left(\frac{1}{n}\right)\Sigma)$ parametreli çok değişkenli normal dağılım göstermektedir.

2. $\sqrt{n}(\bar{X} - M)$ yaklaşık $N_p(0,\sum)$ parametreli çok değişkenli normal dağılım göstermektedir.

3. $n(\bar{x} - \mu)'\sum^{-1}(\bar{x} - \mu)$ yaklaşık p serbestlik derecesi kikare dağılır.

$$n(\bar{x} - \mu)'\sum^{-1}(\bar{x} - \mu) \sim {X^2}_p$$

4. Örnek istatistikleri kullanılarak istatistiklerin dağılımı aşağıdaki gibidir:

$$\sqrt{n}(\bar{x} - \mu) \sim N_p(0, S_n)$$

$$n(\bar{x} - \mu)'^{S^{-1}}(\bar{x} - \mu) \sim {X^2}_p$$

5. Çok değişkenli verilerde eğer normal dağılım varsayımı sağlanmıyorsa incelenen verilere uygun dönüşümler yaparak dağılımları çok değişkenli normallik varsayımına yaklaştırmak gerekmektedir. Bu dönüşüm için uygulanacak bazı dönüştürme yöntemleri şu şekilde verilmiştir:

Çizelge 3.2. Çok değişkenli normal dağılım için dönüştürme teknikleri

Orijinali Veri	Uygun Dönüştürme Tekniği
Sayma ile elde edilen değerler	$\sqrt{y}, \sqrt{y+0.5}, \sqrt{y+1}$
Oransal değerler	$\log it(p) = \left(\frac{1}{2}\right)\log\left(\frac{p}{q}\right), Arcsin\sqrt{p}$
Korelasyonlar	$Fisherz = \left(\frac{1}{2}\right) log\left(\frac{1+r}{1-r}\right)$
Aralıklı, Oransal ölçekli değerler	$ln(y), y^2, \frac{1}{y}, y^{(1/3)}, y^4$

Veri matrisinin çok değişkenli normal dağılıma yaklaştırılmasında uygun dönüştürme tekniğinin seçilmesi için grafiksel yöntemlerden yararlanmak uygun olmaktadır. Dönüştürülmüş ve orjinal verilerin grafiksel gösterimleri izlenerek normale yaklaşımın hangi dönüştürme yöntemi ile sağlandığı görsel olarak izlenmeli ve uygun olan testler yapılmalıdır.

İki değişkenli normal dağılım: Çok değişkenli normal dağılımın basit hali iki değişkenli normal dağılımdır. İki değişkenli normal dağılımın parametleri aşağıda verilmiştir:

$$\mu_1 = E(x_1) \quad \mu_2 = E(x_2)$$

$$\sigma_{11} = Var(x_1) \quad \sigma_{22} = Var(x_2)$$

$$\rho^{12} = Corr(x^1, x^2) = \frac{\sigma^{12}}{\sqrt{\sigma^{11}\sigma^{22}}}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad \Sigma^{-1} = 1/(\sigma_{11}\sigma_{12} - {\sigma_{12}}^2) \begin{bmatrix} \sigma_{22} & -\sigma_{21} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

p boyutlu normal dağılım için sabit yoğunluğun grafiği bir elipsoid olarak çizilmektedir. Bu elipsoid x ile belirlenen ve c^2 alanıyla açıklanan elipsoid olarak belirtilir.

$$c^2 = (x-\mu)'\sum^{-1}(x-\mu)$$

şeklinde hesaplanır.

Bu elipsoidin merkezi μ ortalamalar vektörüdür, eksenleri özdeğer ve özvektörler yardımı ile tanımlanmaktadır ve aşağıdaki gibi belirlenir:

$\pm c\sqrt{\lambda_i e_i}$ $\lambda_i e_i = \sum e_i$ i=1,2,...,p

Buradaki $\lambda_i\ i = 1,2,\ldots,p$ özdeğerleri, $e_i\ i.$ öz değere ilişkin öz vektörü ve $\sum$ kovaryans matrisini belirtir.

Çok değişkenli normal dağılımın birkaç özelliği aşağıda belirtilmiştir.

X çok değişkenli normal dağılıma ait rasgele bir vektör olsun;

1. X çok değişkenli normal dağılım bileşenlerinin doğrusal bileşenleri dahi çok değişkenli normal dağılmaktadır.

2. X çok değişkenli normal dağılım bileşenlerinin bütün altsetleri dahi çok değişkenli normal dağılım göstermektedir.

3. Sıfır kovaryansa sahip olan karşılıklı bileşenler birbirinden bağımsız dağılım göstermektedirler.

4. Bileşenlerin koşullu dağılımları çok değişkenli normal dağılım göstermektedirler.

5. $A\ p*p$ boyutlu simetrik matris ve $x, p*1$ boyutlu rasgele bir vektör olmak üzere;

a) $x'Ax = iz(x'Ax)$

b) $iz(A) = \sum\lambda_i$

şeklinde yazılır.

X veri matrisinin doğrusal bileşenlerinin de çok değişkenli normal dağılım gösterdiği kuralı aracılığıyla doğrudan veriler yerine bu verilerin doğrusal bileşenleri kullanılarak yapılan analizlerin de çok değişkenli normal dağılım varsayımı ile incelenmesini sağlamaktadır.

Çok değişkenli normal dağılım varsayımı aracılığı ile parametre tahminleri yapmak, değişkenlerin dağılımı hakkında kesin bir bilgi olmadığında önemli yardımcı rol oynamaktadır.

3.6.2. Çok değişkenli normal dağılım varsayımı

Çok değişkenli normal dağılımın sağlanıp sağlanmadığının araştırılmasında kullanılan direkt bir test bulunmamaktadır. Çok değişkenli normal dağılımın sağlanabilmesi için bütün değişkenlerin tek değişkenli normal dağılım göstermesi gerekmektedir. Ancak, bu durum yine de değişkenlerin çoklu normal dağılım gösterdiğini garanti etmemektedir. Çok değişkenli süreç kontrol yöntemlerini uygulamaya başlamadan önce değişkenlerin tekli ve çoklu normallik testlerine bağımlı tutularak normallik varsayımının sağlanması gerekmektedir (Orçanlı 2017). Tek değişkenli normallik testlerinde en çok kullanılan teknikler (Ülen 2010); Kolmogorov-Smirnov normallik testi, Shapiro-Wilks normallik testi, Q-Q grafik tekniği, Anderson-Darling normallik testidir. Değişkenlerin normal dağlıma uygunluğunun belirlenmesinde çoğunlukla kullanılan grafiksel yöntemler; Q-Q grafiği, P-P grafiği, histogram, kutu grafiği, dal-yaprak gösterimi grafiğidir (Newbold 2001).

Çok değişkenli normal dağılım verilerinde normallik varsayımının belirlenmesi, tek değişkenli verilerdeki gibi kolay olmamaktadır (Salmona 2004). Çok değişkenli normal dağılımın test edilmesinde birden fazla test çıkartılmış olmakla beraber çıkartılan bu testlerin sonucunda güvenilir bir sonuç elde edilememiştir. Veri kümesi çok değişkenli normal dağılıma uygun olmadığı zamanlarda tek değişkenli normal dağılmayan değişkenler araştırılır ve uygun dönüşümleri yaparak çok değişkenli normallik sağlanabilmektedir (Çilan 2005).

3.6.3. Çok değişkenli normal dağılıma uygunluk testleri

X veri matrisinin çok değişkenli normalliğe uygunluğu için çok fazla sayıda test geliştirilmiştir. Bu testlerin yaygın olarak kullanılanları Mardia tarafından geliştirilmiştir. Mardia tarafından çıkarılan çok değişkenli basıklık (multivariate Curtosis, MVNC_TEST) ve çok değişkenli çarpıklığa (Multivariate Skewness, MVNS_TEST) dayalı testler, çok değişkenli Shapiro-Wilk (MVNSHAPIRO) ve Doornik&Hanses Omnibus (D&H_OMNIBUS) testidir.

Mardia tarafından geliştirilen Skewness ve Curtosis bilinen en eski çok değişkenli normallik testleridir. 2009'da Villasenor Alva ve Estrada tarafından yapılan Monte Carlo simülasyon deneyimlerinde MVNSHAPIRO testinin diğer testlere göre daha güçlü bir çok değişkenli normallik testi olduğu belirlenmiştir.

Çok değişkenli Normal dağılıma uygunluk testi matris cebiri yaklaşımına göre MINITAB ve MATLAB programlarında yapılabilir.

3.7. Hotellıng T² Testi (Çok Değişkenli Hipotezlerin Testi)

Hotelling T^2 testi (kolay olması adına ilerleyen bölümlerde Hotelling T2 testi olarak belirtilecektir) çok değişkenli normal dağılım varsayımı için kurulan çok değişkenli hipotezlerin test edilmesini sağlayan bir yöntemdir.

Hotelling T2 testi $N_p(\mu, \Sigma)$ parametreli çok değişkenli normal dağılımdan tesadüfi bir şekilde çekilen veri matrisi aracılığıyla çok değişkenli hipotezleri test etmede kullanılan bir yöntemdir.

Hotelling T2 testi, bağımsız ve bağımlı olacak biçimde iki örneklem halinde değişken sayısı iki veya ikiden fazla olduğunda $(p \geq 2)$ örnek ortalama vektörlerinin farksız olduğunu veya ortalamalar fark vektörünün sıfır olduğuna dair hipotezlerin önemli olduğunu test etmekte kullanılan bir tekniktir.

Tek değişkenli parametrik hipotezlerin test edilmesinde olduğu gibi Hotelling T2 testinde de benzer biçimde çok değişkenli parametrik hipotezleri test edebilmek için üç farklı model bulunmaktadır.

Hotelling T2 Testi, aşağıda verilen çok değişkenli parametrik hipotezleri test eder:

- Parametreleri bilinen p değişkenli bir topluluğun ortalama vektörüne dayalı hipotezler: Bu model ile p değişkenli toplum ortalama vektöre dayalı hipotezler test edilmektedir. Bu model tek değişkenli parametirk hipotezlerin test edilmesinde kullanılan tek örnek t testinin çok değişkenli genellemesidir.

- Parametreleri bilinen bağımsız iki topluma ait çok değişkenli hipotezler: Bu model ile p değişkenli bağımsız 2 toplum ortalama vektörlerine dayalı hipotezler Hotelling T2 Testi aracılığı ile test edilmektedir. Bu model tek değişkenli parametirk hipotezlerin test edilmesinde kullanılan bağımsız iki örnek t testinin çok değişkenli genellemesidir.

- Parametreleri bilinen bağımlı iki topluma ilişkin çok değişkenli hipotezler: Bu model ile p değişkenli bağımlı iki toplum ortalama fark vektörüne dayalı hipotezler Hotelling T2 Testi aracılığı ile test edilmektedir. Bu model tek değişkenli parametrik hipotezlerin test edilmesinde kullanılan bağımlı iki örnek t testinin çok değişkenli genellemesidir.

3.7.1. Çok değişkenli toplum ortalama vektörüne dayalı hipotezlerin test edilmesi

Normal dağılım gösteren tek toplum parametresine dayanan hipotezleri test ederken ele alınan tek değişkenli sıfır (H_0) ve alternatif (H_1) hipotezleri

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$ test edilir.

$H_0: \mu = \mu_0$ hipotezinin $H_1: \mu \neq \mu_0$ hipotezine karşı test edilebilmesinde

$t = \left(\frac{\bar{x}-\mu}{\sqrt{\left(\frac{S^2}{n}\right)}} \right) t \sim t_{\alpha,n-1}$ test modelinden yararlanılmaktadır.

Çok değişkenli normal dağılım gösteren tek topluma ait H_0 ve H_1 hipotezleri ise aşağıdaki gibi kurulur.

$$H_0: \begin{bmatrix} \mu_1 \\ \vdots \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{01} \\ \vdots \\ \vdots \\ \vdots \\ \mu_{0p} \end{bmatrix} \qquad H_0: \begin{bmatrix} \mu_1 \\ \vdots \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} \neq \begin{bmatrix} \mu_{01} \\ \vdots \\ \vdots \\ \vdots \\ \mu_{0p} \end{bmatrix}$$

Çok değişkenli H_0 ve H_1 hipotezlerinin test edilebilmesi için tek değişkenli tek örnek t testi modelinin çok değişkenli genelleşmiş hali olan Hotelling T2 Testi çok değişkenli tek örnek formu kullanmaktadır.

$$T^2 = n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)$$

Buradaki n örnek hacmi, S örnek kovaryans matrisi, μ toplum ortalama vektörü ve x örnek ortalama vektörü belirtmektedir. Hoteling T2 istatistiği x ile μ arasında verilen istatistiksel aralığı göstermektedir ve bu ifade p boyuttaki gözlemlerin çok fazla farklı bileşenlerinin arasında bulunan istatistiksel aralığı gösterecek biçimde genişletilebilmektedir. Mesela, tek bir gözlem vektörü olan X ile kitle ortalaması μ veya μ nün varsayımı olan X arasında veya i-inci değişken ortalaması X_i ile bütün örneklem ortalaması X arasında bir istatistiksel açıdan uzaklık tanımlanabilmektedir. (Özkale 2004)

Hotelling T2 Test istatistiği T2'nin önemliliği açısıdan F dağılımından yararlanmaktadır. T2'nin F yaklaşım değeri aşağıdaki gibidir;

$$F = \left(\frac{n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)(n - p)}{(n - 1)p} \right)$$

şeklinde veya

$$F = \left(\frac{(n - p)}{(n - 1)p} \right) \times n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)$$

şeklinde hesaplanmaktadır.

F test istatistiği p, $(n - p)$ serbestlik derecesi olan teorik F dağılımını göstermektedir.

$$[F \approx F(\alpha, p, n - p)]$$

F değerinin aşağıdaki belirtilen kritik değerlerine göre T2 test istatistiği değerlendirilmektedir.

$F = \left(\frac{n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu)(n-p)}{(n-1)p}\right) < F(\alpha, p, n-p)$ olduğunda H_0 kabul edilir.

$F = \left(\frac{n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu)(n-p)}{(n-1)p}\right) \geq F(\alpha, p, n-p)$ olduğunda ise H_0 red edilir.

p değişkenli bir topluluğun parametre vektörlerine dayalı olarak kurulan H_0 ve H_1 hipotezlerini n birimlik örnekten elde edilmiş X_{n*p} matrisinin yardımı ile Hotelling T2 Testine göre test etmek için izlenecek aşamalar aşağıda verilmiştir.

1) X_{n*p} veri matrisinin çok değişkenli normal dağılıma uygunluğu test edilmektedir.

X veri matrisinin çok değişkenli normal dağılıma uygunluğu çok değişkenli Basıklık testi (Multivariate Curtosis, $\beta_{2,q}$) ve Mardia çok değişkenli Çarpıklık testine (Multivariate Skewness, $\beta_{1,q}$) göre belirlenmektedir.

2) X veri matrisinin kovaryans matrisi olan S ve ortalama vektörü olan X hesaplanmaktadır.

3) Örnek ortalama vektörüyle toplum ortalama vektörü arasındaki fark alınır ve bu fark vektörü (d) hesaplanır. $d = X - \mu$.

4) Kovaryans matrisinin tersi alınır. (S^{-1}).

5) T2 test istatistiği aşağıdaki verilen şekildeki gibi hesaplanır.

$T2 = n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)$ veya $T2 = n \times d'S^{-1} \times d$

6) T2'nin F dönüşümü elde edilir.

$$F = \left(\frac{T2(n-p)}{(n-1)p}\right)$$

7) F değeri $(p, n-p)$ serbestlik derecesi olan teorik F dağılımının $\alpha = 0.05, 0.01, 0.001$ kritik değerleriyle karşılaştırılmaktadır. $F < F_\alpha$ karşılaştırmasına göre karar verilmektedir.

Veri matrisinde değişkenlerin ölçü birimleri benzerlik göstermiş olsa da ölçüm değerlerinin ortalamaları ve değişim aralığı farklılık gösterebilmektedir. Oluşan bu farklılığı analize yansıtabilmek için ortalama vektöründeki farklılığı analize yansıtacak bir doğrusal bileşenler vektöründen yararlanılması gerekmektedir.

$S * a = (X - \mu)$ yaklaşımıyla a doğrusal bileşen vektörü $a = dS^{-1}$ şeklinde bulunmaktadır. Buradan test istatistiği $T^2 = n * d' * a$ şeklinde hesaplanmaktadır. Bu yaklaşım aracılığı ile elde edilen T^2 değeri direkt orjinal değerler kullanılarak yapılan hesaplama ile aynıdır.

3.7.2. Güven limitleri ve önemli değişkenlerin belirlenmesi

Çok değişkenli X gözlem matrisi $X \approx N_p(\mu, \Sigma)$ parametreli normal dağılım gösteriyorsa a doğrusal bileşen vektörü olmak üzere y vektörü $y = a'x \approx N(a'\mu, a'\Sigma a)$ parametreli çok değişkenli normal dağılımı göstermektedir. Bu halde test edilen hipotezler şu şekildedir;

$H_0(a): a'\mu = a'\mu_0 \quad H_1(a): a'\mu \neq a'\mu_0$

Bu yaklaşım aracılığı ile $a'\mu$ 'nün $1 - \alpha$ olasılıkla güven aralığı aşağıdaki gibidir.

$$\mathrm{P}\left[a'\bar{x} - \sqrt{\left(\frac{1}{n}\right) a'Sa \times F(\alpha, p, n-p)} \leq a'\mu \leq a'\bar{x} + \sqrt{\left(\frac{1}{n}\right) a'Sa \times F(\alpha, p, n-p)}\right]$$

$$= 1 - \alpha$$

Rasgele X gözlem değerlerinin alındığı toplumun $1 - \alpha$ güven ile her bir değişkenin ortalamalarını tahmin etmek mümkündür.

Değişken ortalamalarının güven aralığı yöntemi ile test edilmesi

Güven aralığı yaklaşımının önemli fonksiyonlarından biri de doğrusal bileşenler yardımı ile gözlem setinde yer alan değişkenlerin ortalamalarının güven aralıklarını belirlemektir. Belirlenmiş olan güven aralıkları, varsayılan toplumun ortalama vektöründe bulunan her bir değişken için ortalama parametresini içerip içermediği araştırılarak o değişkenin önemliliğe sebep olan bir değişken olup olmadığına ilişkin karar vermeye yardımcı olmaktadır. Örneğin; μ tahmini parametre vektöründeki güven

sınırları μ_0 parametre vektörü değerlerini kapsıyorsa örnek toplumun rasgele örneğidir. Eğer tahmini parametre vektöründeki bazı değişkenlerin güven aralıkları toplum parametre vektöründeki aynı değişkene ait değeri içermiyor ise o değişkenin önemli olduğu kabul edilmektedir. Bu kararlaştırma bütün değişkenler için tekrarlanmaktadır.

Güven limitlerinin belirlenmesi a doğrusal bileşenlerine göre her bir değişkenin alt sınır değeri ile üst sınır değeri aşağıdaki şekilde hesaplanmaktadır.

$$\mu_{1,2} = \left(a'\bar{x} \mp \left[\left(\frac{1}{n}\right) a'\bar{x}a \right]^{\left\{\left(\frac{1}{2}\right)\right\}} \times \left[\left(\frac{p(n-1)}{n-p} \right) * F(\alpha, p, n-p) \right]^{1/2} \right)$$

Değişken ortalamalarının t dağılımı yöntemi ile test edilmesi

Bilindiği üzere X matrisi $N_p(\mu, \Sigma)$ parametreli çok değişkenli normal dağılımdan çekilen bir örnek olduğundan ℓ doğrusal bileşenler vektörü olmak üzere;

$Z = \ell_1 x_1 + \ell_2 x_2 + \ldots + \ell_p x_p = \ell X$ şeklinde yazılabilmektedir.

$\mu_z = E(Z) = \ell'\mu \; ve \; var(Z) = \sigma_z^2 = \ell'\Sigma\ell$ olarak yazılmaktadır.

t dağılımı yardımıyla doğrusal bileşen ile belirlenen bir ortalama vektörü öğesinin güven aralığı;

$$t = \frac{\bar{z} - \mu}{\frac{S_Z}{\sqrt{n}}} = \frac{\sqrt{n}(\ell'\bar{x} - \ell'\mu)}{\sqrt{\ell S \ell}}$$

şeklindedir.

Ortalama vektöründe yer alan her bir değişken ortalamasının önemliliği;

$$|t| = \left| \frac{\sqrt{n}(\ell'\bar{x} - \ell'\mu)}{\sqrt{\ell S \ell}} \right| \leq t_{\{\alpha, n-1\}}$$

şeklinde veya

$$t^2 = \frac{n(\ell'\bar{x} - \ell'\mu)^2}{\ell' S \ell} = \frac{n\left(\ell'(x-\mu)\right)^2}{\ell' S \ell} \leq t^2_{\alpha, n-1}$$

şeklinde hesaplanır.

Bu test istatistiğinden faydalanılarak ortalama vektöründe yer alan hangi değişkenin önemli olduğuna karar verilmektedir. Aynı zamanda bu eşitlikten faydalanılarak ℓ doğrusal bileşeni yardımıyla ortalama vektöründeki her bir değişken ortalamasının $1-\alpha$ güven aralığı hesaplanmaktadır.

$$P\left(\bar{z}-t_{\alpha,n-1}\times\left(\frac{S_z}{\sqrt{n}}\right)\leq\mu_z\leq\bar{z}-t_{\alpha,n-1}\times\left(\frac{S_z}{\sqrt{n}}\right)\right)=1-\alpha$$

F dağılımı kritik değerlerine göre değişkenlerin önemliliğinin belirlenmesi

Ortalama vektörünün elemanlarına ait güven sınırlarının hesaplanabilmesinde başka bir yaklaşımda F dağılımının kritik değerlerinden faydalanmaktır. Bu yaklaşım sadece n örnek sayısının büyük olduğu durumlarda geçerlidir. Değişkenlere göre F yaklaşımı ile güven aralıkları aşağıdaki gibidir.

x_i ortlama vektörünün $i.$ değişkene ait elemanı s_{ii} değerleri kovaryans matrisinin köşegen elemanları ve $i.$ değişkenin varyansları olmak üzere aşağıdaki şekilde hesaplanır.

$$\mu_{i,1,2}=\bar{x}_i\mp\sqrt{\left(\frac{p(n-1)}{(n-p)}\right)F(\alpha,p,n-p)}\times\sqrt{\left(\frac{S_{ii}}{n}\right)}$$

Farklı değişken ortalamaları arasında önemliliğin test edilmesi

Çok değişkenli normal dağılımdan alınan n birimlik p değişkenli örneklerde $i.$ ve $j.$ değişken ortalamalarının arasındaki farkın önemliliğini test etmek gerekebilir. Değişken ortalamaları arasındaki güven sınırlarının belirlenmesinde F dağılımı yaklaşımından faydalanarak $\mu_i-\mu_j$ popülasyon ortalamaları farklarına bakılarak her iki popülasyondan alınan $i.$ ve $j.$ değişken ortalamaları arasındaki farkın güven aralığı aşağıdaki yaklaşım aracılığıyla belirlenir.

$$\mu_i-\mu_j=\left(\bar{x}_i-\bar{x}_j\right)\mp\sqrt{\left(\frac{p(n-1)}{(n-p)}\right)F(\alpha,p,n-p)}\times\sqrt{\left(\frac{S_{ii}-2S_{ij}+S_{kk}}{n}\right)}$$

3.7.3. Toplum ortalama vektörünün μ₀'ın sıfır olduğu durumlarda hotelling t² testi

Toplum ortalama vektörünün sıfır değeri aldığı $\mu_0 = [0,0,\ldots,0]$ veya ortalamalar arası fark vektörünün $\mu_D = [0,0,\ldots,0]$ olduğu hallerde Hotelling T2 testi aşağıdaki şekilde hesaplanır.

$$T2 = n\bar{x}'S^{-1}x \quad H_0: \mu_0 = [0,0,\ldots,0] \quad H_1: \mu_0 \neq [0,0,\ldots,0]$$

$$T2 = n(\bar{x}_d)'S^{-1}(x_d) \quad H_0: \mu\_\{D\} = [0,0,\ldots,0] \quad H_1: \mu\_\{D\} \neq [0,0,\ldots,0]$$

3.7.4. Çok değişkenli bağımsız iki topluma ilişkin hipotezlerin test edilmesi

Bağımsız iki topluma ait tek değişkenli hipotezler $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$ olarak verilir. Verilen bu hipotezlerin çok değişkenli genellemesi olan $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$ hipotezlerinin test edilmesi için Hotelling T2 testinden faydalanılır.

p değişkenli μ_1 ve μ_2 ortalama vektörüne ait çok değişkenli normal dağılım gösteren iki popülasyondan alınan n_1 ve n_2 birimlik örnek ortalama vektörlerinden faydalanarak $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$ veya $H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$ hipotezlerini test edebilmek için kullanılan Hotelling T2 testi modeli aşağıdaki gibidir:

$$T2 = \left(\frac{n_1 \times n_2}{n_1 + n_2}\right)(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2)$$

S ortak kovaryans matrisi ve S_1 birinci örnek kovaryans matrisi S_2 ikinci örnek kovaryans matrisleri aşağıdaki gibi tahmin edilmektedir.

$$S = \left(\frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2}\right)$$

T2 test istatistiği yaklaşık p, $(n_1 + n_2 - p - 1)$ serbestlik dereceli F dağılımı göstermektedir.

$$F = \left(\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)}\right)T2 \quad F \approx F\big(p, (n_1 + n_2 - p - 1)\big)$$

F değerinin önemliliği, $F(\alpha, p, (n_1+n_2-p-1))$ teorik dağılımın kritik değerleriyle kıyaslanarak bulunur.

Gruplar arasındaki çok değişkenli farklılığı test edebilmek amacıyla dört tane testten faydalanılır. Bu dört test şunlardır; Pillai İz Kriteri (Pillai Trace), Roy en büyük kök kriteri, Hotelling İz Kriteri (Hotelling Trace, Lawley-Hotelling Trace), Wilk's Lambda Kriteridir. Bu istatistiklerin hesaplanması ve önemliliğinin değerlendirilmesinde kullanılacak yaklaşımlar aşağıda verilmiştir. Her bir test istatistiği $H = BW^{-1}$ matrisinin özdeğerleri yardımıyla hesaplanmaktadır.

H: Hipotez matrisi ve genel kareler ve çapraz çarpım toplamı matrisi

B: Gruplararası KÇÇT matrisi

W: Hata KÇÇT matrisi

- Pillai Kriteri, H matrisinin sıfırdan büyük özdeğerleri yardımıyla aşağıdaki şekilde hesaplanır.

$$\sum_{i=1}^{s} \lambda_i/(1+\lambda_i)$$

(Pillai 1967)

Burada s sıfırdan büyük olan öz değer sayısını göstermektedir.

T nin önemliliğini belirtmek için Pillai Kriteri'nde F yaklaşımından faydalanılmaktadır.

$$F = \left(\frac{(n_e - p - s)T}{\big(b(s-T)\big)}\right)$$

n_e: Hata KÇÇT matrisi serbestlik derecesi,

n_h: Hipotez KÇÇT matrisi serbestlik derecesi

$b = max(p, n_h)$ p: Değişken sayısı

s: Sıfırdan büyük özdeğer sayısı

F değerinin önemliliği $b,s(n_e - p + s)$ serbestlik dereceli teorik F dağılımının kritik değerleri aracılığı ile belirlenir.

• Hotelling İz Kriteri HT istatistiği, H matrisinin sıfırdan büyük özdeğerleri kullanılarak aşağıdaki şekilde hesaplanır.

$$HT = \sum_{i=1}^{s} \lambda_i$$

HT nin önemliliğinin belirlenmesi için aşağıdaki F yaklaşımı kullanılmaktadır.

$$F = \left(\frac{2(sn+1)HT}{s^2(2m+s+1)} \right)$$

Burada $n = \frac{|n_e - p| - 1}{2}$ ve $m = \frac{|n_h - p| - 1}{2}$ şeklinde hesaplanmaktadır.

F değerinin önemliliği $s(2m+s+1)$, $2(sn+1)$ serbestlik dereceli teorik F dağılımının kritik değerleri aracılığı ile belirlenir.

• Wilk's Lambda Kriteri (L), H matrisinin sıfırdan büyük özdeğerlerinden faydalanılarak aşağıdaki gibi hesaplanır (Rao 1973).

$$L = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

L nin önemliliğinin belirlenmesi için aşağıdaki F yaklaşımı kullanılmaktadır.

$$F = \left(\frac{\left(1 - L^{1/\ell}\right)\left(M\ell + 1 - \frac{n_{\{h\}p}}{2}\right)}{L^{1/\ell_{nhp}}} \right)$$

F değerinin önemliliğinin belirlenmesinde n_{hp} ve $\left(m\ell + 1 - \frac{n_{hp}}{2}\right)$ serbestlik dereceli teorik F dağılımının kritik değerlerinden faydalanılmaktadır.

Buradaki ℓ²=(p²n_{h}²-4)/(p²+n_{h}²-5) ve M=n_{e}-(p+1-n_{h})/2 şeklinde hesaplanmaktadır.

• Roy En büyük kök kriteri (R) ise aşağıdaki gibi hesaplanır.

$$R = \lambda_1/(1+\lambda_1)$$

3.7.5. Hotelling T2 istatistiğinin dağılım özellikleri

Ayrı ayrı hepsinin m büyüklükte p değişkeni kapsayan n tane birim olduğunu ve bu birimlerin Σ kovaryans matrisli, μ ortlama vektörlü bir normal dağılımdan edinildiğini kabul edelim. Hesaplamalarda zıttı belirtilmediği sürece p bileşeninin üzerinde tek bir gözlemin olduğu ve $m = 1$ varsayılmaktadır.

Hoteling T^2 testi istatistiğini açıklamada değişik olasılık fonksiyonları kullanılabilmektedir. Aşağıda üç farklı yaklaşım ele alınmıştır.

1) Çok değişkenli normallik dağılımına ait olan μ ve Σ'nin bilindiğini varsayalım.

Her bir gözlemin vektörü X'e ait olan T^2 istatistiği aşağıdaki gibi

$$T^2 = n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \sim x_p^2$$

olup x_p^2, p serbestlik dereceli ki-kare dağılımını göstermektedir. T^2 dağılımı yalnızca X gözlem vektöründeki belirtilen değişken sayısı olan p ye bağlıdır.

2) Çok değişkenli normallik varsayımı içinde, anakütle μ ve Σ parametreleri bilinmeden X ve S varsayımlarını kullanarak hesapladığımızı farzedelim. X ve S değerleri n gözlemi kapsayan önceden verilen veri kümesinden bulunur.

$$\bar{X} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} X_i$$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

$\bar{X}$ ve S den bağımsız gözlem vektörü X için T^2 istatistiği aşağıdaki gibi

$$T^2 = n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) - \left[\frac{p(n+1)(n-1)}{n(n-p)}\right] F_{(p,n-p)}$$

olup p ve $(n-p)$ serbestlik dereceli F dağılımına sahiptir.

3) Hotelling T^2 testinin dağılımı, geçmişte kontrol altına alınan bir süreçten raslantı olarak belirlenen p değişkenine ve n gözlemine bağlıdır. μ ve Σ nin belli olamayıp X ve S ile düşünüldüğünü X gözlem vektörünün X ve S den ayrı olmadığını farzedelim. Bundan dolayı T^2 istatistiğinin dağılımı ve yapısı şu şekildedir:

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) - \left[\left(\frac{(n-1)^2}{n}\right)\right] B_{(\frac{p}{2}, \frac{n-p-1}{2})}$$

3.8. Varyans Analizi (Anova, Manova)

Çok değişkenli varyans analizi iki veya ikiden fazla ortalamanın birbirinden farklı olup olmadığıyla ilgili hipotezleri test etmede kullanılan bir yöntemdir. İki ortalamanın arasında anlamlı bir farkın olup olmadığını test edebilmek için kullanılan diğer bir yöntemde t testidir. Ancak t testi ikiden çok ortalamanın karşılaştırıldığı durumlarda sorun oluşturmaktadır. Ancak, ikiden çok ortalamanın karşılaştırıldığı durumlarda da, ortalamaları ikişer ikişer ortalamaları t testiyle kıyaslamak muhtemel olsa dahi, yapılan bu yöntem 1. tip hata oranının fazla bir şekilde artmasına neden olacaktır. Varyans 1. tip hata oranını arttırmada ikiden çok ortalamanın karşılaştırılmasında kullanılan bir testtir. Varyans analizinde H_0 hipotezine göre, tüm popülasyonların ortalamaları eşit kabul edilebilir.

$H_0: \mu_1 = \mu_2 = \ldots = \mu_n\}$ Ortalamalar eşittir aralarında fark yoktur.

H_a: Ortalamaların minimum ikisi arasında anlamlı fark vardır.

H_0 hipotezi, iki varyans tahminini karşılaştırma yoluyla test edilmektedir. Birinci varyans, grupların içindeki varyanstır (Mean square error MSE). MSE H_0 hipotezi eğer doğru olsun ya da olmasın, varyansın tahmini olmaktadır. İkinci varyans, grupların ortalamalarının varyansları üzerine yerleştirilmiştir (Mean Square Between MSB). MSB, sadece H_0 hipotezi eğer doğruysa varyansın tahmini olmaktadır. H_0 hipotezi eğer yanlışsa MSB varyanstan daha büyük bir rakam varsaymaktadır. Varyans analizindeki hipotez testi şu şekilde yazılır:

MSE ile MSB yaklaşık eşit ise → H_0 hipotezi doğru

MSB, MSE' den büyük ise → H_0 hipotezi yanlış

Varyans analizinde asıl amaç ortalamaların arasındaki farkın varlığını belirlemektir.

Varyans analizinde hipotezi test edebilmek için aşağıdaki F değeri kullanılmaktadır;

$$F = \left(\frac{MSB}{MSE}\right)$$

Eğer F değeri, belirtilen anlamlılık düzeyinde tablodaki değerden küçükse H_0 hipotezi reddedilmez. Bu durumda ortalamaların arasında anlamlı fark yoktur.

F değeri, eğer tablodaki değerinden büyükse H_0 hipotezi reddedilir. Bu durumda ortalamaların arasında anlamlı bir fark vardır.

Varyans analizinde bağımlı ve bağımsız değişkenlerden bahsedilmektedir. Bağımsız değişkenlerin, bağımlı değişkenler üzerinde oluşturduğu etkiler araştırılır. Bağımsız değişken kategorik olması gerekirken, bağımlı değişken ise metrik olmalıdır.

3.8.1. Tek yönlü varyans analizi (anova)

Tek yönlü Anova'da esas olan iki varsayım bulunmaktadır. Bu varsayımlara göre gruplar normal dağılımdan gelmekte ve bu grupların varyansları homojen olmaktadır. Çalışmalarda genelde varyansların homojenliğine bakılır. Ancak varyanslar homojense varsayımların tamamının sağlandığı kabul edilir.

Normal dağılım gösteren bağımsız g grup için $(g > 2)$ tek değişkenli hipotezlerin test edilebilmesinde tek yönlü varyans analizinden (TYANOVA) faydalanılır. TYANOVA'da bağımsız g topluluğuna ait kurulan H_0 ve H_1 hipotezleri test edilebilmektedir. TYANOVA'da H_0 hipotezi g toplum ortalamalarının birbirine eşit olduklarını varsayarken, H_1 hipotezi ise g toplum ortlamasından en az birinin diğerlerinden farklı olduğunu varsaymaktadır.

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_g$$

$$H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_g$$

H_1:"En azından bir populasyon ortalaması diğerlerinden farklıdır."

3.8.2. İki yönlü anova

Bağımsız olan iki değişkenin, bir bağımlı değişkenin üzerine olan etkisi araştırılırken, bunu tek tek araştırmaktansa, ikisini tek bir işlem de birleştirmek daha verimli olamktadır. Böyle bir yaklaşım, bağımlı değişkenler üzerinde, bağımsız olan değişkenlerin tek tek etkilerini hesaplamanın yanı sıra, bağımsız değişkenin birbirleri arasında oluşan etkileşimi de düşünmek gerekir.

3.8.3. Çok değişkenli varyans analizi (manova)

Çok değişkenli varyans analizi (MANOVA) iki ve ikiden fazla bağımsız ve bağımlı değişkenlerde çok değişkenli normallik dağılımına dayalı olan hipotezleri belirlemek için geliştirilen bir tekniktir. Grup sayısı ikiden çok olduğu zaman T2 yöntemi çalışmamaktadır ve MANOVA aracılığı ile analizler yapılmaktadır. MANOVA, Hotelling T2 testinin fonksiyonlarını da yapan bir tekniktir.

MANOVA üç farklı şekilde kurulan hipotezleri test etmektedir. MANOVA test edilen hipotez tipine göre farklı isimler ile hatırlanmaktadır. Bunlar aşağıdaki gibi verilebilir.

- İki veya ikiden fazla $(g \geq 2)$ bağımsız grupta çok değişkenli normal dağılımını gösteren toğluluğa ait kurulan hipotezlerin test edilebilmesinde Tek Yönlü Çok Değişkenli Varyans Analizi (TYMANOVA) kullanılır.

- İki veya ikiden fazla $(g \geq 2)$ bağımlı grupta çok değişkenli normallik dağılım gösteren topluluğa ait kurulan hipotezlerin test edilebilmesinde İki Yönlü Çok Değişkenli Varyans Analizi (IYMANOVA) kullanılır.

- İki veya ikiden fazla bağımsız veya bağımlı çok değişkenli çok faktörlü denemelerin sonuçlarının analizinde Faktöriyel MANOVA (FMANOVA) tekniği kullanılmaktadır.

Aşağıda grup sayısı $g > 2$ ve faktör sayısı $f \geq 2$ için manova türleri açıklanacaktır.

Tek yönlü çok değişkenli varyans analizi (tek yönlü manova, TYMANOVA)

TYMANOVA yöntemi çok değişkenli normal dağılımı belirten g toplum ortalama vektörlerine ait hipotezleri belirtmektedir. TYMANOVA'da H_0 hipotezi g toplum ortalama vektörlerinin birbirlerine eşit olduğunu varsayarken, H_1 hipotezi ise g

toplum ortlama vektörlerinden en az birisinin diğerlerinden farklı olduğunu belirtmektedir. Bu hipotezler aşağıdaki şekilde verilebilir.

$$H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_g \end{bmatrix} = \begin{bmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \vdots \\ \mu_{0g} \end{bmatrix} \qquad H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_g \end{bmatrix} \neq \begin{bmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \vdots \\ \mu_{0g} \end{bmatrix}$$

Bilindiği üzere tek değişkenli g toplumuna ait oluşturulan hipotezler, toplumlardan rasgele çekilen g örnek grup verilerine dayanarak test edilmektedir. MANOVA'da da temel şart gruplarda veri matrislerinin çok değişkenli normal dağılım göstermektedir.

Tek değişkenli g bağımsız grupta $i.$ grup ve $j.$ birime ait gözlem değeri; $X_{ij} = \mu + \alpha_i + e_{ij} \quad i = 1,2,\dots g; \quad j = 1,2,\dots n_i$ şeklinde belirlenir. Buradaki μ genel ortalamayı, α_i $i.$ grup etkisini ve e_{ij} rasgele hatayı belirtir.

Rasgele hata vektörü e_{ij}; $e_{ij} \sim N_p$ (0,Σ) parametreli çok değişkenli normal dağılımı göstermektedir.

α_i grup etkilerini belirtmektedir ve şu özelliği taşımaktadır

$$\sum_{i=1}^{g} n_i \alpha_i = 0$$

TYMANOVA'da "kovaryans matrisinin tüm toplumlarda benzer olduğu, örneklerin alındığı varsayılan toplumların ortak kovaryans matrisli çok değişkenli normal dağılım gösterdiği" varsayılmaktadır.

TYANOVA'daki yaklaşıma benzer biçimde, TYMANOVA'da da g gruptan elde edilen p değişkenli gözlem vektörünün elemanlarının genel ortalamadan olan farkları iki bileşene ayrılır;

$$x_{ij} - \bar{x} = (\bar{x}_i - x) + (\bar{x}_{ij} - \bar{x}_i)$$

I II

Bu eşitlikteki I. eleman grup ortalama vektörünün genel ortalama vektöründen farkını göstermekte ve grupların arasındaki etki farklılığını belirtir. II. eleman her gruptaki gözlemler vektöründeki elemanların kendi grup ortalama vektöründen ayrılışlarını belirtir ve hatları belirtir.

Yukarıdaki bu bileşenler Kare ve Çarpımlar Toplamı terimleri cinsinden ele alınırsa, gözlemler ile genel ortalama vektörü arasında olan genel değişimi iki bileşene ayırarak MANOVA uygulaması yapılabilmektedir.

I. eleman aracılığı ile belirlenen değişim kaynağı gruplar arası kareler ve çarpımlar toplamı matrisiyle (B) belirlenir.

II. eleman ise Hata Kareler Toplamı Toplamı ve Çapraz Çarpımlar Toplamı matrisi (W) aracılığı ile belirlenir. Toplam değişim $T = B + W$ biçiminde belirlenir.

- Gruplar arası değişimi gösteren B matrisi aşağıdaki şekilde hesaplanır;

$$B = \sum_{i=1}^{g} n_{i(\bar{x}_i - \bar{x}) - (\bar{x}_i - \bar{x})'}$$

- Gruplariçi değişimi gösteren W matrisi aşağıdaki gibi hesaplanır;

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$$

- Toplam değişimi gösteren T matrisi $T = B + W$ biçiminde hesaplanır.

- Gruplararası serbestlik derecesi ise $sd_1 = g - 1$ şeklinde belirlenir.

- Gruplariçi serbestlik derecesi ise $sd^2 = \sum_{i=1}^{g} n_i - g$ şeklinde belirlenir.

- Toplam serbestlik derecesi gsd $sd^2 = \sum_{i=1}^{g} n_i - 1$biçiminde hesaplanır.

- Genel değişimi belirten terimler ve bu değişimin, değişim kaynaklarına göre bileşenleri Çok Değişkenli Tek Yönlü Varyans Analizi tablosunda aşağıdaki gibi

verilmiştir. Bu tabloda $g > 2, p \geq 2$ için veri matrisinin istatisitikleri test edilebilmektedir.

- Gruplararası değişimin hataya göre önemliliğinin test edilebilmesinde Wilk's Lambda istatistiği kullanılır. Wilk's Lambda (Λ) aşağıdaki gibi hesaplanır.

$$\Lambda = L = \left(\frac{|W|}{|B+W|}\right) = \left(\frac{\left|\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'\right|}{\left|\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})'\right|}\right)$$

- L nin önemliliği p değişken ve g grup sayısına bağlı olarak farklı şekillerde değerlendirilse de toplam birim sayısı $N(\Sigma n_i)$ yeterince büyükse aşağıdaki gibi hesapanır.

$$LL = -\left(N - 1 - \frac{(p+g)}{2}\right) ln\left(\frac{|W|}{|B+W|}\right)$$

- L istatistiğinin önemliliği, gruplardaki birim sayılarına göre değerlendirilmektedir. $n_i \geq 10$ olduğunda Kikare yaklaşımı, $n < 10$ olduğunda ise F dağılımı yaklaşımı tercih edilmektedir.
- $n_i \geq 10$ ise LL test istatistiği, $p(g-1)$ serbestlik dereceli Kikare dağılımı göstermektedir ($LL \approx \chi^2(p(g-1))$). LL'nin önemliliği ise $\chi^2(p(g-1))$ dağılımının kritik değerlerine göre belirlenmektedir.
- $n_i < 10$ 0 olduğunda ise LL test istatistiği aşağıdaki şekilde hesaplanır

$$LL = \left(\frac{\sum n_i - p - 2}{p}\right)\left(\frac{1-\sqrt{L}}{\sqrt{L}}\right)$$

ve $2p, 2(N-p-2)$ serbestlik dereceli F dağılımı göstermektedir

($LL \approx F(2p, 2(\sum n\_\{i\} - p - 2))$

- $n_i < 10$ ise LL test istatistiği $F(\alpha, 2p, 2(N-p-2))$ teorik dağılımının kritik değerlerine göre belirlenmektedir. Buradaki $N = n_1 + n_2 + \ldots + n_g\}$ şeklinde hesaplanan genel gözlem sayısıdır.

TYMANOVA uygulamasında başka bir temel varsayım ise grup kovaryans matrislerinin homojen olması şartıdır. Grup kovaryans matrislerinin homojenliği Box's M testi aracılığı ile belirlenir.

Box's M testi aşağıdaki şekilde hesaplanır.

$M = \sum_{i=1}^{g}(n_i - 1)ln|S| - \sum_{i=1}^{g}(n_i - 1)ln|S_i|$ Verilen bu eşitlikte S, ortak kovaryans matrisini, S_i her bir grubun kovaryans matrisini, $|S_i|$ grup kovaryans matrislerinin determinantını belirtmektedir. S ortak kovaryans matrisi örnek kovaryans matrislerinden faydalanılarak aşağıdaki şekilde hesaplanır.

$$S = \left(\frac{\sum_{i=1}^{g}(n_i - 1)S_i}{\sum_{i=1}^{g}(n_i - 1)}\right) = \left(\frac{\sum_{i=1}^{g}(n_i - 1)S_i}{N}\right)$$

İki yönlü çok değişkenli varyans analizi (IYMANOVA)

IYMANOVA $i.$ işlemde $j.$ birimin gözlenen değeri; $X_{ij} = \mu + \alpha_T + \beta_B + e_{ij}$ olarak ele alınır.

Buradaki μ genel ortalamayı, α_T işlem etkisini, β_B birim etkisini ve e_{ij} rasgele hatayı belirtir. Modelde birim ve işlem arasında bir etkileşimin olmadığı varsayılmaktadır.

IYMANOVA, IYANAOVA'nın $p \geq 2$ için çok değişkenli bir genellemesidir.

$p \geq 2$ değişken sayısı, j birim sayısı, i işlem sayısı olmak üzere bağımlı gözlem aşağıdaki gibi ifade edilir;

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

Buradaki μ genel ortalama vektörünü, α_i işlem etkilerini, β_j birim etkilerini, $\alpha\beta_{ij}$ birim ile işlem etkileşimini ve e_{ijk} rastgele hatayı belirtir. $i = 1,2,\ldots,t;\ j = 1,2,\ldots,b;$ $k = 1,2,\ldots,n$ olacak şekilde değerler almaktadır.

İki yönlü bir deneme sonucuyla genel ortalama arasındaki fark şu şekildedir;

$$X_{ijk} - \mu = \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

Bu ifadeyi karesel terimler cinsinden açık bir şekilde aşağıdaki gibi yazabiliriz.

$$X_{ijk} - \mu = \sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk} - \bar{X}\right)\left(X_{ijk} - \bar{X}\right)'$$

$'+= \sum_{i=1}^{t} b_n\,(\bar{X}_{i.} - \bar{X})(\,\bar{X}_{i.} - \bar{X})' + \sum_{\mathrm{j}=1}^{\mathrm{b}} \mathrm{tn}\left(\bar{X}_{j.} - \bar{X}\right)\left(\,\bar{X}_{j.} - \bar{X}\right)' +$

$$+\sum_{i=1}^{t}\sum_{j=1}^{b} n\left(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{j.}+\bar{X}\right)\left(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{j.}+\bar{X}\right)'$$

$$+\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{n} n\left(\bar{X}_{ijk}-\bar{X}_{.j}\right)(\bar{X}_{ijk}-\bar{X}_{ik})'$$

Burada yer alan 4 elemandan her biri genel çeşitliliğin birer bileşeni olan birim etkisini, işlem etkisini, hata değişimini ve birim ile işlem etkileşimini belirtmektedir. Bu bileşenleri Kareler Toplamları (KT) şeklinde aşağıdaki gibi hesaplayabiliriz.

- İşlem KT= $\sum_{i=1}^{t} b_n\,(\bar{X}_{i.}-\bar{X})(\,\bar{X}_{i.}-\bar{X})'$
- Birey KT= $\sum_{\mathrm{j}=1}^{\mathrm{b}} \mathrm{tn}\left(\bar{X}_{j.}-\bar{X}\right)\left(\,\bar{X}_{j.}-\bar{X}\right)'$
- Birim×İşlem KT= $\sum_{i=1}^{t}\sum_{j=1}^{b} n\left(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{j.}+\bar{X}\right)\left(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{j.}+\bar{X}\right)'$
- Hata KT= $\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{n} n\left(\bar{X}_{ijk}-\bar{X}_{.j}\right)\left(\bar{X}_{ijk}-\bar{X}_{ik}\right)'$

KÇÇT matrislerinden faydalanarak; işlem etkilerinin önemliliği, birim×işlem önemliliği, ve birim etkilerinin önemliliğine ait hipotezler test edilmektedir. İlk olarak etkileşim olup olmadığına ait hipotezler test edilir.

H_0: $\alpha\beta_{ij}$=0; H_1: En azından bir etkileşim $\alpha\beta_{ij} \neq 0$

$$\Lambda_{INT} = \left(\frac{|K\text{ÇÇ}T_{RES}|}{|K\text{ÇÇ}T_{INT} + K\text{ÇÇ}T_{RES}|}\right)$$

biçiminde hesaplanır.

Λ_{INT} 'nın önemliliği için;

$$LL_{\{INT\}} = -\left[tb(n-1) - \frac{p+1-(t-1)(b-1)}{2}\right] ln\Lambda > \chi^2_{a,(t-1)(b-1)p}$$

İşlem etkisinin önemliliği için;

$$\Lambda_{TR} = \left(\frac{|K\text{ÇÇ}T_{RES}|}{|K\text{ÇÇ}T_{TR} + K\text{ÇÇ}T_{RES}|} \right)$$

Λ_{TR}'nın önemliliği için

$$LL_{\{TR\}} = -\left[tb(n-1) - \frac{p+1-(t-1)}{2} \right] ln\Lambda > \chi^2_{a,(t-1)p}$$

Birim etkisinin önemliliği için;

$$\Lambda_{BR} = \frac{|K\text{ÇÇ}T_{\{RES\}}|}{|K\text{ÇÇ}T_{\{BR\}} + K\text{ÇÇ}T_{\{RES\}}|}$$

Λ_{BR}'nın önemliliği için

$$LL_{BR} = -\left[tb(n-1) - \frac{p+1-(b-1)}{2} \right] ln\Lambda > \chi^2_{a,(b-1)p}$$

yaklaşımlarından faydalanılır.

Yukarıda verilen modellerde birim ve işlem yerine iki faktör ele alınarak faktöriyel MANOVA yapılabilmektedir.

Eğer modelde etkileşim yoksa bu durumda etkileşim kaynağına ait KÇÇT matrisi $K\text{ÇÇ}T_{RES}$ içinde yer alır. Böylelikle interaksiyon önemliliği test edilemez.

3.9. Çok Değişkenli Doğrusal Regresyon Analizi

Regresyon analizi aralarında bağlantı olan iki veya ikiden fazla değişkenlerden birisinin bağımlı değişken diğerinin bağımsız değişkenler olacak şekilde ayrımıyla bu değişkenler arasındaki ilişkiyi matematiksel eşitlik ile inceleyen bir yöntemdir.

Regresyon analizi modeldeki bağımlı değişken sayısına göre farklı şekillerde isimlendirilir:

- Basit Regresyon: Bağımlı bir Y değişkeniyle bağımsız bir X_1 değişkeni arasındaki ilişkiyi araştıran teknik basit regresyon olarak adlandırılır. Basit regresyon modeli $Y = f(X)$ şeklinde yazılır.

• Çoklu Regresyon: Bağımlı bir Y değişkeniyle iki veya ikiden fazla bağımsız $X^1, X^2, \ldots, X_P$ değişkeni arasında bulunan ilişkiyi araştıran teknik çoklu regresyon olarak adlandırılır. Çoklu regresyon modeli $Y = f(X_1, X_2, \ldots, X_p)$ şeklinde yazılır.

• Çok Değişkenli Regresyon: En az iki bağımlı değişken ve en az bir bağımsız değişken arasındaki bağıntıyı inceleyen yönteme çok değişkenli regresyon denir. Çok değişkenli regresyon modeli $(Y_1, Y_2, \ldots, Y_q) = f(X_1, X_2, \ldots, X_p)$ şeklinde yazılır.

Bağımlı değişkenle bağımsız değişkenler arasındaki ilişkiyi araştıran modeller doğrusal bağıntı şeklindeyse bu tipteki regresyon modellerine doğrusal (linear) regresyon yöntemi denir. Bağımsız değişkenlerle bağımlı değişkenler arasında bulunan ilişki eğer doğrusal değil ise bu tipteki regresyon modellerine doğrusal olmayan (nonlinear) regresyon yöntemi denir.

Aşağıda doğrusal model çeşitleri verilmiştir;

$Y = \beta_0 + \beta_1 X_1 + e_{ij}$ Basit doğrusal model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e_{ij}$ Çoklu doğrusal model

$(Y_1, Y_2, \ldots, Y_q) = f(X_1, X_2, \ldots, X_p)$ Çok değişkenli doğrusal model

Çoklu regresyon modeli çok değişkenli regresyon modeli olarak ifade edilemez. Yukarıda verilen model açıklamalarında da olduğu gibi çoklu regresyon modelinde bir bağımlı, iki veya daha fazla bağımsız değişken vardır. Çok değişkenli modelde ise en az iki bağımlı, bir veya birden fazla bağımsız değişken vardır.

3.9.1. Basit doğrusal regresyon

Basit doğrusal regresyon tekniğinde bir Y bağımlı değişkeniyle bir X bağımsız değişkeni arasındaki matematiksel bağıntı incelenmektedir.

$Y = \beta_0 + \beta_1 X + \varepsilon$ şeklinde ifade edilen modele basit doğrusal regresyon modeli denir.

β_0: doğrunun y-eksenini kestiği nokta

β_1: doğrunun eğimi

ε: şansa bağlı hata terimi

Basit doğrusal regresyon teknikte bağımlı ve bağımsız değişken arasındaki işlevsel bağıntının matris biçimi aşağıdaki gibidir. Y=βX+ε

Bu teknikte; $Y, (n \times 1)$ boyutlu bağımlı değişken vektörünü, X, $(n \times (p+1))$ boyutlu bağımsız değişken matrisini, β, $((p+1) \times 1)$ boyutlu katsayılar vektörünü ve ε, $(n \times 1)$ boyutlu rasgele hata vektörünü belirtmektedir.

Regresyon teknikleri için değişkenlerin uyması gereken bazı önemli şartlar bulunmaktadır. Önemli şartlar aşağıdaki şekilde verilmiştir.

1- Y normal dağılım göstermelidir.

2- X gözlemleri birbirinden bağımsız olmalıdır.

3- X hatasız ölçümleri içermelidir.

4- $e, N(0, \Sigma)$ parametreli normal dağılım göstermelidir.

5- Hata terimleri arasındaki kovaryans sıfır olmalıdır $\left(Cov(e_i, e_j) = 0\right)$.

Basit doğrusal regresyon modeli matris formunda aşağıdaki gibi ifade edilir.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

şeklinde veya $Y = X\beta + \varepsilon$ biçiminde yazılır.

Bu modeldeki $E(\varepsilon) = 0$ ve $Cov(\varepsilon) = \sigma^2 I$ dir.

Bu modelde Y'nin beklenen değeri $E(Y) = X\beta$ şeklinde belirtilir.

Modelde yer alan $p+1$ regresyon katsayısı En küçük Kareler Yöntemine göre şu şekilde tahmin edilir.

$$b = (X'X)^{-1}X'Y$$

3.9.2. Çoklu doğrusal regresyon

Çoklu doğrusal regresyon bağımlı bir Y değişkeniyle iki ve ikiden fazla bağımsız $(X_1, X_2, \ldots, X_p)$ değişkeni arasındaki doğrusal bağıntıyı inceleyen bir tekniktir.

Çoklu regresyon modelinde bağımlı değişkenle bağımsız değişken arasında bulunan işlevsel ilişki aşağıdaki matris formu ile gösterilir.

$$Y = X\beta + \varepsilon$$

$$Y: (n \times 1)$$

boyutlu bağımlı değişken gözlem vektörü

$X: (n \times (p+1))$ boyutlu bağımsız değişkenler gözlem matrisi

$\beta: ((p+1) \times 1)$ boyutlu katsayılar vektörü

ε: (n×1) boyutlu rastgele hata vektörü

Verilere çoklu regresyonun uygulanabilmesi için, genel regresyon uygulamaları için verilerin uyması gereken koşullara ek olarak bağımsız değişkenlerin arasında çoklu bağımlılık olmaması gerekmektedir.

Gözlemlere göre yukarıdaki p değişkenli çoklu doğrusal regresyon modeli aşağıdaki gibidir.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \left[\begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_3 \end{bmatrix} \right]$$

veya $Y = X\beta + \varepsilon$ şeklinde yazılır.

Bu modeldeki $E(\varepsilon) = 0$ ve $Cov(\varepsilon) = \sigma^2 I$ dir.

Böylece Y'nin beklenen değeri $E(Y) = X\beta$ şeklinde belirlenir.

Modelde yer alan p+1 regresyon katsayıları Enküçük Kareler Yöntemi ile aşağıdaki şekilde tahmin edilir.

$b = (X'X)^{-1}X'Y$Bu yaklaşımda p+1 sayıda regresyon katsayısına ait tahminler elde edilmektedir $(b_0, b_1, \ldots, b_p)$.

Çoklu doğrusal regresyon tekniğinde parametre tahminleri ve Kareler Toplamları basit doğrusal regresyon bölümünde kullanılan matris formundaki formüller ile aynıdır.

3.9.3. Regresyon katsayılarının önemliliğinin test edilmesi

Regresyon analizinde regresyon denkleminin Y'yi açıklamadaki yeterliliğinin yanı sıra regresyon katsayılarının sıfırdan önemli ölçüde farklılık gösterip göstermediğinin test edilmesi de gerekmektedir.

Basit ve çoklu doğrusal regresyon modellerinde H₀:β_{i}=0 hipotezinin H₁:β_{i}≠0 hipotezine karşı test edilebilmesi söz konusu iken çok değişkenli regresyon modelinde $H_0: \beta_{ij} = 0$ hipotezi $H_1: \beta_{ij} \neq 0$ hipotezine karşı test edilir.

Basit ve çoklu regresyon modelinde regresyon katsayılarının önemliliğini belirlemek için her bir parametre tahmininin varyansını hesaplamalıyız. Regresyon katsayılarının varyansları $var(b_i)$ $s^2(X'X)^{-1}$ matrisinin diyagonal elmanlarından oluşmaktadır.

Her bir regresyon katsayısının önemliliğinin belirlenmesinde t testinden faydalanılır. t testi modeli aşağıda verilmiştir.

$$t_i = \left(\frac{b_i - \beta_i}{\sqrt{var(b_i)}}\right) \quad sd = n - 1$$

t_i test istatistiğinin önemlililiğini belirlemede $t(\alpha, sd)$ kritik değerlerinden yararlanılır.

$H_0: \beta_0 = 0$ hipotezinin test edilmesinin amacı doğrunun merkezden geçip geçmediğini belirlemektir. Eğer doğru merkezden geçiyorsa yani H_0 hipotezi kabul ediliyor ise o zaman modele sabit katsayının eklenmesine gerek yoktur. Ama doğru merkezden geçmiyorsa yani H_0 hipotezi reddediliyor ise o zaman modele

b_0 eklenmelidir. b_0 testi aracılığı ile modelde sabite yer verilip verilmeyeceğine karar verilir. Hipotez gereği doğru merkezden geçer anlamına $H_0: \beta\_\{i\} = 0$ gelir.

$H_0: \beta_1 = 0$ hipotezinin test edilmesinin amacı regresyon doğrusunun eğiminin sıfır olup olmadığının belirlenmesidir.

Çoklu regresyon modelinde regresyon katsayılarının önemliliği, basit doğrusal regresyonda izlenen yönteme benzer biçimde test edilmektedir. Her bir değişkenin katsayılarının önemliliği, o değişkenin Y değişkeninin değişimi üzerine olan etkilerinin ne derecede önemli olup olmadığının test edilmesidir.

Çoklu regresyonda her regresyon katsayısının sıfırdan önemli düzeyde farklı olup olmadığının belirlenmesi için aşağıdaki modelden yararlanılır.

$t_i = \left(\frac{b_i - \beta_i}{\sqrt{var(b_i)}}\right)$ Hipotez gereği $\beta_i = 0$ olur.

$t_i$istatistiğinin önemliliği için $sd = n - 1$ serbestlik dereceli kuramsal t dağılımının kritik değerleri $t(\alpha, sd)$ ile karşılaştırılır.

$Var(b_i)$ değerleri $A = inv(X' * X)$ olmak üzere, $VARb = s2 * A$ biçimin de hesaplanır.

Çok değişkenli regresyon modelinde regresyon katsayılarının önemliliğinin belirlenmesi için RKT ve AKT matrislerinden faydalanılır. Bu matrisler yardımı ile H matrisi aşağıdaki gibi hesaplanır.

$$H = inv(AKT) * RKT$$

H matrisinin öz değerleri hesaplanabilmektedir. Bu öz değerler aracılığı ile aşağıda verilen bir yaklaşımla L test istatistiği hesaplanır.

$$L = \Pi(1 + \lambda\_\{i\})^{-1}$$

L test istatistiği $x^2 = -(n - p - 1 - (1/2) * (p - q + 1)) * log(L)$ şeklinde kikare test istatistiğine dönüştürülür. χ^2 değerinin önemliliği, $\chi^2(\alpha, pq)$ teorik dağılımının kritik değerleri aracılığı ile karşılatırılarak bulunur. q : bağımlı değişken sayısı; p : bağımsız değişken sayısı

3.10. Ana Bileşenler Analizi (ABA, PCA)

Ana bileşenler Analizi, aralarında korelasyon ilişkisi bulunan p tane gözlemsel değişkenin varyansını açıklayan daha küçük sayıda, birbirlerinden bağımsız ve orjinal değişkenlerin doğrusal bileşenleri olan gizil değişkenlerle belirtme yöntemidir.

ABA bir veri indirgeme tekniğidir. Ana bileşenler analizinde temel fikir, aralarında ilişki bulunan p sayıda değişkenden ibaret olan veri kümesinin, mevcut çeşitliliği için olabildiğince önlem alarak, n boyuttan k boyuta indirgemeye yarayan tekniktir. Bu durum ilişkisiz olan ve ilişkisiz olmayan ana bileşenlerde özgün değişkenlerinden var olan çeşitliliği koruyarak, ayarlanmış doğrusal kombinasyonlarının yüksek varyansının kontrol edilerek ve orijinal değişkenin sayısından daha az yani $k < p$ olacak biçimde özgün bir değişken grubuna çevrilerek gerçekleştirilir. Aynı zamanda bu özgün değişkenler öbür çok değişkenli analizlerde de kullanılabilmektedir (Jolliffe 2002; Özdamar 2013).

Ana bileşenler analizinin dört genel amacı vardır (Özdamar 2013):

1) Değişken indirgemesi yapmak.

2) Hesaplama yapmak

3) Diğer tekniklerin analiz edilebilmesinde yardımcı olmak (Regresyon ve faktör analizi).

4) Birbiriyle ilişki içinde olan değişkenlerden türetilecek ana bileşen skorlarını açıkladığı maksimal varyansa dair birimleri bu skorlara göre büyükten küçüğe doğru veya küçükten büyüğe doğru olacak şekilde sıralamak.

Genel manada ana bileşenler analizi boyut indirgeme ve bağımlılık yapısını ortadan kaldırmak için kullanılan bir teknik olsa da bazı diğer yöntemler içinde veri hazırlama yöntemi olarak kullanılabilmektedir (Tatlıdil 1992). Farklı bir biçimde anlatmak gerekirse ana bileşenler bizzat kendileri sonuç olmaktan ziyade sonuç almayı hedefleyen bir özelliğe sahiptir. Örneğin, veri setine çoklu regresyon uygulayabilmek için çoklu regresyon varsayımlardan biri olan "bağımsız değişkenler arasında çoklu doğrusal bağımlılığın olmaması "şartının yerine getirilmesi gerekir. Eğer veri matrisi bu

şartı sağlamıyor ise çoklu regresyon analizi uygulanamaz. Böyle bir durumda ana bileşenler analizi aracılığı ile yapay değişkenler türetilerek çoklu doğrusal bağlantı problemi ortadan kaldırılır, yeni yapay değişkenlerle çoklu regresyon analizi uygulanır. Aynı zamanda ana bileşenler analizi faktör analizi modelleri için ölçeklenmemiş hallerde kovaryans matrisi çarpanları olarak kullanılmaktadır (Özdamar 2013).

Ana bileşenler analizinde değişkenlere bağımsız değişken ve bağımlı değişken ayrımı yapılmadan, bütün gözlemlerin birbirinden bağımsız olduğu, bütün değişkenlerin normal dağılım gösterdiği ve değişkenlerin arasında bir doğrusal ilişkinin olması gerekmektedir (Bayram, 2009). Ana bileşenler analizindeki veri seti içinde gözlemlenen değişkenlerin bağlantıları belirtilmek için çok değişkenli normallik varsayımlarına ihtiyaç yoktur. Fakat anakütle de, çok değişkenli normallik varsa numune olan bileşimlerden intikal yapılabilir ve analizin açıklama gücü artabilir. Aynı zamanda ana bileşenler analizinde yararlanılan matrislerin tersine gerek olmadığından için çoklu doğrusallık için sorun oluşturmamaktadır (Tabachnick ve Fidell 2015).

3.10.1. Ana bileşenlerin elde edilmesi

Cebirsel olarak ana bileşenler analizi, p sayıdaki rastgele değişkenin $X_1, X_2, \ldots, X_p$ doğrusal bileşenleridir. Geometrik olarak ise, doğrusal bileşenleri $X_1, X_2, \ldots, X_p$ koordinat eksenlerini orijinal olan düzeneği döndürerek elde edilen, özgün bir koordinat sisteminin seçimini gösterir. Yeni olan eksenler maksimal kararsızlığı gösteren gidişatı belirtmekte ve kovaryansın çok basit ve çok az sayıda değişken aracılığıyla açıklanmasını sağlamaktadır. Bu açıdan bakıldığında ana bileşenler $X_1, X_2, \ldots, X_p$, kovaryans matrisi $\sum$ veya korelasyon matrisi p' ye bağlıdır. Daha önceden de belirtildiği gibi anabileşenler analizindeki X veri matrisinde çoklu normallik varsayımının olması gerekmemektedir. Öte yandan çoklu normallik gösteren bir popülasyon söz konusu ise türetilen ana bileşenler sabit yoğunluk elipsoitleri cinsinden faydalı yorumlamalarda bulunabilir. Aynı zamanda toplum çoklu normallik gösteriyorsa örnek bileşimlerden çoklu normallik dağılımlarına ilişkin çıkarsamalarda bulunabilir (Johnson ve Wichern 2007). X veri matrisi p sayıdaki değişkenin doğrusal bileşenlerini bulabilmek için korelasyon veya kovaryans matrisinin özdeğerlerini ve özvektörlerini kullanması gerekir. Öz değerlere karşılık gelen öz vektörler birbirinden

bağımsızdır. Ana bileşenler analizi bu özellikten yararlanarak, öz değerlerin büyüklük sırasına göre önem derecelerini belirlemektedir.

Bir rastgele vektör $X' = [X_1, X_2, \ldots, X_p]'$ nin $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ şeklinde öz değerleriyle beraber kovaryans matrisine $\sum$ sahip olsun. Öz vektörler yükler matrisi $(E = L)$ aşığıdaki gibi hesaplanır.

$$E = L = \begin{bmatrix} \ell_{11} & \ell_{21} & \ldots & \ell_{p1} \\ \ell_{12} & \ell_{22} & \ldots & \ell_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \ell_{1p} & \ell_{2p} & \ldots & \ell_{pp} \end{bmatrix}$$

Bu verilen varsayıma göre doğrusal bileşenler aşağıdaki şekilde hesaplanmaktadır.

$$Y_1 = \ell_1'X' = \ell_{11}X_1 + \ell_{21}X_2 + \ldots + \ell_{p1}X_P\}$$

$$Y_2 = \ell_2'X' = \ell_{12}X_1 + \ell_{22}X_2 + \ldots + \ell\_\{p2\}X_P$$

$$Y_P = \ell_p'X' = \ell_{1p}X_1 + \ell_{2p}X_2 + \ldots + \ell_{pp}X_P$$

Her bir ana bileşenin varyansı;

$$Var(Y_i) = \ell_i' \Sigma \ell_i \quad i = 1,2,\ldots,p$$

Kovaryansı;

$$Cov(Y_i, Y_k) = \ell_i' \Sigma \ell_k \quad i,k = 1,2,\ldots,p$$

şeklinde tanımlanır. Anabileşenler, $Y_1, Y_2, \ldots, Y_p$ birbirleriyle ilişkili değildir ve varyansların büyüklüklerine göre sıralanırlar. Birinci anabileşen, maksimum varyanslı doğrusal kombinasyon olmaktadır. Yani, $Var(Y_1) = \ell_1'\Sigma\ell_1$ en üst düzeye çıkarılır. $Var(Y_1) = \ell_1'\Sigma\ell_1$ birim uzunluğunda verilen bu ifade ile birtakım sabit vektörler aracılığıyla rastgele bir ℓ_1 çarpıldığında artabileceği bellidir. Böylece anabileşenler analizi şu şekilde belirlenir:

1.ana bileşen: $Var(\ell_1'X')'i\ \ell_1'\ell_1 = 1$ olması gereken şekilde maksimize edilerek bulunan $\ell_1'X'$ doğrusal kombinasyonudur.

2.ana bileşen: $Var(\ell_2'X')'i\ \ell_2'\ell_2 = 1$ ve $Cov(\ell_1'X', \ell_2'X') = 0$ olması gereken şekilde maksimize edilerek bulunan ℓ₂′X′ doğrusal kombinasyonudur.

i.ana bileşen $\ell_1'\ell_1 = 1$ ve $k < i$ için $Cov(\ell\_\{i\}'X', \ell\_\{k\}'X') = 0$ olması gereken şekilde maksimize edilerek bulunan $\ell_1'X'$ doğrusal bileşenidir (Johnson ve Wichern 2007).

Ana bileşenler, veri setinde bulunan değişkenlerin ölçü birimlerinin farklı olduğu ve değişkenliklerinin farklı olduğu zamanlarda korelasyon matrisinden veya standardize veri matrisinden hesaplanır.

$z_{ij} = (X_{ij} - \mu)/\sigma_{ii}$ yaklaşımıyla veri setindeki her bir değişkene ait veri vektörü standardize edilmektedir. R korelasyon matrisi, gerek standardize değerler olsun, gerekse orjinal değerlerin üzerinden hesaplansın, kovaryans matrisi S ve korelasyon matrisi R benzerdir. R matrisi standardize ve orjinal veri matrisleri için benzer olduğu için orjinal değişken için R matrisine standardize değişkenler için ise S ve R matrislerine anabileşenler uygulanarak benzer sonuçlar çıkarılır.

Eğer standardize değişkenlerin kovaryans matrisleri üstünden ana bileşenler hesaplanacaksa veri matrisi standardize edilebilir. $S = R$ matrisi hesaplanabilir. Eğer orjinal değerler aracılığı ile hesaplanan korelasyon matrisine göre ana bileşenler hesaplanacak ise değişkenler arasındaki ikili ilişki düzeyleri hesaplanarak R matrisi hesaplanabilir. Standardize ve orjinal değerlerin korelasyon matrisleri birbirlerine eşittir.

X veri matrisinde değişkenlerin arasında çok fazla yüksek korelasyonlar var ise, orjinal değerlerin ölçü birimleri çok farklı ise veya değişkenlerin değişim aralıkları analizi etkileyecek kadar çok farklılık gösteriyorsa böyle bir durumda verilerin standardize edilerek kullanılması uygun olmaktadır.

Verilerin standardize edilebilmesinde standart sapma köşegen matrisinin elemanlarından ve ortalama vektöründen faydalanılır.

3.10.2. Ana bileşenlerin elde edilmesinde kullanılan matrisler

Ana bileşenler analizi, hem orijinal veri setinden hesaplanan kovaryans ve korelasyon matrisleri kullanılarak hem de standardize edilmiş veri setlerinden hesaplanan korelasyon veya kovaryans matrisleri kullanılarak uygulaması yapılabilmektedir. Her iki veri tipinde ve her iki matristen de elde edilen ana bileşenler farklılık gösterebilir. Bundan dolayı farklı sonuçlar verebilen bu iki veri tipi ve matristen hangisinin seçileceği verilerin uygun olan ölçü birimlerine bağlıdır.

Veri setinde yer alan p tane değişkenin ölçü birimlerinde ve değişim aralıklarında birbirlerinden çok farklılık gösteriyor ise söz konusu olan durumu anabileşenler analizinde korelasyon matrisini kullanmak, aksi halde ise yani farklılık göstermiyor ise kovaryans matrisini kullanmak yerinde olacaktır. Standardize veri setlerinde korelasyon veya kovaryans matrislerinden yararlanılarak elde edilen ana bileşenleri kullanabilmek için verilerin çok farklı değişkenliğe ve değişim genişliğine sahip olması gerekir (Özdamar 2013).

3.10.3. Ana bileşen sayısını belirlemede kullanılan yöntemler

Ana bileşenler analizi, aynı zamanda daha az değişken ile çalışma yöntemidir. Bundan dolayı p değişken içeren verilerin kaç ana bileşen ile belirtilmesi gerektiği önemli olan sorunlardan biridir. Ana bileşen sayısını belirlemede kullanılan en popüler yöntemler aşağıdaki gibidir.

- **Öz değer sayısına göre ana bileşen belirleme (kaiser kriteri):** S matrisinin veya R matrisinin 1 veya 1'den çok büyük özdeğer sayısı ($\lambda_i \geq 1$) kadar anabileşen seçmek genellikle en yaygın olarak kabul edilmiş kurallardan biridir. Anabileşenler analizinde uygulanacak matrisin öz değerleri hesaplanır ve değeri 1'e eşit ve çok daha büyük olan kaç tane öz değer var ise o kadar sayıda ana bileşen belirlenmesi için çözümlemeler yapılmaktadır.

- **Açıklanan varyans yüzdesine göre ana bileşen belirleme**: Orjinal değişkenlerin yerine geçebilecek sayısı belirlerken en az ana bileşen sayısının toplam varyansın $2/3$'üne yani $\%67$'sini açıklayacak şekilde anabileşen belirlemek uygun olmaktadır. Açıklanan varyans oranı $\%67$'den daha büyük değerler seçilirse bu yöntem

anabileşenler analizi çözümlerinin tutarlılığı açısından gerekli olabilir. Böyle bir durumda açıklanan varyans oranını %95'lere çıkan değerlerde seçmek gerekebilir. Fakat %67'den daha büyük oranlar çok sayı da anabileşenin seçilmesini gerektirir. Optimal oran araştırmanın boyutuna, veri indirgemenin derecesine göre araştırıcı aracılığıyla belirlenmelidir.

$\Sigma\lambda_i \geq \%67$ şartını sağlayan 2 tane özdeğer ($\lambda_1 ve\ \lambda_2$) bulunmaktadır. Verilen bu kurala göre örnek veriler için iki tane anabileşen seçmek uygundur.

- **Öz değerler eğim grafiğinden faydalanarak ana bileşen sayısını belirleme**: S veya R matrisinin özdeğerli bulunduktan sonra bulunan bu öz değerlerin büyüklük sırasına göre dizilmiş halleri bir çizgi grafiği şeklinde gösterilir. Koordinat düzleminde X ekseninde bileşen sayısı $(1,2,\ldots,p)$, Y ekseninde özdeğerler olmaktadır. Bu çizgi grafiğinin eğimine bakılmaktadır. Eğim grafiği, çoğunlukla negatif yönlü bir eğilim göstermektedir. İlk bileşen en yüksek olmak üzere, takip eden diğer değişkenlerde orta ve düşük düzeyde azalan bir eğimle sıralanmaktadır. Azalan değerlere göre bir eğim izleyen eğilim çizgisinin eğiminin sabitleştiği veya çok küçük azalan kuvvetlere eriştiği noktaya kadar olan sayıda ana bileşen seçilmektedir ve çözümlemeler bu öz değerlerden yola çıkarak uygulanmaktadır. Ancak eğim grafiği kesin bir sonuç vermemektedir. Araştırmacıların seçimlerinde ideal güvenilirlik söz konusu olmadığından öz değerdeki kesilmelerin nerede olduğu kısmen belirlenmektedir (Tabachnick ve Fidell 2015).

Kural olarak, ana bileşen sayısını belirleme yöntemlerinin birlikte ele alınması fayda sağlamaktadır. Araştırmacı problemin özelliğine ve veri matrisinin yapısına göre üç yöntemi birlikte ele alarak ana bileşen sayısının kaç tane olduğuna karar vermelidir. Genel bir kural olarak ikiden daha az sayıda ana bileşen ile çalışılmamalıdır.

3.11. Faktör Analizi (FA)

Faktör Analizi, 1930 ile 1950 yıllarında matematiğin yaygın olan gelişimi içerisinde ilerleyen ve değişen çok değişkenli analizlerden birisidir. Öncelikli olarak psikoloji alanında yaygın bir kullanım alanı keşfetmiş ve 1950 yılından sonra bilgisayarların gelişmesi sonucunda sosyal alanlarda ve bunun yanı sıra farklı alanlarda da kullanılmaya başlanmıştır.

Faktör analizi (FA) birbirleri ile ilişkili çok sayıdan oluşan veri yapılarını birbirinden bağımsız, daha az sayıda özgün veri yapılarına çevirmek, yapılan veri yapısında elde edilen değişkenleri kümeleyerek bir olayı anlattıkları varsayılan ortak faktörlerin belirlenmesini, bir oluşumu etkileyen değişkenleri grup haline getirmek, majör ve minör faktörlerini tanımlamak için kullanılan bir tekniktir.

Diğer bir ifade ile faktör analizi, ortak boyutların belirlenebilmesi ve bağımlılık yapısının yok edilebilmesi için kullanılan bir tekniktir. Faktör analizinin hedefi fazla sayıda incelenen değişken ile daha küçük sayıdaki faktör denilen daha küçük ölçüdeki gizli değişkenlerin arasında kovaryans bağlantısı oluşturur. Faktör analizi, birbirleri ile bağlantılı sayısız değişkeni bir arada toplayarak, kavramsal olacak şekilde anlamlı olan daha az sayıda özgün veri setleri keşfetmeyi hedefleyen istatistiksel yöntem olarak da tanımlanmaktadır (Tavşancıl 2002).

Diğer bir araştırmacıya göre Faktör analizi incelenebilen ve ölçülebilen oldukça fazla sayıdaki özelliğin ardında bulunan gerçek sebeplerin gizil boyutların ortaya çıkarmaktadır (Hair, Rolph, Ronald & William 1998). Faktör analizi, çok karışık ve çok boyutlu bir bağlantı ile karşılaşıldığı hallerde, çok boyutlu ölçekleme analizi, kümeleme analizi ve kanonik korelasyon analizi gibi kullanılabilen bir tekniktir (Albayrak 2005).

Faktör analizi yorumlanabilmesi zor, birbirleriyle ilişkili çok fazla sayıda değişkenden en az bilgi kaybı ile bağımsız, kavramsal bakımdan az sayıda değişkenler bulmayı, ortaya koymayı hedefleyen çok değişkenli yöntemdir. Faktör analizi anlaşılabilmesi zor olan çok değişkenli bir verinin, daha az değişkenle açıklanıp açıklanmayacağının belirlenmesinde kullanılan bir tekniktir (Alpar 2011).

Faktör Analizinin iki temel amacı aşağıdaki gibidir:

1-Değişkenlerin toplam sayısını azaltmak

2-Orjinal değişkenlerin arasındaki ilişkilerden faydalanarak, aralarında ilişki olmayan fakat bir grup oluşturan değişkenlerin açıkladığı düşünülen bazı yeni yapılar ortaya çıkarmak (Özdamar 2018).

Faktör analizi, Ana bileşenler analizine benzemektedir. Faktör analizinde de Ana bileşenler analizinde de verilerin indirgenmesi mevzubahisti. Ancak Faktör Analizi Ana bileşenler analizinden farklı olarak birimleri kümeleyerek ortak faktörleri açıklama özelliğine sahiptir.

Faktör Analizi; özellikle sosyoloji, tıp, eğitim bilimleri, psikoloji, sosyal bilimler gibi alanlarda, değişkenlerin çok sayıda birbirleri ile bağlantılı olan özellikler arasından, birbirleriyle ilişkisiz ancak bir olayı açıklamada kullanılabilecek olan değişkenleri kümeleyerek yeni bir adlandırma faktör olarak açıklamayı sağlayan ve sürekli olarak kullanılan bir tekniktir.

Faktör analizi gözlenen ve aralarında korelasyon bulunan X veri matrisindeki p tane değişkenden gözlenemeyen ancak değişkenlerin bir araya gelmesiyle oluşan sınıflamayı yansıtan rastgele faktörleri ortaya çıkarmada kullanılır. Belirlenen bu yeni değişkenlere faktör adı verilmektedir.

k sorudan oluşan bir ölçek aracılığı ile ölçülen sosyo-kültürel, bilişsel ya da duyuşsal fenomenin hangi alt faktörlerden oluştuğunu belirlemek ve bu faktörlere birer isim vererek subjektif durumu objektif biçimde belirtmek için faktör analizinden yararlanılır.

Faktör analizi uygulama amacına göre farklı isimlerle ifade edilmektedir. Bunlar hakkında aşağıda kısaca bilgi verilmiştir.

Açıklayıcı faktör analizi (AFA): Birbirleriyle ilişkili p sayıda değişkenden oluşmuş veri setinin Kovaryans veya korelasyon matrisinden faydalanılarak daha az veya eşit sayıda ve birbirinden bağımsız yeni faktörleri belirlemek için kullanılan bir tekniktir.

Doğrulayıcı faktör analizi (DFA): Açıklayıcı faktör analizi aracılığıyla belirlenen faktörlerin, hipotez ile belirlenen ya da teorik faktör yapılarına uygunluğunu tespit etmek için kullanılan faktör analizi çeşididir.

Q-Tipi faktör analizi: Birimlerin birbirleriyle benzerliklerini inceleyerek bu benzerliklerden daha az sayıda homojen birim gruplamaları ortaya çıkaran bir analiz yöntemidir.

R-Tipi faktör analizi : Değişkenlerin R matrisinden yararlanarak yapılan bir faktör analizi çeşididir.

O-Tipi faktör analizi: Veri matrisinde sütunların yılları, satırların ölçümleri belirttiği durumlarda ölçümlerin hangi yıllarda kümelenme gösterdiğini araştıran faktör analizi yöntemidir.

T-Tipi faktör analizi: Veri matrisinde sütunların yılları, satırların birimleri ifade ettiği durumlarda tek değişkenli yapılarda birimlerin yıllara göre kümelenmesini göstermek için kullanılan faktör analizi yöntemidir.

S-Tipi faktör analizi: Veri matrisinde sütunların kategorileri, satırların yılları ve hücrelerde ise bir değişkene ait ölçümlerin bulunduğu durumlarda kategorilerin zaman periodlarına göre kümelenmesini inceleyen bir faktör analizidir.

Günümüzde tek değişkenli olan yöntemlerden O-tipi, Q-tipi, R-tipi, S-tipi, ve T-tipi faktör analizleri veri analizinde kullanılmamaktadır.

3.12. Kümeleme Analizi (Cluster Analysıs)

Çok değişkenli analiz tekniklerinden biri olan kümeleme analizinin temel amacı nesnelerin veya bireylerin temel özelliklerini dikkate alarak onları gruplandırmaktır. Kümeleme analizi X veri matrisinde bulunan ve doğal gruplamaları kesin olmayan değişkenleri, birimleri veya hem değişkenleri hem birimleri birbirleriyle benzer sınıflara ayırmada kullanılan bir yöntemdir. Kümeleme analizi birimleri değişkenleri ve hem birim hem değişkenleri uzaklık benzerlik farklılık ve yakınlıklarına göre hesaplanan ölçümlerden yararlanarak homojen gruplara ayırmak için kullanılır.

Kümeleme analizinin dört temel amacı şu şekildedir;

1- n sayıda nesneyi, oluşumu, birimi, p değişkenine göre kendi içinde homojen ve kendi içinde heterojen olan alt gruplara ayırmak,

2- p sayıda değişkeni, n sayıda birimden üretilen değerlere göre ortak özeliklerin açıklandığı varsayılan alt kümelere ayırmak ve kümelerin ortak yapılarını tespit etmek,

3- Hem değişkenleri hem de birimleri beraber ele alarak n birimini p değişkenli ortak özellikli alt gruplara ayırmak,

4- Birimlerin, p değişkene göre belirtilen yapılar vasıtasıyla toplumdaki doğal veya olası oluşturdukları düşünülen biyolojik ve topolojik sınıflamayı ortaya koyar.

3.12.1. Kümeleme analizinde dikkat edilmesi gereken hususlar

Analizde kullanılan uzaklık ölçüleri ile bu ölçüler üzerine kurulan bir takım yöntemlerin olması bu tekniğin kullanılmasında bazı mecburi genellemeleri ortaya çıkarmıştır (Hamarat 1998).

- Kümeleme analizi tekniklerinin birçoğu istatistikçi bir temele dayandırılmadığından dolayı kolay bir işlem süreci içermektedir.

- Yöntemlerin seçiminde kullanılacak veri setine en uygun olan kümeleme yöntemi seçilmelidir. Uygulamalarda farklı bilim dallarının seçilmesi durumunda her bilim dalının kendine özgü eğilime sahip olduğu ve verilerin farklılık gösterebileceğinden dolayı analiz aşamasında bu eğilime dikkat edilmelidir.

- Aynı veri setine farklı kümeleme yöntemleri uygulandığından farklı sonuçlar ortaya çıkabilmektedir. Kümeleme analizinin hedefi karşılaşılan her türlü özel sorunlara bir çözüm yolu bulmaktır.

3.12.2. Kümeleme analizi uygulamada izlenecek adımları

1. Veri matrisini elde etmek: Değişkenlerin veya birimlerin doğal sınıflamaları için kesin bilgilerin olmadığı topluluklardan alınan n sayıda örnek birimlerin p sayıdaki değişkenlerine ait verileri elde etmek.

2. Uzaklık, farklılık veya benzerlik matrisini hesaplamak: Değişkenlerin, birimlerin veya hem değişkenlerin hem de birimlerin birbirleriyle olan uzaklıklarını, farklılıklarını veya benzerliklerini belirten uygun bir uzaklık, farklılık veya benzerlik, matrisini hesaplamak.

3. Kümeleri belirtmek: Uygun kümeleme yöntemi aracılığı ile uzaklık, farklılık, benzerlik matrisine göre birimleri uygun olacak şekilde kümelere ayırmak.

4. Kümeleri araştırmak ve test etmek: Elde edilen kümelerin değerlendirilmesi ve kümeleme şekline göre kurulan varsayımları doğrulamak için gerekli olan yöntemleri irdelemek ve test etmek.

Buradan anlaşılacağı üzere kümeleme analizi çok sayıda farklı işlemi ele alan yöntem topluluğudur. Yani farklı amaçlar için farklı yöntemler içermektedir.

3.12.3. Değişkenlerin transformasyonu (dönüştürülmesi)

Verilerin belirli aralıklara dönüştürülmesi için birçok yöntem kullanılmaktadır. Bu yöntemlere aşağıda kısaca değinilmektedir.

- Z skorlarına dönüştürme: Kümeleme yapılacak birimler için tanımlanan değişkenler oransal veya aralıklı ölçekte olup normal gösterdiği düşünülen verilere uygulanan yöntemdir. En çok tercih edilen dönüştürme yöntemidir.

$Z_i = ((x_i - x)/S)$ p değişken sayısı olduğundan, $i = 1,2,\dots,p$ S, i. değişkenin standart sapması

- $-1 \leq X \leq +1$ aralığına dönüştürme: Aşırı sapan değerlerin bulunduğu, artı ve eksi değerlerin bulunduğu durumlarda ve verilerin heterojen yapıda olduğu durumlarda tercih edilen bir dönüştürme yöntemidir. X_{max} en büyük değer olmak üzere dönüştürme şu şekilde ifade edilir;

$x_i = \left(\frac{X_i}{X_{max}}\right)$ $0 \leq X \leq +1$ aralığına dönüşümü: Heterojen yapıdaki veri setlerinde, aşırı sapan değerlerin bulunduğu durumlarda değerleri pozitif ve $0-1$ arasında değişecek biçime dönüştürmede kullanılan bir dönüştürme yöntemidir. X_{max} en büyük değer X_{min} en küçük değer ve R değişim genişliği olmak üzere dönüştürme şu şekilde ifade edilir.

$X_i = \frac{X_i - X_{min}}{R}$ $\qquad R = X_{max} - X_{min}$

- Maksimum değeri 1 olacak biçimde dönüştürme: Veri setindeki değerlerin maksimum değerinin 1 olması istenildiği durumlarda kullanılan bir dönüşüm yöntemidir. Dönüştürme;

$$x_i = \frac{X_i}{X_{max}}$$

şeklindedir.

Eğer veri setindeki maksimum değer yani X_{max} sıfır ise dönüşüm;

$$x_i = \frac{X_i}{|X_{min}|} + 1$$

şeklinde yapılır.

- Dönüştürülmüş veri setinin ortalaması 1 olacak biçimde dönüştürme: Dönüştürülmüş veri setinin ortalamasını 1 yapabilmek için kullanılan bir dönüşüm yöntemidir. Dönüşüm;

$$x_i = \frac{X_i}{x}$$

şeklinde yapılır

Eğer veri setinin ortalaması yani x sıfır ise dönüşüm;

$$x_i = \frac{X_i + 1}{x + 1}$$

şeklinde yapılır.

- Dönüştürülmüş veri setinin standart sapması 1 olacak biçimde dönüştürme: Veri setinin dönüşümü yapıldıktan sonra veri setinin standart sapmasını 1 yapabilmek için kullanılan bir dönüşüm yöntemidir. Dönüşüm;

$$x_i = \frac{X_i}{S}$$

şeklinde yapılır.

Eğer veri setindeki değerler $-$ ve $+$ değerlerden oluşuyorsa ve standart sapması yani S sıfır ise veri setinde dönüşüm yapılamaz. Yapılamaması durumunda ise yukarıda

belirtilen bir diğer dönüşümlerden biri uygulanarak dönüşüm gerçekleştirilir. oluşturdukları düşünülen biyolojik ve topolojik sınıflamayı ortaya koyar.

3.12.4. Uzaklık ölçüleri

Bir veri setinde bulunan tüm değişkenler sürekli olduğu zaman birimlerin arasındaki yakınlıklar karakteristik olarak, farklılık ya da uzaklık ölçümleriyle ölçülür. nxn tipindeki matrislerde D uzaklıkları d_{ij} bu uzaklıkların elemanlarını göstermektedir. p sayıdaki değişkenler veya birimler için kullanılan uzaklık ölçüsü Minkowski uzaklık ölçüsüdür. Minkowski uzaklık ölçüsünün formülünde yer alan m değerinin 1 olması durumunda oluşan uzaklık ölçümü Manhattan uzaklığı, m değerinin 2 olması durumunda oluşan uzaklık ölçümü Öklid uzaklığı olarak adlandırılır. Sıklıkla kullanılan uzaklık ölçüleri Çizelge 3.3'de verilmiştir (Özdamar 2013).

Çizelge 3.3. Uzaklık ölçüleri

Uzaklık Ölçüleri	Denklemleri
Minkowski Uzaklığı	$d_{ij} = \left[\sum_{k=1}^{p} \left\lvert X_{ik} - X_{jk} \right\rvert^{m}\right]^{1/m} \quad m \geq 1$
Manhattan Uzaklığı	$d_{ij} = \sum_{k=1}^{p} \left\lvert X_{ik} - X_{jk} \right\rvert \qquad m = 1$
Öklid ve Karesel Öklid Uzaklığı	$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2}$ $d_{ij}{}^2 = \sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2 \quad m = 2$
Mahalanobis Uzaklığı	$d_{ik} = \sqrt{(X_i - X_k)^T (X_i - X_k)}$

Çizelge 3.3'ün devamı

Gower Uzaklığı	$d_{ij} = \frac{1}{n}\sum_{k}\frac{\lvert X_{ik} - X_{jk}\rvert}{max_k - min_k}$
Pearson ve Karesel Pearson Uzaklığı	$d_p(i,j) = \sqrt{\frac{\sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^2}{S_k^2}}$ $d_p^2(i,j) = \frac{\sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^2}{S_k^2}$

3.12.5. Benzerlik ölçüleri

Bütün değişkenlerin kategorik olduğu verilerde genellikle benzerlik ölçüleri kullanılır. Ölçümler çoğunlukla $[0,1]$ aralığında ölçülürken bazı durumlarda $0-100$ aralığında yüzde biçiminde de ifade edilir. Benzerlikte bulunan sıfır değeri iki değişkenin bütün değişkenler için maksimum açıdan farklı olduğunu belirtir. Benzer olan öğelerin eğer karakteristiği var ise 1 değerini ve karakteristiğin olmaması halinde 0 değerini alırken, ikili değişken ilave ederek matematiksel açıdan daha çok karakteristiğe sahip olabilir. İki değişkenin benzeşme değerinin bilinmesi olasılık tablosunda beraber olup olmadıklarına göre Çizelge 3.4'te verilmiştir.

Çizelge 3.4. Benzeşme değeri için olasılık tablosu

i. gözlem	j. Gözlem	
	Var(1)	Yok(0)
Var(1)	a	b
Yok(0)	c	d

a: İki gözlemdeki değişken sayısı

b: $X_{ik} = 1$ ve $X_{jk} = 0$ olan değişkenin sayısı

c: $X_{ik} = 0$ ve $X_{jk} = 1$ olan değişkenin sayısı

d: İki gözlemde 0 bulunan değişken sayısı

İkili değişkenler için önerilen benzerlik ölçüleri Çizelge 3.5'te verilmiştir.

Çizelge 3.5. İkili değişkenler için benzerlik ölçüleri

Benzerlik Ölçüleri	Denklemleri
Sneath ve Sokal	$\frac{a}{[a+2(b+c)]}$
Gower ve Legendre	$\frac{(a+d)}{\left[a+\frac{1}{2}(b+c)+d\right]}$
Eşleştirme Katsayısı	$\frac{(a+d)}{p}$
Jaccard Katsayısı	$\frac{a}{a+b+c}$
Rogers ve Tanimoto	$\frac{(a+d)}{[a+2(b+c)+d]}$

3.12.6. Kümeleme yöntemleri

Temel kümeleme yöntemlerine Çizelge 3.6'da yer verilmiş ve bu yöntemlerin özeliklerine değinilmiştir (Han vd.2012; s 450; Güler, 2006: s 56).

Çizelge 3.6. Kümeleme Yöntemleri

Yöntemler	Özellikleri
Izgara Tabanlı	Çoklu sonuç üretir. Hızlı işlem yapma imkanı sağlar.
Hiyerarşik	Hiyerarşik bir işleyiş öngörür. Bağlantı teknikleri ya da mikro kümelenme gibi tekniklere dahil olmayı sağlar.

Çizelge 3.6'nın devamı

Hiyerarşik Olmayan	Mutlak ayrık kümeleri bulabilmek için kullanılan yöntemdir. Uzaklık ölçümü aracılığı ile işlemler yapılır. Küçük ve orta ölçekli kümelerde etkin rol oynar.
Yoğunluk Tabanlı	Uç değerleri ayırır. Rastgele olan şekillerdeki kümeleri oluşturur.
Bulanık	Uzaklık ölçülerini kullanmada esneklik gösterir. Yorumlama kısmında kullanışlı üyelik değerlerini sağlar.

3.13. Diskriminant Analizi (Ayırma Analizi)

Gerçek anlamda ayırma olup değişkenlere ait p tane özellikten faydalanarak dahil oldukları grupların belirlenmesinde ve var olan grupları birbirinden ayıran fonksiyonu belirlemede kullanılan istatistiksel bir tekniktir (Çamdeviren 2000). Birimlerin gruplanmasında diskriminant fonksiyonu adı verilen eşitliklerden yararlanılır. Bu eşitlikler birbirlerine benzeyen grupları belirlemede yarar sağlayacak biçimde grupların ortak özelliklerini belirtmek için kullanılır. Bu grupları ayırmak için yararlanılan karakteristikler diskriminant değişkeni olarak adlandırılır. Özetlemek gerekirse iki veya daha çok sayıdaki grupların farklılıklarının diskriminant değişkenleri ile ortaya konulmasıdır (Klecka 1980).

3.13.1. Diskriminant analizi kullanımı ve varsayımları

X veri setinde yer alan değişkenlerin iki ya da daha çok gerçek gruplara ayrılmasında kullanılan yöntemdir (Özdamar 1999).

Diskriminant Analizi, gruplar arasında bulunan çeşitli değişkenlere bağlı kalarak farklılıkları ortaya çıkarmaya olanak sağlar. Birimler ise en az hatayla ait oldukları

birimlere ayrılır. İki ya da daha çok gruptaki birimler arasındaki etkileşimin ne seviyede olduğu, diğer değişkenlerin arasında ne tarz bir farklılık olduğunu ortaya çıkarmaktadır (Tümer 2001).

Diskriminant Analizi ayırma fonksiyonu belirlemek için uygulanmışsa tanımlayıcı diskriminant analizi, sınıflamak için uygulanmışsa tahmin edici diskriminant analizi olarak adlandırılır.

Diskriminant analizinin kullanılabilmesi için veri setlerinin ANOVA ve MANOVA yöntemleri için aşağıdaki varsayımları sağlaması gerekir.

1) X veri seti çok değişkenli normal dağılıma sahip olmalıdır.

2) Veri setindeki değişkenlerin varyansları ve ortalamaları arasında bağlılık bulunmamalıdır.

3) Değişkenlerin kovaryans matrisleri türdeş olmalıdır. X veri matrisinde bulunan değişkenler ortak kovaryans matrisine ait çok değişkenli normal dağılımdan oluşan ana kütleden çekilmelidir.

4) Değişkenlerin arasında yüksek korelasyon olmamalıdır.

5) Veri setindeki değişkenlerin arasında çoklu bağımlılık olmamalıdır.

6) X veri matrisi grupların birbirlerinden ayrılmasını sağlayacak şekilde doğru ve gerekli olan değişkenleri içermelidir. Gereksiz değişkenlere yer verilmemelidir.

Diskriminant analizinin varsayımları:

1) Anakütle belirli özelliklere göre gruplandırılabilir (Tatsuoka 1976). Birbirlerinden farklı iki ya da daha çok grup olmalıdır (Polat 1995).

2) Bütün gruplarda en iki durum söz konusu olmalıdır. Yani $n_i \geq 2$ olmalıdır.

3) Veriler anakütlenin içerisinden rastgele seçilmiştir.

4) Ayırt edici değişkenler 1.ölçek düzeyinde ölçülmelidir.

5) Herhangi bir ayırt edici değişken başka bir ayırt edici değişkenin lineer bileşimi olmamalıdır.

6) Her grup için herhangi bir özel formül kullanılmadığı takdirde kovaryans matrisleri eşittir. Gruplar için kovaryans ve ortalama matris önceden bilinmelidir. Grupların sapma matrisleri de eşittir (Karels-Prakash 1987). Bu varsayım sağlanmadığı sürece Diskriminant analizinin karesel olan yöntemi kullanılır.

7) Eğer gruplar eşit sayıda birimden oluşmuyorsa bu durumda üyelerin tahmini olasılıklarının bilindiği farzedilir.

3.13.2. İki grup için doğrusal diskriminant analizi

Doğrusal diskriminant analizinin temel sorunu p adet değişkenin hangi lineer bileşimin örneklerini en iyi biçimde ayırt edeceğidir. Sonuçlar gruplar arası oranları, grup içi oranları ve toplam varyansı kapsayan bir kurala göre belirlenmektedir.

Bu analiz için açıklanan varyansın yüzde yüzüne tekabül eden bir öz değer ve bir diskriminant fonksiyonu vardır. Birden çok diskriminant analizi olması durumunda bunların ilki bunların en büyüğü ve en önemli olanıdır. İkinci fonksiyon açıklayıcı anlamda ikinci olarak en önemlisidir.

Doğrusal diskriminant fonksiyonunu kullanabilmek için verilerin ortak varyans-kovaryans matrisine sahip olmaları gerekir ve aynı zamanda normal dağılımlı olması gerekir. Varsayımların sağlanmaması durumunda bir takım çalışmalar yapılmıştır. Fisher'in belirlediği klasik diskriminant analizinde p değişkenli iki grup için iki fonksiyon belirlenmiştir.

g_1 ve g_2 gruplarından alınann_1 ve n_2 hacimli çok değişkenli veri grupları olsun. X_1 ve X_2 bu veri gruplarının gözlem matrisi, S_1 ve S_2 ise kovaryans matrisleri olsun. g_1 ve g_2 Σ ortak korvaryans matrisine sahip olan toplumun rastgele örnekleri olsun. Böylece $S_1 = S_2$ olur. Bu grupların ortak varyansları olan S matrisi aşağıdaki gibi hesaplanır.

$$S = \frac{(n_1 - 1) * S_1 + (n_2 - 1) * S_2}{n_1 + n_2 - 2}$$

Ortak kovaryans matrisi aşağıdaki biçimde hesaplanır.

$$\hat{\Sigma} = E(X - \bar{x}_\iota)(X - \bar{x}_\iota)'$$

Diskriminant analizi her grup için birer ayırma fonksiyonu hesaplamayı hedefler.

$Y_i = b_{0i} + b_{1i}X_1 + b_{2i}X_2 + \cdots + b_{pi}X_p \quad i = 1,2$ grup sayısı

Burada b_{0i} sabit değeri b_{ij} ise doğrusal bileşenleri belirtir. Doğrusal bileşenlere aynı zamanda kanonik değişkenlerde denir.

Doğrusal bileşenler aşağıdaki şekilde de hesaplanır:

$$b_{ij} = S^{-1}(\bar{x}_i) \quad i = 1,2,\ldots g \quad j = 1,2,\ldots,p$$

b_{ij} katsayılarına doğal değişkenler veya grup ayırma fonksiyonunun katsayıları da denir. Bazen bu katsayılar belli bir kritere göre ölçeklendirilerek veya normalize edilerek kullanılır. İki tür ölçeklendirme veya normalizasyon vardır.

$b^* = \frac{b}{\sqrt{b'b}}$ veya $b^* = \frac{b}{b_{max}}$ b_{max}: en büyük doğrusal bileşen

İki tür ölçeklendirme veya normalizasyon yaklaşımında amaçlanan durum katsayılardan birinin diğer katsayılara göre önemli olacak şekilde ağırlığını arttırmak ve diğer katsayılara göre çok önemli olan değişkenleri modelde etkin biçimde görmektir.

Ayırma fonksiyonu aracılığı ile gruplar arasındaki farklılığı maksime ederek grupları birbirinden ayırmak mümkündür. Bundan dolayı ortak olan bir ayırma fonksiyonu belirlenir. i ve j gruplarının arasında oluşan diskriminant fonksiyonu;

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

Burada ki $b_i$ler ortalama fark vektörü yardımıyla $b_i = S^{-1}(\bar{x}_i - \bar{x}_j)$ $i = 1,2,\ldots,g$ eşitliği ile bulunur.

b_0 sabit değeri ise $b_0 = -(1/2)\bar{x}'S^{-1}\bar{x}$ ile hesaplanır.

P değişkenli ve g grup arasında bulunan karesel uzaklığı veren ve Mahalanobis tarafından belirlenen D^2 uzaklığı ${D_{ij}}^2 = (\bar{x}_i - \bar{x}_j)' S^{-1}(\bar{x}_i - \bar{x}_j)$ ile hesaplanır.

i ve j gruplarının birbirinden ayrılmasında D^2 uzaklığının etkin rol oynayıp oynamadığı tespit edilir. Bu nedenden dolayı Hotelling T^2 ile D^2 nin önemliliği belirlenir. T^2 nin önemliliği içinde F yaklaşımından faydalanılır.

$$T^2 = \left(\frac{n_1 * n_2}{n_1 + n_2}\right) D^2$$

$$F = \frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} T^2$$

3.13.3. Çoklu doğrusal diskriminant analizi

$p \geq 2$ ve $g > 2$ şeklinde olan veri setlerine doğrusal ayırma analizini uygulayabilmek için doğrusal ayırma analizinin $g = 2$ geçerli olan kuralları $g > 2$ kuralı için genellenmelidir. Böylelikle grup sayısının çokluğu kadar ayırma fonksiyonu hesaplanmış olur.

g grubuna ait ortak kovaryans matrisi aşağıdaki şekilde hesaplanır.

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{n_1 + n_2 + \cdots + n_g - g}$$

Doğrusal ayırma fonksiyonu şu şekilde hesaplanır;

$$y_i(x) = \bar{x}_i' S^{-1} x - \frac{1}{2} \bar{x}_i' S^{-1} \bar{x}_i + lnp_i$$

p_i her bir grubun ilk olasılığını göstermektedir. p_i verilmediği durumlarda $p_i = \frac{n_i}{N}$ oranı olarak alınır.

Gruplar arasındaki karesel uzaklığı veren Mahalanobis uzaklığı aşağıda belirtildiği gibi hesaplanır.

$$D = (\bar{x}_i - \bar{x}_j)' S^{-1} (\bar{x}_i - \bar{x}_j)$$

x gözlem matrisinin i. gruba ait ortalama vektöründen uzaklığını belirten karesel uzaklık ise şu şekildedir;

$$D_i{}^2(x) = (x - \bar{x}_i)' S^{-1} (x - \bar{x}_i)$$

4. BULGULAR VE TARTIŞMA

Çalışmanın bu bölümünde çok değişkenli istatistik metotlarının bazıları kullanılarak örnek bir uygulama yapılmıştır. Veriye uygun analiz metotları seçilmiştir. Uygulama örneği için kullanılan veri, Elazığ'ın Ağın ilçesinde bulunan deri fabrikası civarından alınmış sediman örneklerinin (EADF) kimyasal özelliklerine aittir. Tüm örneklerin verileri excel programında düzenlenerek SPSS programına aktarılmıştır. Çalışmanın çok değişkenli istatistikleri SPSS programında yapılmıştır. Grafiklerin bir kısmı Excel de bir kısmı SPSS de yapılmıştır. Ham veriler göz önünde bulundurularak K, Ti, P, Na, Fe, Ca ve Si elementleri % cinsine çevrilmiştir. Mg, Mn, Al, Zn, Ba, Co, Nb, Rb, Sr, Pb, Cr elementleri ppm cinsinden alınmıştır.

4.1. Betimsel İstatistik

Çizelge 4.1. EADF Sediman verilerinin tanımlayıcı istatistikleri

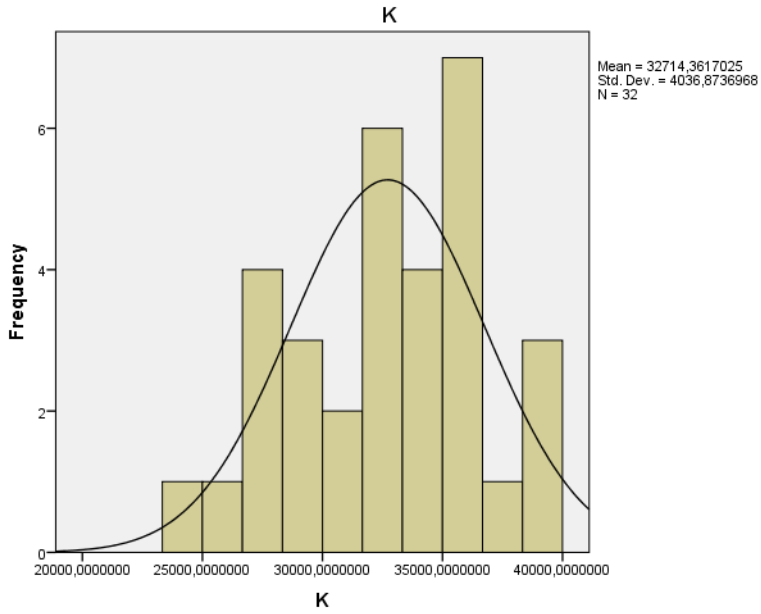
Kimyasal elementler	Ortalama	Medyan	Mod	Standart sapma	Çarpıklık	Basıklık	Minimum	Maximum
K	3,2714	3,3274	2,73[a]	0,40369	-0,26	-0,628	2,48	3,97
Ti	1,1044	1,029	,74[a]	0,49765	1,029	1,64	0,19	2,43
P	0,2021	0,1877	,05[a]	0,13767	0,837	-0,099	0,04	0,52
Na	2,5448	2,5485	1,82[a]	0,50781	0,129	-0,506	1,74	3,73
Mg	0,1568	0,15	,14[a]	0,05547	1,073	2,272	0,07	0,34
Fe	1,4193	1,379	1,16	0,38004	0,473	0,412	0,71	2,26
Ca	0,0000957	0,0000957	0,0001	1,566E-05	-0,594	1,702	0,00005	0,00013
Si	0,00337	0,00352	0,004	0,000292	-0,799	-0,84	0,003	0,004
Mn	0,0605	0,062	0,07	0,01155	-1,091	0,784	0,03	0,08
Al	6,035	6,0565	5,28[a]	0,64378	-0,176	0,159	4,59	7,41
Zn	64,3438	71,5	77	26,85037	1,03	4,186	23	160
Ba	550,5938	555,5	521	75,09128	-0,544	0,515	346	689
Co	202,0938	213	234	50,60496	-0,544	0,08	75	298
Nb	17,8406	15,5	,00[a]	11,32368	0,662	0,474	0	46
Rb	152,7813	141,5	134	37,14812	1,246	1,626	101	264
Sr	59,6875	64	65	16,84165	-0,792	1,322	12	95
Pb	75,4688	67,5	58,00[a]	18,34611	0,978	0,198	53	121
Cr	217,1563	222	212	35,22289	-0,811	1,658	132	298

a. Çoklu modlar mevcuttur. En küçük değer gösterilir

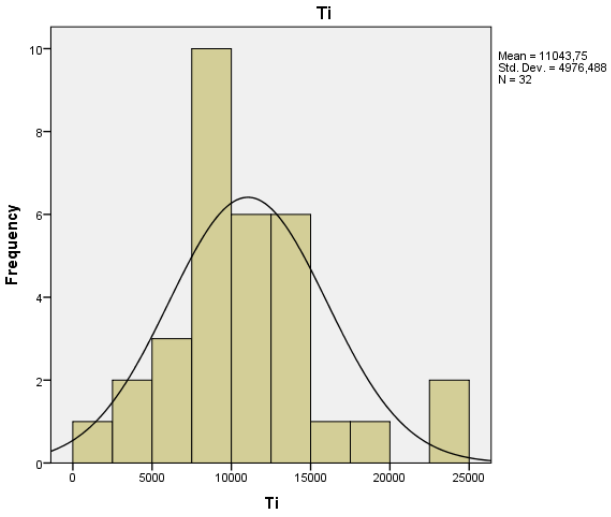
Çizelge 4.1'de Elazığ'ın Ağın ilçesinden farklı lokasyonlardan alınan sediman örneklerinin kimyasal değerlerinin tanımlayıcı istatistikleri olan ortalama, standart

sapma, basıklık, çarpıklık, minimum ve maksimum değerleri verilmiştir. Bu ve bu gibi verilerde veri setine hatalı veri girişi yaptığımız zaman analiz de çok farklı sonuçlar ortaya çıkmaktadır. Bunun önlenmesi için maksimum ve minimum değerleri hesaplanmaktadır. Çarpıklık alınan 32 sediman verisinin normal dağılımdan ne ölçüte saptığı hakkında bilgi vermektedir. Eğer çarpıklık değeri pozitif ise sağa çarpık, negatif ise sola çarpık dağılımı ifade etmektedir. Bu veriler için basıklık değerinin 0 dan büyük olması normal dağılıma göre daha dik bir dağılım, 0 dan küçük olması normal dağılıma göre daha basık bir dağılım olduğunu ifade etmektedir. Çizelge 4.1'de K elementinin ortalaması 3,2714, medyanı 3,3274, modu 2,73, standart sapması 0.40369,maksimum değeri 3.97 ve minimum değeri ise 2,48 bulunmuştur. K elementinin çarpıklık değeri -0,26 olduğundan dolayı büyük değerlerin fazlalıkta olduğunu ve sola çarpık dağılım olduğunu göstermektedir. K elementinin basıklık değeri -0,628 olduğundan dolayı normalden daha düz bir dağılıma sahiptir. Fe elementinin ortalaması 1,4193, medyanı 1,379, modu 1,16, standart sapması 0,38004,maksimum değeri 2,26 ve minimum değeri ise 0,71 bulunmuştur. Fe elementinin çarpıklık değeri 0,473 olduğundan dolayı küçük değerlerin fazlalıkta olduğunu ve sağa çarpık dağılım olduğunu göstermektedir. Fe elementinin basıklık değeri 0,412 olduğundan dolayı normalden daha dik bir dağılıma sahiptir. Cr elementinin ortalaması 217,156, medyanı 222, modu 212, standart sapması 35,222, maksimum değeri 298 ve minimum değeri ise 132 bulunmuştur. Cr elementinin çarpıklık değeri -0,811 olduğundan dolayı büyük değerlerin fazlalıkta olduğunu ve sola çarpık dağılım olduğunu göstermektedir. Cr elementinin basıklık değeri 1,658 olduğundan dolayı normalden daha dik bir dağılıma sahiptir.

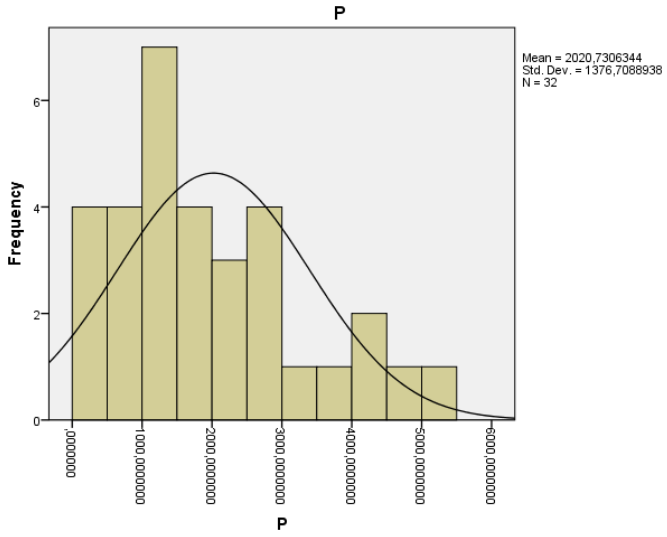
EADF Sediman örneklerine ait kimyasal verilerin dağılım grafikleri şekil 4.1 ile şekil 4.17 arasında verilmiştir.



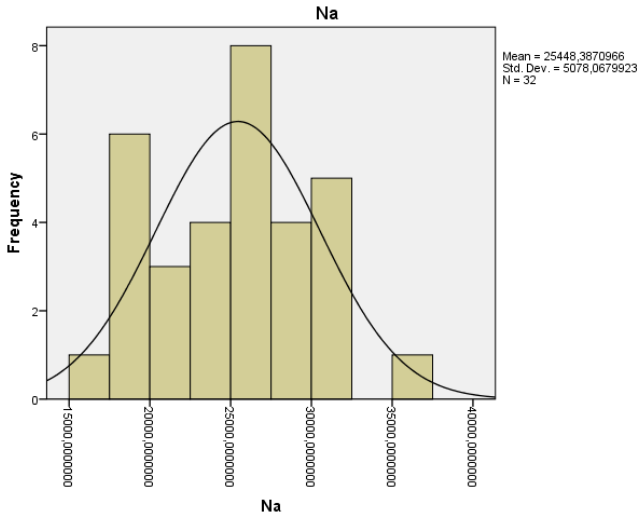
Şekil 4.1 K kimyasal elementi dağılım grafiği



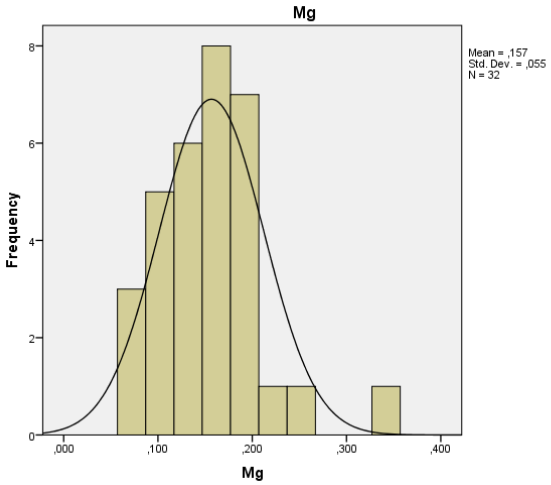
Şekil 4.2. Ti kimyasal elementi dağılım grafiği



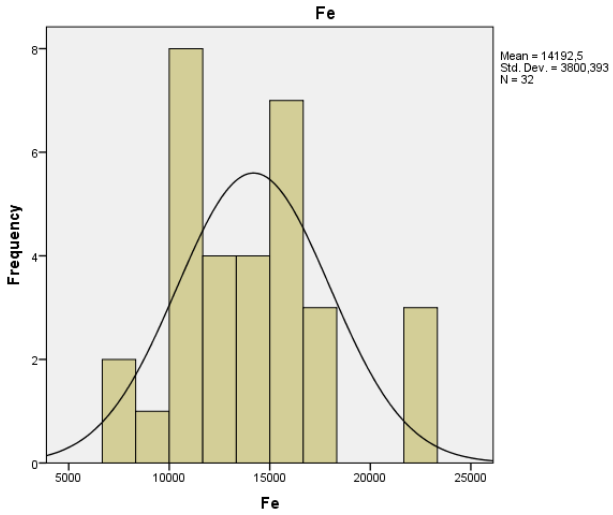
Şekil 4.3. P kimyasal elementi dağılım grafiği



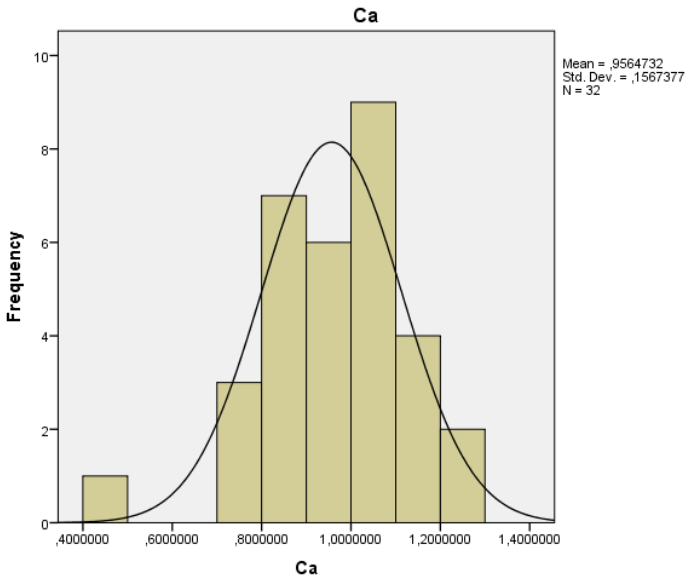
Şekil 4.4. Na kimyasal elementi dağılım grafiği



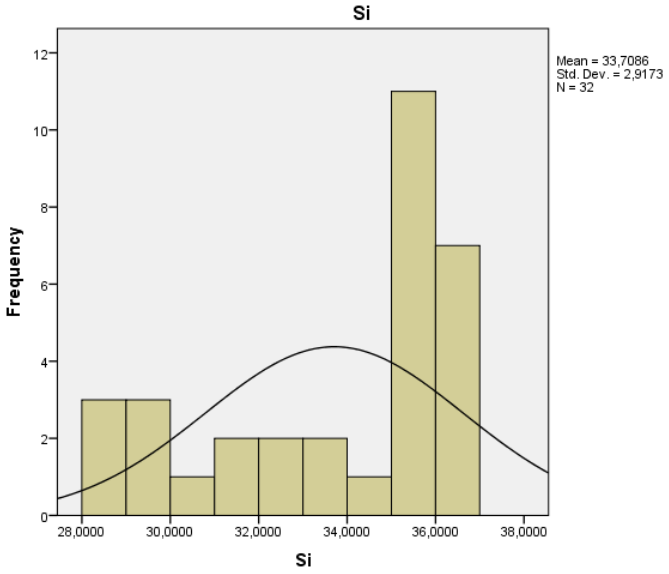
Şekil 4.5. Mg kimyasal elementi dağılım grafiği



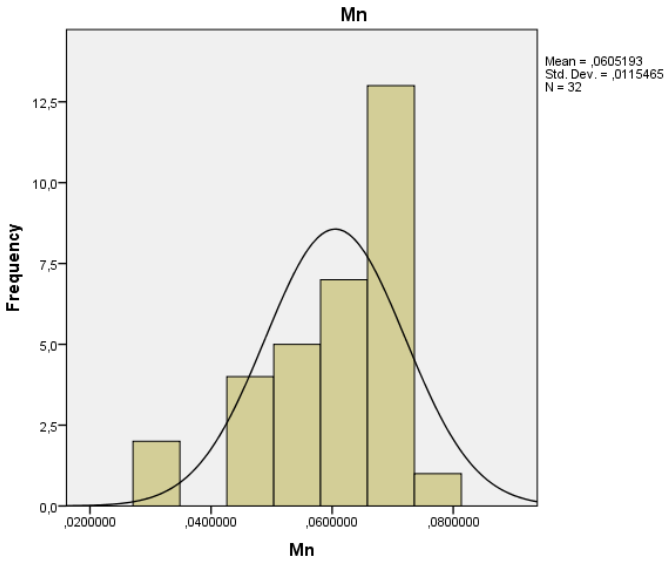
Şekil 4.6. Fe kimyasal elementi dağılım grafiği



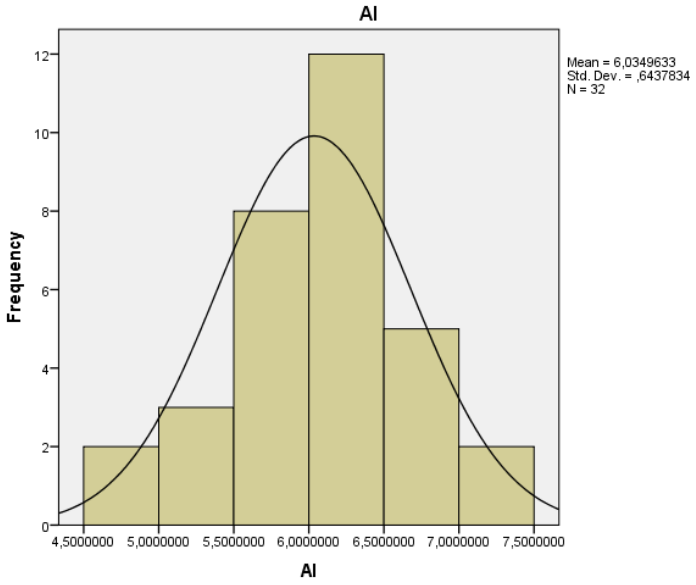
Şekil 4.7. Ca kimyasal elementi dağılım grafiği



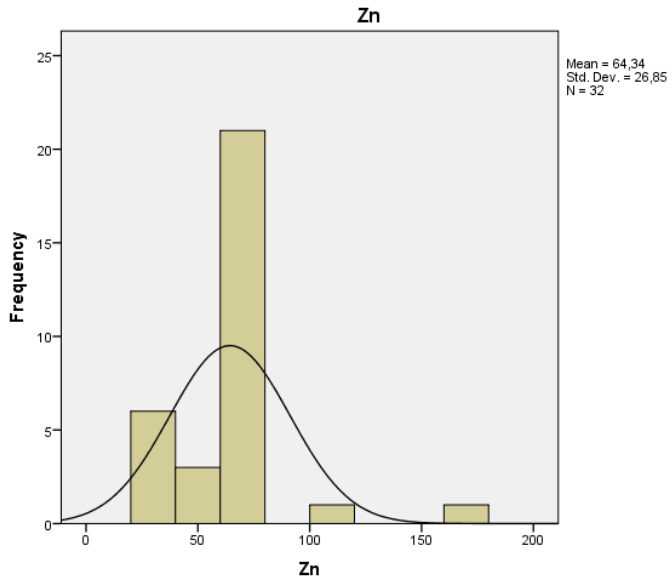
Şekil 4.8. Si kimyasal elementi dağılım grafiği



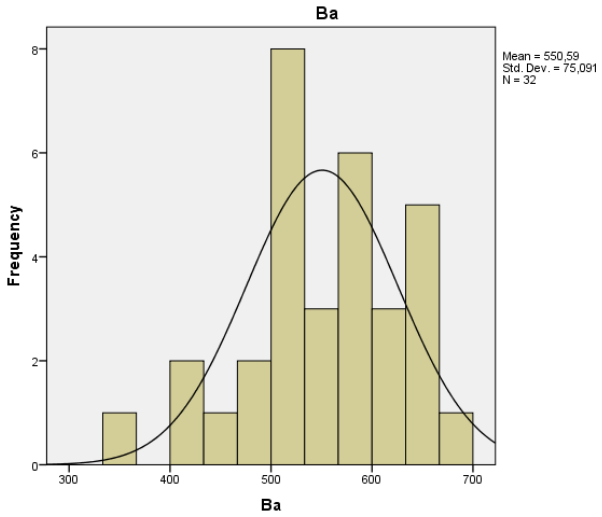
Şekil 4.9. Mn kimyasal elementi dağılım grafiği



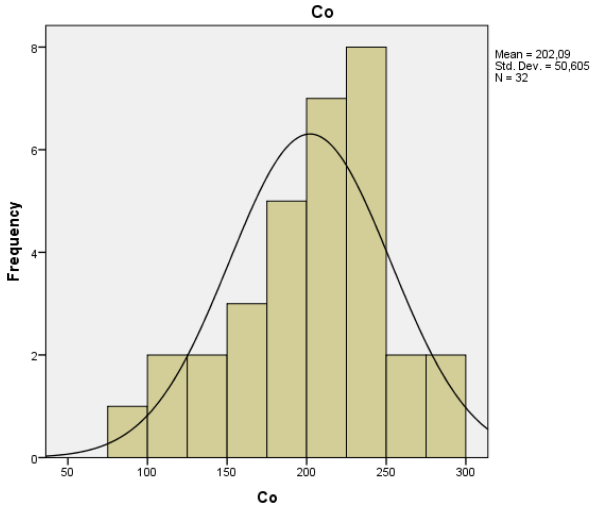
Şekil 4.10. Al kimyasal elementi dağılım grafiği



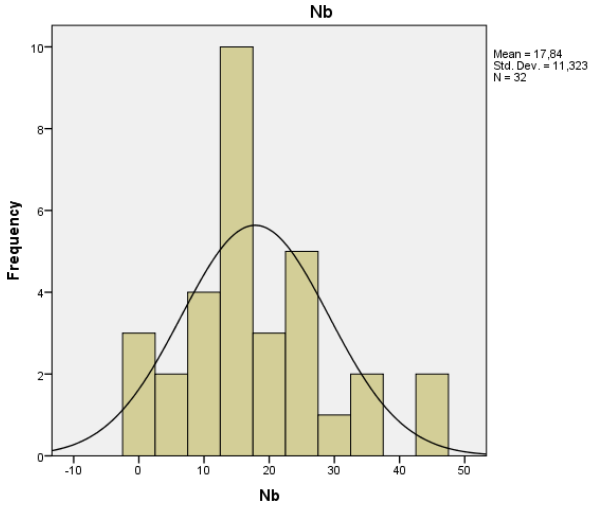
Şekil 4.11. Zn kimyasal elementi dağılım grafiği



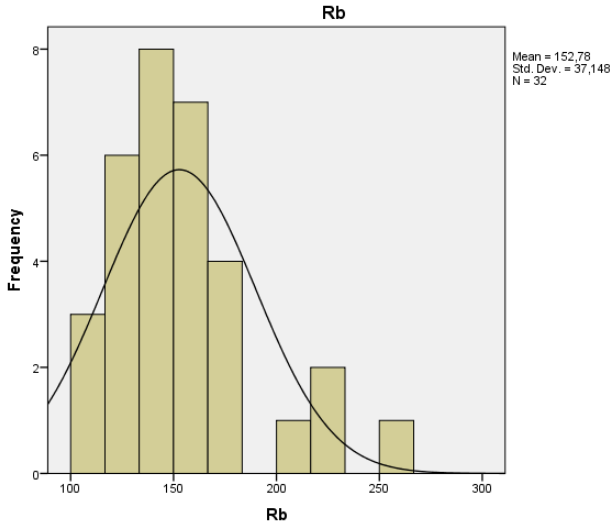
Şekil 4.12. Ba kimyasal elementi dağılım grafiği



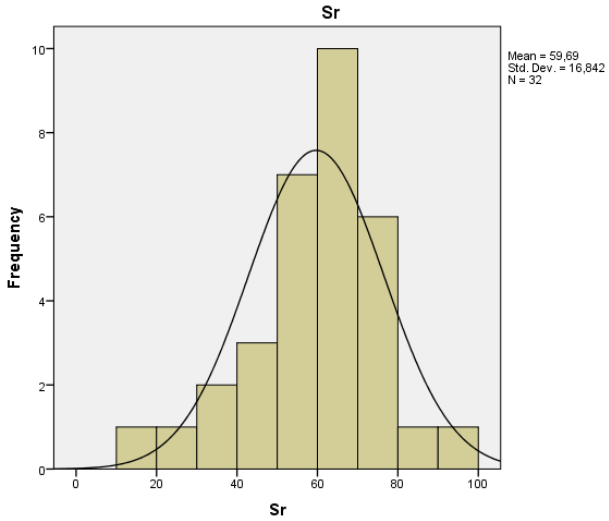
Şekil 4.13. Co kimyasal elementi dağılım grafiği



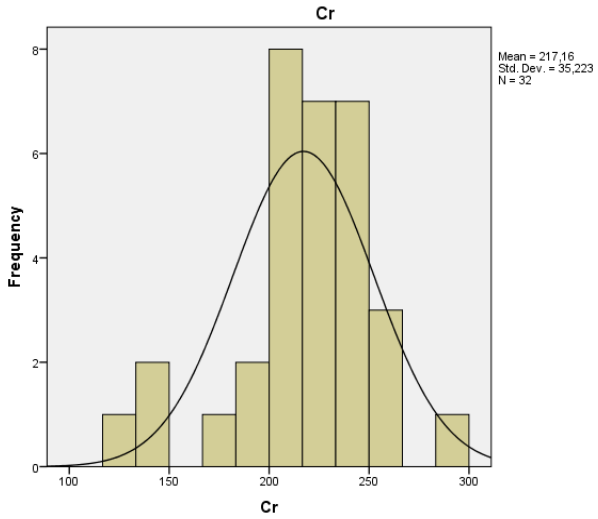
Şekil 4.14. Nb kimyasal elementi dağılım grafiği



Şekil 4.15. Rb kimyasal elementi dağılım grafiği



Şekil 4.16. Sr kimyasal elementi dağılım grafiği



Şekil 4.17. Cr kimyasal elementi dağılım grafiği

4.2. Normallik Varsayımı

EADF Sediman örneklerine ait kimyasal verilerin çok değişkenli istatistik yöntemlerinde analizi için öncelikle varsayımların sağlanıp sağlanmadığı incelenmiştir. İstatistiksel araştırmalarda testlerin yapılabilmesi için dağılımın normal ya da normale yakın olması gerekir. Bunun nedeni verilerin normalden uzak olması analiz sonuçlarının yanlış çıkmasına sebep olur. Özellikle regresyon analizi yapmadan önce mutlaka verilerin normallik varsayımını sağlayıp sağlamadığı kontrol edilmelidir. Veri seti 30 dan büyük olduğu durumlarda merkezi limit teoremine göre normal dağıldığı kabul edilebilir. Tezdeki örnek veri 32 farklı veriden oluştuğu için merkezi limit teoremine göre normal dağılmaktadır. Bu tür yaklaşımlar birçok araştırmacı tarafından kullanılmaktadır.

4.3. Korelasyon Analizi

Çok değişkenli analiz metotlarından korelasyon analizi sediman örneklerinde sıkça kullanılan bir metottur. Elementler arasındaki ilişkinin gücünün belirlenmesinde kullanılmaktadır. Korelasyon analizi iki değişken arasındaki doğrusal ilişkiyi test etmek için kullanılan bir yöntemdir. EADF Sediman verilerinin kimyasal analizlerinin korelasyon analizi sonuçları Çizelge 4.2'de verilmiştir. Bu veriler için korelasyon analizi iki kimyasal element arasındaki ilişkiyi belirlemektir. Korelasyon analizi

sonuçlarında 0,5 ile 0,65 arasında çıkan değerler orta derece kabul edilmiştir. 0,5'ten küçük olan değerler düşük derece kabul edilmiştir.

Çizelge 4.2. EADF Sediman verilerinin kimyasal analizlerinin korelasyon tablosu

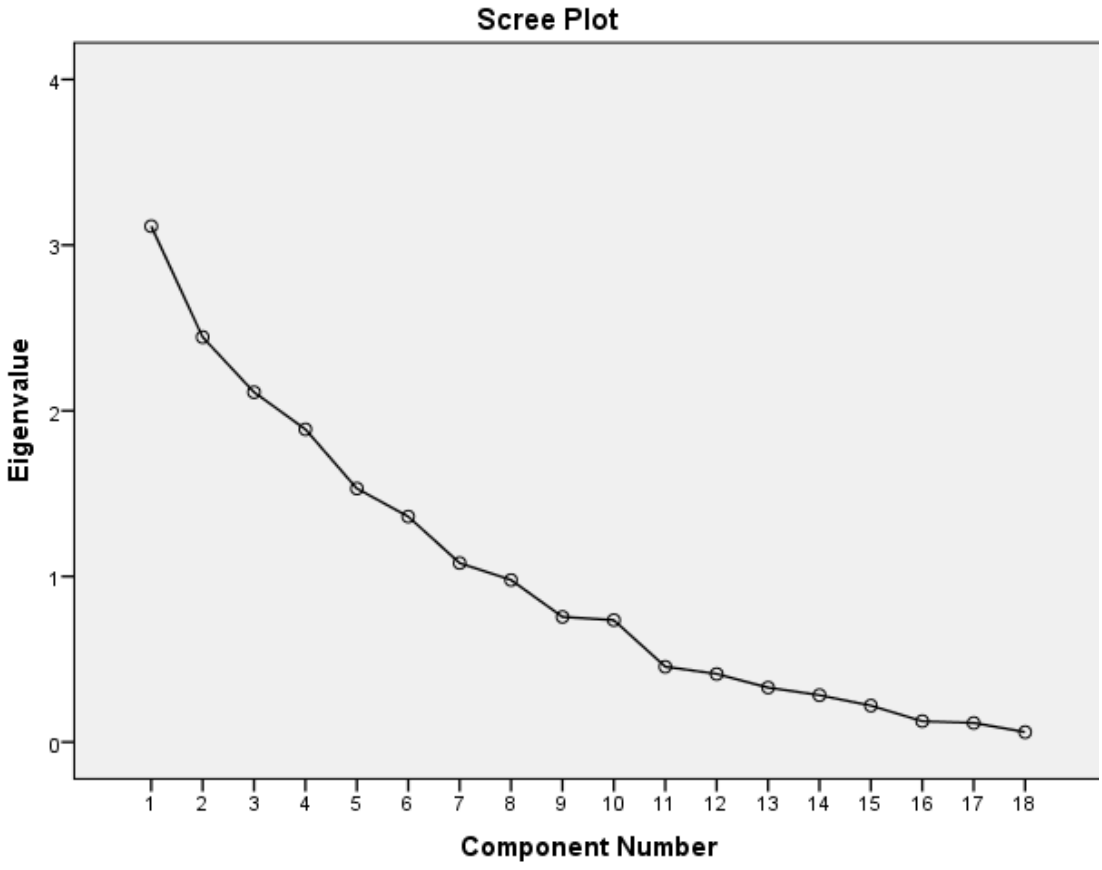
Pearson Korelasyo	K	Ti	P	Na	Mg	Fe	Ca	Si	Mn
K	1								
Ti	-0,109	1							
P	-0,185	0,157	1						
Na	0,038	-0,029	-0,205	1					
Mg	-0,148	0,434*	0,029	-0,024	1				
Fe	0,014	-0,102	0,089	0,107	-0,175	1			
Ca	-0,212	-0,016	-0,352*	-0,226	0,372*	-0,055	1		
Si	0,151	0,078	-0,475**	0,072	0,127	0,363*	0,378*	1	
Mn	-0,011	0,440*	0,323	-0,370*	-0,022	0,186	-0,235	0,214	1
Al	0,147	0,132	0,002	-0,195	0,268	0,118	0,432*	0,234	0,122
Zn	-0,136	0,099	0,016	0,081	0,194	0,067	0,246	0,127	-0,337
Ba	0,244	-0,161	-0,235	-0,055	-0,148	0,019	0,233	0,220	-0,015
Co	0,098	0,390*	0,002	-0,436*	0,316	-0,079	0,191	0,175	0,407*
Nb	0,230	-0,094	0,130	-0,213	-0,128	0,055	-0,125	0,259	0,268
Rb	-0,068	-0,152	-0,284	0,259	-0,047	-0,143	-0,129	0,032	-0,180
Sr	-0,065	0,151	-0,241	-0,030	0,234	-0,291	0,365*	0,261	-0,127
Pb	-0,215	-0,080	-0,059	0,045	-0,173	-0,013	-0,078	-0,253	-0,030
Cr	0,126	-0,312	0,035	-0,095	-0,182	-0,127	-0,044	0,051	-0,103

Çizelge 4.2'nin devamı

Pearson Korelasyonu	**Al**	**Zn**	**Ba**	**Co**	**Nb**	**Rb**	**Sr**	**Pb**	**Cr**
K									
Ti									
P									
Na									
Mg									
Fe									
Ca									
Si									
Mn									
Al	1								
Zn	0,075	1							
Ba	0,343	-0,225	1						
Co	0,435*	0,161	-0,001	1					
Nb	0,246	-0,068	0,075	0,329	1				
Rb	-0,251	-0,242	-0,108	-0,447*	-0,252	1			
Sr	0,180	0,066	0,390*	0,226	0,225	-0,061	1		
Pb	0,059	-0,065	0,269	-0,152	-0,155	0,080	0,222	1	
Cr	-0,386*	-0,042	-0,113	0,004	0,216	-0,059	-0,007	-0,614**	1

Çizelge 4.2'de yapılan Pearson korelasyon analizi sonucuna göre Si elementi ile p elementi arasında korelasyon sayısı -0,475 çıktığından dolayı yüksek mertebeden negatif düşük korelasyon ilişkisi vardır. Cr elementi ile Pb elementi arasında korelasyon -0,614 çıktığından dolayı yüksek mertebeden negatif orta korelasyon ilişkisi vardır. Mg ve Ti arasındaki korelasyon değeri 0,434 çıktığından dolayı düşük mertebeden pozitif düşük korelasyon ilişkisi vardır. Mn ve Ti arasındaki korelasyon değeri 0,440 çıktığından dolayı düşük mertebeden pozitif düşük korelasyon ilişkisi vardır. Co ve Ti arasındaki korelasyon değeri 0,390 çıktığından dolayı düşük mertebeden pozitif düşük korelasyon ilişkisi vardır. Ca ve P arasındaki korelasyon değeri -0,352 çıktığından dolayı düşük mertebeden negatif düşük korelasyon ilişkisi vardır. Mn ve Na arasındaki korelasyon değeri -0,370 çıktığından dolayı düşük mertebeden negatif düşük korelasyon ilişkisi vardır.

4.4. Faktör Analizi



Şekil 4.18. EADF Sedimanın verilerinin Scree Plot grafiği

Şekil 4.18'deki grafik verilerin kaç faktör altında toplandığının bilgisini veren ilk grafiktir. Doğrunun eğiminin 0 a yaklaştığı nokta verilerin kaç faktör altında toplandığını veren noktadır. Eğimin 0'a yaklaştığı nokta olarak 7'nciden sonra eğim azalıyor. Buradan hareketle verilerin kesin olmamakla birlikte 7 faktör altında toplanabiliyor olacağının bilgisini verir. Bu kesinlik Çizelge 4.3'deki açıklanan toplam varyans analizi tablosu ile belirlenir.

Çizelge 4.3. EADF sediman verilerinin açıklanan toplam varyans tablosu

Açıklanan Toplam Varyans									
	İlk özdeğerler			Kare yüklemelerin çıkarma toplamları			Kare yüklemelerin döndürme toplamları		
Bileşen	Toplam	Varyansın %si	Kümülatif %	Toplam	Varyansın %'si	Kümülatif %	Toplam	Varyansın %'si	Kümülatif %
1	3,114	17,298	17,298	3,114	17,298	17,298	2,547	14,149	14,149
2	2,445	13,581	30,879	2,445	13,581	30,879	2,407	13,370	27,519
3	2,112	11,733	42,612	2,112	11,733	42,612	2,036	11,311	38,830
4	1,888	10,488	53,100	1,888	10,488	53,100	1,952	10,847	49,677
5	1,530	8,502	61,602	1,530	8,502	61,602	1,625	9,030	58,707
6	1,361	7,562	69,164	1,361	7,562	69,164	1,564	8,689	67,397
7	1,080	6,002	75,167	1,080	6,002	75,167	1,399	7,770	75,167
8	0,979	5,437	80,604						
9	0,756	4,198	84,802						
10	0,736	4,088	88,890						
11	0,455	2,527	91,417						
12	0,411	2,285	93,702						
13	0,329	1,827	95,529						
14	0,283	1,573	97,102						
15	0,220	1,223	98,325						
16	0,126	0,700	99,025						
17	0,115	0,641	99,666						
18	0,060	0,334	100,000						
Ekstraksiyon Yöntemi: Temel Bileşen Analizi.									

Çizelge 4.3'de veriler 7 faktör altında toplanmıştır. Kümülatif değeri %75.16'dır çıkmıştır. 1. Faktör tüm faktöriyelin %14,149'unu karşılarken 1 ve 2. Faktörler tüm faktörlerin %27,51 ini karşılıyor. 1,2 ve 3. Faktörler tüm faktörlerin %38.83'ünü karşılarken 1,2,3, ve 4. Faktör tüm faktörlerin %49,67'sini karşılıyor. 1,2,3,4,5,6 ve 7.

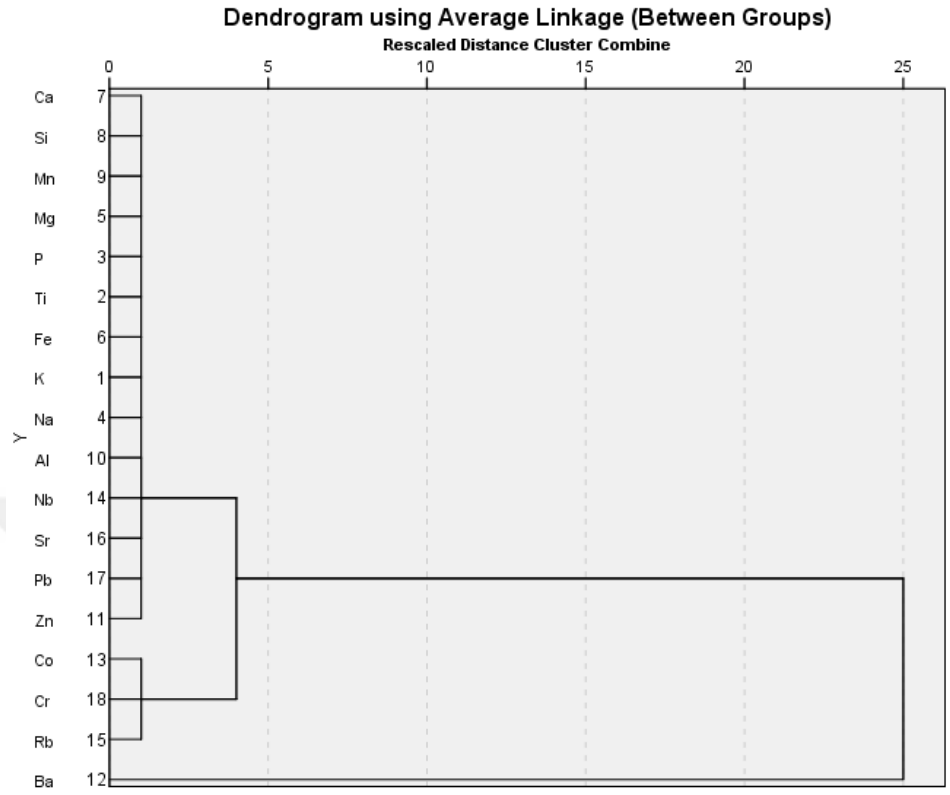
Faktörler tüm faktörlerin %75,16 karşılamaktadır. Kümülatif değerin 75 den büyük olması önem arz etmektedir. Kullandığımız veriler için kümülatif değer %75,16 olduğundan dolayı kabul edilebilir varyans oranıdır.

Çizelge 4.4. EADF sediman verilerinin bileşen matris tablosu

Bileşen matris[a]							
	Bileşen						
	1	2	3	4	5	6	7
Co	0,783	-0,282	-0,016	-0,094	-0,114	0,054	0,102
Al	0,698	0,147	-0,005	0,273	0,106	-0,209	0,118
P	-0,050	-0,719	-0,291	0,077	-0,103	-0,287	0,035
Mn	0,376	-0,633	0,020	0,353	0,188	0,364	-0,258
Ca	0,531	0,562	-0,067	-0,265	-0,054	-0,201	-0,385
Sr	0,461	0,470	0,018	0,037	-0,410	0,156	0,084
K	0,067	-0,038	0,591	0,092	0,045	0,142	0,588
Nb	0,380	-0,311	0,527	0,079	-0,129	-0,045	0,144
Ti	0,442	-0,248	-0,521	-0,038	0,179	0,454	0,198
Mg	0,473	0,136	-0,487	-0,370	0,042	0,236	0,088
Pb	-0,110	0,299	-0,363	0,729	-0,147	-0,162	0,013
Cr	-0,164	-0,226	0,565	-0,601	-0,294	0,010	-0,147
Ba	0,274	0,423	0,361	0,519	-0,193	-0,061	0,020
Fe	0,025	-0,123	0,207	0,201	0,774	-0,334	-0,196
Si	0,471	0,317	0,430	-0,127	0,524	0,230	-0,188
Rb	-0,493	0,340	-0,036	0,021	0,073	0,526	-0,200
Zn	0,194	0,166	-0,270	-0,474	0,211	-0,511	0,256
Na	-0,437	0,364	-0,050	-0,069	0,409	0,135	0,467
Ekstraksiyon Yöntemi: Temel Bileşen Analizi.							
a. 7 Bileşen çıkarıldı.							

Çizelge 4.4'de bu faktörleri açıklayan elementler içeriklerin neler olduğunu vermektedir. Bu tablonun da 7 sütuna bölündüğü görülmüştür. Co için baktığımızda en yüksek değerini veren 1.faktördür. Yani Co 1. Faktörün elemanıdır. 1. Faktörü oluşturan elementler Co, Al, Mn, Mg. 2. Faktörü oluşturan elementler Ca, Sr. 3. Faktör oluşturan elementler Nb, Cr, K. 4. Faktörü oluşturan elementler P, Pb, Ba. 5. Faktörü oluşturan elementler Fe, Si. 6. Faktörü oluşturan elementler Ti, Rb. 7. Faktörü oluşturan elementler Zn, Na.

4.5. Kümeleme Analizi



Şekil 4.19. EADF Sediman verilerinin dendogram grafiği

Dendegrom grafiğinin üst kısmındaki skala 0'dan 25' kadar çıkmıştır. Burada 0 ile 5 arasında çıkan ilişkilerin yani kümelemelerin daha yüksek benzerlik oranlarına sahip iken 0 ile 10 arasında bundan biraz daha az benzerlik oranına sahiptir. 0 ile 15 arasında daha düşük yani orta, 0ile 20 arasındaki benzerlik oranı zayıf, 0 ile 25 aralığına düşen kümelemelerin benzerlik oranı ise daha da zayıftır yani burada güçsüz bir kümeleme oluşturmuştur. Buna göre 7'inci, 8'inci, 9'uncu, 5'inci, 3'üncü, 2'inci, 6'ıncı, 1'inci, 4'üncü, 10'uncu, 14'üncü, 16'ıncı, 17'inci, 11'inci lokasyonlar 0 ile 5 arasında kümelenmiş ve yüksek bir benzerlik göstermektedir. Bunlara ek olarak bu küme içerisinde yer alan 14'ün dışarıdan 18 ile yüksek benzerliği vardır. 18numaralı örneğin diğer lokasyonların örnek numaralarıyla benzer ilişkisi olmadığından sadece 14 ile ilişkisi olduğundan 18 numaralı lokasyon bu kümenin içerisinde yer alamamıştır. 18 aynı zaman 13 ve 15 ile benzer ilişki göstermiş bunlarda kendi aralarında kümelenmişlerdir. 12 numaralı örneğin hiçbir örnek numarası ile kümelenmesi yoktur. Dışarıdan sadece 17 numaralı örnek ile bir bağlantısı var. Buda çok zayıf bir ilişki

çünkü 0 ile 25 aralığındadır. Buradan hareketle deri fabrikası için kimyasal analiz verilerine göre 3 tane kümelenme göstermiştir.1. küme 11'den 7'ye kadar olan yer, 2.kümelenme 13, 15, 18 kısmı 3.kümelenme ise 12 tek başına bir kümelenme göstermiştir.

5. SONUÇLAR

Çok değişkenli istatistiksel teknik, nesnelerin, bağımlı ve bağımsız değişken grupları arasında bağlantı, ilişki kurabilmek için kullanılan bir tekniktir (Timm, 2002). Çok değişkenli istatistik farklı anlarda kullanılmaktadır. Araştırmacılar verilerine göre farklı çok değişkenli istatistik yöntemlerinden yararlanmıştır.

Sağlık Bilimleri, Biyoloji, Ekonomi, Jeoloji, Kimya, Çevre ve Doğa Bilimleri, Mühendislik, Eğitim vb. gibi çok geniş alanlarda kullanılan çok değişkenli istatistik yöntemler bazı varsayımlara sahiptir. Bu varsayımların en başında gelen çoklu normal dağılımdır.

Bu yüksek lisans tez çalışmasının birinci amacı olan çok değişkenli istatistiksel metotların matematiksel alanda incelenmesi ve birçok değişkenli istatistiğe ait olan veriler ile bir uygulama yapılarak verilerin uygunluğuna bakılmıştır. İlk olarak veri setimiz Elazığ'ın Ağın ilçesinde bulunan deri fabrikası civarından alınmış sediman örneklerinin verileridir. Modelimizin öncelikle bulgular ve tartışma kısmında bulunan Çizelge 4.1.1'de tanımlayıcı istatistik değerlerine bakılmıştır. Her bir kimyasal elementin ortalama, standart sapma, çarpıklık, basıklık, minimum ve maksimum değerleri hesaplanmıştır.

EADF Sediman örneklerine ait kimyasal verilerin çok değişkenli istatistik yöntemlerinde analizini yapmadan önce veri setindeki varsayımların sağlanıp sağlanmadığı incelenmiştir. Burada veri setimizdeki değerler 30 dan fazla olduğu için merkezi limit testine göre normal dağıldığı belirlendi.

Çizelge 4.3'e göre incelenen verilerden toplamda 7 tane faktör üretilmiştir. Çizelge 4.3'e göre ilk öz değerlerdeki toplam değeri 1 in üzerinde olanların her birini bir faktör olarak kabul etmiştir. Böylelikle incelenen veri seti ile 7 tane faktör üretilmiştir. Yapılan bu sonuçta kümülatif değer %75 çıktığı için açıklanan varyans değeri verilerin güvenli ve geçerli olduğunu göstermektedir.

Şekil 4.19'daki dendogram grafiğinde 0 ile 5 arasında çıkan ilişkilerin yani kümelemelerin daha yüksek benzerlik oranlarına sahip iken 0 ile 10 arasında bundan biraz daha az benzerlik oranına sahiptir. 0 ile 15 arasında daha düşük yani orta, 0ile 20

arasındaki benzerlik oranı zayıf, 0 ile 25 aralığına düşen kümelemelerin benzerlik oranı ise daha da zayıftır yani burada güçsüz bir kümeleme oluşturmuştur. Burada 7'inci, 8'inci, 9'uncu, 5'inci, 3'üncü, 2'inci, 6'ıncı, 1'inci, 4'üncü, 10'uncu, 14'üncü, 16'ıncı, 17'inci, 11'inci lokasyonlar 0 ile 5 arasında kümelenmiş ve yüksek bir benzerlik göstermektedir. Buradan hareketle deri fabrikası için kimyasal analiz verilerine göre 3 tane kümelenme göstermiştir.1. küme 11'den 7'ye kadar olan yer, 2.kümelenme 13, 15, 18 kısmı 3.kümelenme ise 12 ise tek başına bir kümelenme göstermiştir.

Çok değişkenli istatistikler farklı disiplinlerde oldukça yaygın olarak kullanılmaktadır. Matematiksel modelleme üzerine yapılan bu tez literatüre fayda sağlayacak niteliktedir. Bu veri üzerine daha önceden herhangi birçok değişkenli istatistiksel model çalışılmamıştır.

6. KAYNAKLAR

Afıfı, A.A., CLARK,V., 1997. Computer- Aided Multivariate Analysis, New York.

Akbulut, Ö. Tablolar ve Grafikler, Biyoistatistik, 1-21.

Aldemir, S. 2019. Türkiye'deki İllerin Hayvancılık İstatistikleri Bakımından Çok Değişkenli Analiz Teknikleri İle İncelenemesi. Yükseklisans tezi, Afyon Kocatepe Üniversitesi, Afyonkarahisar, 113 s.

Albayrak, A., S., Eroğlu, A., Eroğlu,Ş., Küçüksille,E., AK, B., Karaatlı, M., Ünlü, H., N., Kesk, Çiçek, E., Kayış,A., Öztürk,E., Antalyalı, Ö., Uçar, N., Demirgil, H., İşler,D. B., Sungur, O., 2005. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri.Asil Yayın Dağıtım Ltd. Şti, Ankara.

Albayrak, A. S. 2003. Türkiye'de İllerin Sosyoekonomik Gelişmişlik Düzeylerinin Çok Değişkenli İstatistik Yöntemlerle İncelenmesi. Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Alpar, R. 2011. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Detay Yayıncılık. Ankara.

Anonim1:https://acikders.ankara.edu.tr/pluginfile.php/73919/mod_resource/content/2/Unite_1.pdf , [Son erişim tarihi: 06.09.2022]

Anonim 2: https://tr.wikipedia.org/wiki/Kovaryans, [Son erişim tarihi: 06.09.2022]

Anonim 3: https://tr.wikipedia.org/wiki/Korelasyon, [Son erişim tarihi: 06.09.2022]

Anonim 4: https://dergipark.org.tr/tr/download/article-file/100977, [Son erişim tarihi: 06.09.2022]

Anonim5:https:///C:/Users/pc/Downloads/Jeodezide%20kestirim%20y%C3%B6ntemleri%20V.%20Hafta%20(3).pdf, [Son erişim tarihi: 06.09.2022]

Anonim6:http://biyoinformatiktr.blogspot.com/2013/04/testin-gucu-power-test.html#:~:text=G%C3%BC%C3%A7%20analizi%2C%20hipotez%20testlerinde%20genellikle,%C3%B6rneklem%20b%C3%BCy%C3%BCkl%C3%BC%C4%9F%C3%BCn%C3%BC%20belirlemek%20i%C3%A7inde%20kullan%C4%B1l%C4%B1r, [Son erişim tarihi: 06.09.2022]

Arslan, O. 2008. Su Kalitesi Verilerinin CBS ile Çok Değişkenli İstatistik Analizi (Porsuk Çayı Örneği). *Jeodezi, Jeoinformasyon ve Arazi Yönetimi Dergisi* 2008/2, Sayı 99.

Bayram, Nuran, 2009. Sosyal Bilimlerde SPSS ile Veri Analizi, Ezgi Kitabevi, Bursa.

Blinstrub, M.J. 2002. Spatial and Temporal Differences in Surface Water Quality in the Newport Bay Watershed, Orange County, California, from 1977 through 2000. M. S. Thesis California State Univercity, Fullerton

Bulut, H. 2014. Çok Değişkenli İstatistiksel Analizde Robust İstatistiklerin Kullanımı. Yükseklisans tezi, Ondokuz Mayıs Üniversitesi, Samsun, 55 s.

Cochrane, D., G.H. Orcutt 1949. Application of Least Squares Regressions to Relationships Containing Autocorrelation Error Term, *Journal of the American Statistical Association*, Vol. 44, s. 32-61.

Çamdeviren, H., 2000. Lojistik Regresyon ve Diskriminant Analizi, Doktora Tezi, Ankara Üniversitesi, Ankara, 89-91 s.

Çelik, H. C. 2004. Çok Değişkenli İstatistiksel Yöntemlerden Kümeleme Yöntemi Ve Kronik Sigara İçiciler Üzerine Bir Uygulama. Doktora tezi, Dicle Üniversitesi, Diyarbakır, 168 s.

Çilan, Ç. A. (2005). Kalite kontrol diyagramlarında varsayımların sağlanması ve cam sanayinde bir uygulama, VII. Ulusal Ekonometri ve statistik Sempozyumu, Alındıgı tarih: 10.08.2011 adres: http://www.ekonometridernegi.org/bildiriler/o7s1.pdf

Çörek, E. T. ve Akın, H. B. 2005. Müşteri Memnuniyetinde İstatistiksel Yöntemler Ve Bir Uygulama. Öneri, Cilt 6, Sayı 23, 259-266 s.

Durbin, J. 1960. Estimating of Parameters in Time Series Regression Models, *Journal of the Royal Statistical Society,* Ser. B, Vol. 22, s. 139-153.

Demirci, E. 2017. İstatistikte Geometrinin Uygulanması. Yükseklisans tezi, Ordu Üniversitesi, Ordu, 76 s.

Everitt, B.S. 1979. A Monte Carlo Investigation of the Robustness of Hotelling's One and Two Sample T2 Tests, *Journal of the American Statistical Association,* 74, s. 48-51.

Gatty, R., 1966. Multivariate Analysis for Marketing Research. An Evaluation, Applied Statistics, 15: 3,157-172.

Glass, G.V., P.D. Peckham, and J.R. Sanders 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance, Review of Educational Research, 42, s. 237-288.

Gnandesikan, R. 1977. Methods for Statistical Analysis of Multivariate Observations, John Wiley, New York.

Gujarati, D.N., 1995. Basic Econometrics, 3rd Ed., McGraw-Hill, New York.

Han, Jiawei, KAMBER, Micheline ve PEI, Jian, (2012), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers is an imprint of Elsevier, USA

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. 1998. Multivariate Data Analysis, Fifth Edition, Prentice Hall International Editions, New Jersey.

Hamarat, Bahattin, 1998. "Türkiye'de Sağlık Açısından Homojen İl Gruplarının Belirlenmesine İlişkin İstatistiksel Bir Yaklaşım", Yüksek Lisans Tezi, Anadolu Üniversitesi Fen Bilimleri Enstitüsü, Eskişehir.

Helberg, Clay 2002 Pitfalls of Data Analysis, http://www.execpc.com/helberg/pitfalls.

Holloway, L.N., O.J. Dunn 1967. The Robustness of Hotelling's T2, *Journal of the American Statistical Association,* 62, s. 124-136.

Hüyüktepe, B. 2018. Türkiye'deki İllerin Sosyo-Ekonomik Gelişmişlik Göstergelerinin Çok Değişkenli İstatistiksel Yöntemlerle İncelenmesi. Yüksek lisans tezi, Kütahya Dumlupınar Üniversitesi, Kütahya, 106 s.

Jama Abdı, S. 2017. Unıversıty Websıte Desıgn Wıth Multıvarıate Statıstıcal Technıques In Kanseı Engıneerıng. Master Thesis, Anadolu University,

Eskişehir, 67 s.

Johnson, R.A, Wichern D.W, 1992. Applied Multivariate Statistical Analysis, Prentice Hall, NJ.

Johnson, Richard A. ve WICHERN Dean W., 2007. Applied Multivariate Statistical Analysis, Pearson Prentice-Hall, New Jersey.

Johnston, J. 1984. Econometric Methods, 3rd ed., McGraw-Hill, New York.

Jollıffe, I.T., 2002. Principal Component Analysis, Springer, New York.

Joseph F., Hair, J.R., Robert P. B. ve David J. O. 2002. Marketing Research Within a Changing Information Environment, The McGraw-Hill Companies, New York.

Kalaycı, Ş., 2010. SPSS Uygulamalı Çok Değişkenli İstatisitik Teknikleri, Asil Yayın Dağıtım LTD. ŞTİ., Ankara.

Karels, V. G., Prakash, J. A., 1987. 'Multivariate Normality and Forecasting of Business Bankruptcy', J. Of Business Finance &Accounting, 14(4), pp. 573-593.

Klecka, W. R., 1980. Discriminant Analysis, Sage Pub., Beverly Hills.

Köksal, G. 2019. Matematiksel Düşünmenin Matematik Kaygısı Üzerine Etkisinin Çok Değişkenli İstatistiksel Yöntemlerle İncelenmesi. Yükseklisans tezi, Çukurova Üniversitesi, Adana, 106 s.

Mertler, C.A. ve Vannatta Reınhart, Rachel 2017. Advanced and Multivariate Statistical Methods Practical Application and Interpretation, Routledge Taylor & Francis Group, New York and London.

Mardia, K.V. 197. The Effect of Non-Normality on Some Multivariate Tests and Robustness to Non-Normality in the Linear Model, Biometrika, 58, s. 105-212.

Newbold, P. 2001. İşletme ve İktisat İçin İstatistik, (çev.Ü. Şenesen), Prentice Hall.

Norusis, M. J., and SPSS Inc. 1993. SPSS for Windows: Base System User's Guide, Rel.

Orhunbilge, N., 2000. Tanımsal İstatistik Olasılık ve Olasılık Dağılımları, Avcıol-Basım, İstanbul.

Olson, C.L. 1974, Comparative Robustness of Six Tests in Multivariate Analysis of Variance, *Journal of American Statistical Association*, 69 (348), s.894-907

Orçanlı, K. 2017. Çok Değişkenli Kalite Kontrol Grafikleri ve Yapay Sinir Ağları Yöntemiyle Döküm Sanayisinde Bir İstatistik Süreç Kontrol Uygulaması. Yayımlanmamış Doktora Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Erzurum.

Özkale M. R. 2004. "İstatistiksel Kalite Kontrol Yöntemleri ve Uygulamalar", Yüksek Lisans Tezi, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü.

Özdamar, Kazım, 2013. Paket Programlar ile İstatistiksel Veri Analizi Cilt 2, Nisan Kitabevi, Eskişehir.

Özdamar, K., 1999. Paket Programlar ile İstatistiksel Veri Analizi, Kaan Kitabevi, Eskişehir.

Pamuk, M. 2005. Öğrencilerin Öğretim Üyesini Değerlendirmesine Ait Bir Uygulama. Ekonometri ve İstatistik Sayı:1, 15-05, İstanbul.

Polat, E., 1995. Türk Bankacılık Sisteminde Problemli Kredileri Önceden Belirleyecek Model Geliştirilmesi İçin Bir Uygulama, Pamukbank T. A. Ş. Eğitim Yayınları, İstanbul, 66.

Sharma, S., 1996. Applied Multivariate Techniques. John Wiley & Sons, Inc, New York.

Shaw, P. J. A., 2003. Multivariate Statistics for the Environmental Sciences. Hodder Arnold, New York.

Sheth, J.N., 1971. The Multivariate Revolution in Marketing Research. *Journal of Marketing*, 35,13-19.

Shın, K., 1996. SPSS Guide for DOS Version 5 and Windows 6.1.2., Chicago.

Salmona, M.Ö.A. 2004. Multivariate Statistical Quality Control: An Industrial Application, (yüksek lisans tezi), Marmara Üniversitesi, İstanbul.

Sağır, S. 2020. Üniversite Hastanesi Acil Servisine Gelen Hasta Sayılarına Uygun İhtimal Dağılımlarının Belirlenmesi Ve İstatistiksel Analiz. Yüksek lisans tezi, Sivas Cumhuriyet Üniversitesi, Sivas, 117 s.

Tatlıdil, H., 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz, Cem Web Ofset Ltd. Şti., Ankara.

Tımm, N.H., 2002. Applied Multivariate Analysis. Springer -Verlag, New York.

Tacg, J. 1997. Multivariate Techniques in Social Sciences, Sage Pub. Ltd., London.

Toktay, Y. 2017. Çok Değişkenli İstatistik Analiz Yöntemler: Faktör Analizi Ve Diskriminant Analizinin Iğdır Üniversitesi Öğrencileri Üzerine Uygulaması. Yüksek lisans tezi, Iğdır Üniversitesi, Iğdır, 104 s.

Tatlıdil, Hüseyin, 1992. Uygulamalı Çok Değişkenli İstatistiksel Analiz, Hacettepe Yayınları, Ankara.

Tabachnıck, Barbara G. ve Fıdell Linda S., 2015. Çok Değişkenli İstatistiklerin Kullanımı, Baloğlu, Mustafa, (Çeviri Editörü), Nobel Akademik Yayıncılık, 6.Baskı, Ankara.

Tekin, B. 2015. Temel Sağlık Göstergeleri Açısından Türkiye'deki İllerin Gruplandırılması: Bir Kümeleme Analizi Uygulaması. *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, Cilt 5, Sayı 2, 389-416 s.

Tatlı, H. 2015. Akıllı Telefon Seçiminin Belirleyicileri: Üniversite Öğrencileri Üzerine Bir Uygulama. *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, Cilt 5, Sayı 2, 549-567 s.

Tatsuoka, M. M., 1976. 'Discriminant Analysis', Data Analysis Strategies and Designs for Substance Abuse Research, pp. 201-220.

Tavşancıl, E., 2002. Tutumların Ölçülmesi ve SPSS ile Veri Analizi. Ankara: Nobel Yayıncılık.

Tümer, M., 2001. Kuzey Kıbrıs Türk Cumhuriyeti İmalat Sanayinde Faaliyet Gösteren Kobileri Ayrıştıran Faktörlerin Tespiti, Doğu Akdeniz Üniversitesi, 296-303 s.

Ülen, M. (2010). Çok Değişkenli İstatistiksel Kalite Kontrolünün İlaç Endüstrisine Uygulanması. Yayımlanmamış Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Ünlükaplan, Y. 2008. Çok Değişkenli İstatistiksel Yöntemlerin Peyzaj Ekolojisi Araştırmalarında Kullanımı. Doktora tezi, Çukurova Üniversitesi, Adana, 143 s.

Webster, A. 1995, Applied Statistics for Business and Economics, 3rd ed.

Yavuz, S. 2009. Hatalari Ardişik Bağimli (Otokorelasyonlu) Olan Regresyon Modellerinin Tahmin Edilmesi, *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 23(3), 123-140 s.

Yılmaz, H. 2012. Çok Değişkenli İstatistiksel Süreç Kontrolü: Bir Hastane Uygulaması Yükseklisans tezi, İstanbul Teknik Üniversitesi, İstanbul, 138 s.

Yılmaz, V. 2009. Türkiye Akarsuları Su Kalitesi Parametrelerinin Çok Değişkenli İstatistiksel Analiz Yöntemleri ile İncelenmesi. Yüksek lisans tezi, Selçuk Üniversitesi, Konya, 94s.

ÖZGEÇMİŞ

Esra ŞİMŞEK

ÖĞRENİM BİLGİLERİ

Yüksek Lisans 2018-2022	Akdeniz Üniversitesi, Fen Bilimleri Enstitüsü, Matematik A.B.D., Antalya
Lisans 2014-2018	Akdeniz Üniversitesi, Fen Fakültesi, Matematik Bölümü, Antalya

ESERLER

Uluslararası hakemli dergilerde yayımlanan makaleler

1- Simsek E. Ve Yalcin F. (2022). Assumptıons And Examınatıon Of Multıvarıate Statıstıcal Technıques, ISADET 2022 (Özet Bildiri)