KAYIP VERİ ELE ALMA YÖNTEMLERİNİN t-TESTİ VE ANOVA PARAMETRELERİ ÜZERİNE ETKİSİNİN İNCELENMESİ [*]

İbrahim Alper KÖSE [**]
Begüm ÖZTEMUR [***]

ÖZET
Bu araştırmanın amacı, kayıp veri sorunu giderme yöntemlerinin t-testi ve ANOVA parametreleri üzerine etkisinin incelenmesidir. Araştırma 50, 100, 200, 400 birimlik yapay veri setleri üzerinden yürütülmüştür. Veri setleri düşük ve yüksek korelasyonlu normal dağılıma uygun olarak oluşturulmuştur. %5, %10, %20 kayıp olacak şekilde rastgele koşullar altında eksiltilmiş veriler Tamamıyla Rassal Olarak Kayıp (TROK) yapısına uygun oluşturulmuştur. Türetilen veri setlerine kayıp veri giderme yöntemlerinden silme, yerine ortalama koyma, regresyon ve beklenti maksimizasyonu yöntemleri uygulanmıştır. Çalışma sonucunda kullanılan yöntemlerin ortaya koyduğu değerler farklı korelasyona ve farklı büyüklükteki veri setlerinde oldukça değişiklik göstermiştir. Düşük birimli veri setlerinde regresyon ve Beklenti Maksimizasyonu (BM) yöntemleri en yakın sonuçları verirken, yüksek birimli veri setlerinde regresyon ve yerine ortalama koyma yöntemi tam veri setlerine uygulanan analiz değerleriyle daha tutarlı sonuçlar vermiştir.
Anahtar Kelimeler: Kayıp Veri Analizi, Atama Yöntemleri, BM, Yerine Ortalamayı Koyma, Regresyon.

EXAMINING THE EFFECT OF MISSING DATA HANDLING METHODS ON THE PARAMETERS OF t-TEST AND ANOVA

ABSTRACT
The purpose of this study was to examine the effect of missing data handling methods on the parameters of t-test and ANOVA. The study was conducted with simuated data sets. These data sets were produced in a way that they would have normal distributions in high and low correlation and their sizes were 50, 100, 200, 400 units. Under random conditions, data sets were reduced %5, %10, %20 in the form of MCAR. In the simulated data sets, mean substitution method, regression method, expectation-maximization (EM) method and deletion method were applied. Results showed that in different sample sizes and correlations, findings were differentiated. It is observed that in data sets with low sample sizes, regression and EM application were usefull on the other hand in data sets with larger sample sizes, mean substitution method instead of regression method had more consistent results.
KeyWords: Missing Value Analysis, ImputationMethod, EM, MeanSubstitutionMethod, Regression.

---

[*] "Kayıp Veri Ele Alma Yöntemlerin Varyans Analizi Parametreleri Üzerine Etkisinin İncelenmesi" isimli tez çalışmasının bir bölümüdür.
[**] Yrd. Doç. Dr., AIBU Eğitim Fakültesi, Eğitim Bilimleri Bölümü, i.alper.kose@gmail.com
[***] MEB, Düzce Endüstri Meslek Lisesi, matbegum@hotmail.com

1.GİRİŞ

Yapılan birçok araştırmada verilerde kayıplar olabilmektedir. Bu kayıp verilerin araştırma sonuçlarını etkilediği ve en yaygın problemlerden birisi olduğu bilinmektedir. Kayıp veri içeren veriler ile yapılan analizlerin doğru ve güvenilir olmadığı bilinen bir gerçektir. Stekhoven&Bühlmann (2011)'a göre istatistiksel analizler için veri setlerinin eksiksiz (tam) olması büyük önem taşımaktadır; çünkü çoğu istatistiksel veri analizi paket programları verilerin kayıpsız olduğu varsayımına dayanmaktadır. İstatistiksel kestirimlerde kayıp veriler olası bir yanlılık durumunu doğurmaktadır. Veri toplamadaki başarısızlık analiz için geçerli durum sayısının azalmasına, bu azalma değerli bilgi kaybının oluşmasına sebep olmaktadır. Kayıp veri probleminin ciddiyeti verilerin ne kadar ve neden kayıp olduğuna bağlıdır (Tabachnick ve Fidell, 2001).

Araştırmalarda süreç boyunca tam veri setleri elde etmeye çalışmasına rağmen farklı sebeplerden dolayı veri kayıpları yaşanabilmektedir. Örneğin, hane halkı araştırmalarında aile bireyleri gelirlerini belirtmekten kaçınabilir, bir endüstriyel denemede bazı sonuçlar deney süreci ile ilgisiz mekanik sorunlar nedeniyle kayıp olabilir. Bir kamuoyu araştırmasında bazı bireyler, birkaç aday içerisinden tercihlerini açıkça ifade etmemiş olabilirler (Yazıcı, 2005). Yazıcıoğlu ve Erdoğan (2007)'a göre ise bu eksiklikler deneklerden kaynaklanmaktadır. Kalaycı (2008)'ya göre bazı veri kayıpları da veri toplayan kişiden kaynaklanmaktadır. Bu durumda kayıp veri oluşum süreci, araştırmacının denetimi altındadır ve tanımlanabilir. Bu gibi durumlarda kayıp veri "ihmal edilebilir kayıp veri" olarak adlandırılır ki bu sorunu çözmek için kayıp veri probleminin çözümünde bilinen yöntemlerin kullanılmasına gerek kalmaz. Kayıp verinin ihmal edilebilir olarak değerlendirilmesi için kayıp veri sürecinin rastgele olması, dolayısıyla eldeki verinin tam ve kayıp değerler setinin rastgele örneklemi olmasıdır (Alpar, 2003).

Kayıp veri çalışmaları 1966'da Afiffi ve Elashoff'un araştırmalarıyla başlamıştır. 1977'de Dempster, Laird ve Rubin beklenti maksimizasyonu (BM) üzerinde yoğunlaşmışlardır. 1987'de Rubin'in veri ataması gibi algoritmik hesaplarla kayıp veri sorununa çözüm aranmıştır. 1990'lı yıllarda çoklu atama (multipleimputation) ve Monte Carlo yöntemleri ile Bayes teknikleri üzerine çalışmalar artmıştır. Gildea ve Hofmann bilişim teknolojisi (1999), Iturria ve Blangero genetik dalı (2000), Schneider iklim verileri (2001), Krishner, Cadez, Smyth ve Kamath uzay bilimleri (2003), Evens (2003) doğa bilimleri analarında BM yöntemini kullanmışlardır.

Kayıp verilerin varlığı ve miktarı kadar, içerdiği mekanizma da büyük önem taşımaktadır. Kayıp veri mekanizmasının belirlenmesi ve kayıp veri sorununun giderilmesinde kullanılacak uygun yöntemin seçilip sınıflandırması da son derece önemlidir. Bu konudaki ilk çalışma Rubin (1976) tarafından yapılmıştır. Little ve Rubin (2002) kayıp veri mekanizmalarını;

(1) Tamamıyla Rassal Olarak Kayıp (MissingCompletely at Random, MCAR, TROK)
(2) Rassal Olarak Kayıp (Missing at Random, MAR, ROK)
(3) İhmal Edilemez Kayıp (Noignorable, NI, İEK) olacak şekilde üç kategoriye ayırmıştır.

1.1. Tamamıyla Rassal Olarak Kayıp (MissingComletely At Random, MCAR, TROK): X ve Y farklı değişkenler olsun. Yanıtın X ve Y değişkenlerinden bağımsız olduğunda yani iki değişkenin de gözlenme olasılığı birbirinden etkilenmediğinde; X ve Y'nin gözlenememiş olması tamamen rassal olarak kayıp durumunu oluşturmaktadır (Little ve Rubin, 2002; Allison, 2001). Diğer bir deyişle kayıp olma durumu analizde yer alan özel değişkenlerle ilgili değildir. TROK koşulunun sağlanıp sağlanmadığı, yanıt verenler ile yanıt vermeyenler arasındaki gözlenen verilerin karşılaştırılmasıyla sağlanabilir. Eğer veriler TROK değilse, sonuçlar yanlı olacağından daha güçlü tekniklerin kullanılması uygun olacaktır (Little ve Rubin, 2002; Alpar, 2003; Yazıcı, 2005). Özetle TROK yapısında rastgele oluşmuş bir kayıp durumu, diğer değişken ölçütleriyle ilgisiz olduğu zaman ortaya çıkar (Enders, 2011). Bir başka ifade ile, TROK yapısı veri setlerini oluştururken tamamen istek dışı oluşan kayıplardır. Bir anket çalışmasında soruyu görmeyip cevaplamama veya verilerden bazılarının kaybolması TROK mekanizması olarak nitelendirilebilir (Sezgin & Çelik, 2013). Örneğin, anket çalışmasında anketörün soru sormayı unutması, bir koşu yarışmasında atletin kural ihlali nedeniyle bitişi görememesi, laboratuvar deneyinde numunenin yere düşmesi durumlarında oluşan kayıplar da TROK yapısını oluşturur (Yozgatlıgil vd., 2011).

1.2. Rassal Olarak Kayıp (Missing At Random, MAR, ROK): X ve Y gibi değişkenlerin arasından X'in yanıt olasılığının Y'ye bağlı fakat Y'nin yanıt olasılığının X'e bağlı olmaması hâlinde, bu durumdaki kayıp veri rassal kayıp veri olarak adlandırılır (Allison, 2001). Bir anket çalışmasında örneklem bireylerini oluşturan grubun sorulan soruları bilerek atlaması veya yanlış cevap vermesi ROK olarak nitelendirilebilir (Sezgin ve Çelik, 2013).

1.3. İhmal Edilemez Kayıp/Rassal Olmayan Kayıp (Noignorable, NI, IE): ROK ve TROK olmayan süreçler rastlantısal olmayan kayıp olarak tanımlanır. Bir başka deyişle kayıp olma durumu rassal değildir ve veri tabanındaki bir diğer değişkenden tahmin edilemez (Allison, 2001). IE mekanizması kayıp veri olasılığının diğer değişkenlerle olan ilişkisi doğrultusunda ortaya çıkmıştır (Enders, 2011). Örneğin anketteki bir sorunun yanlış sorulmasından dolayı doğru cevabın çözülemediği durumlardaki kayıp durumu IE olarak tanımlanabilir (Sezgin ve Çelik, 2013). Veri setlerinde, kayıp değerlerin rastgele koşullar altında olup olmadığı belirlendikten sonra bu kayıplılığın giderilmesine yönelik çözümlemeler yapılmasını gerektirir. Veri tabanındaki değerlerin ROK veya TROK olması durumunda kullanılacak yöntemler de farklı olacaktır.

Kayıp Veri Sürecinde Rastgeleliğin Sorgulanması: Baygül (2007) ve Alpar (2003)'a göre kayıp veri analizinde kullanılacak yöntemin belirlenmesi için kayıp veri sürecinin rastgeleliğinin aşağıdaki yöntemlerle irdelenmesi gerekir:

1. Veri setindeki bir değişkene ilişkin gözlemler kayıp veri içerenler ve içermeyenler olarak iki gruba ayrılmalı ve ilgilenilen diğer değişkenlerin aldığı değerler açısından bu iki grup arasında anlamlı bir fark olup olmadığı iki ortalama arasındaki farkın anlamlılığı test eden "t testi" kullanılarak yapılabilir. Anlamlı fark, rastgele olmayan kayıp veri sürecinin varlığını gösterir.
2. Veri setindeki değişkenler kayıp değer içeren ve içermeyenler olmak üzere iki gruba ayrılıp, tam veriler 1, kayıp veriler 0 olarak kodlandıktan sonra değişkenler arasındaki Pearson korelasyon katsayısı hesaplanır. Bulunan korelasyon katsayıları

her bir değişken çifti için kayıp veriler arasındaki ilişki miktarının derecesini belirtir. Küçük korelasyon katsayısı rastgeleliği işaret eder.

3. Little (1998)'ın TROK testi rastgeleliğin araştırılmasında sıklıkla kullanılan bir χ^2 testidir. p< 0, 05 olması durumunda veri yapısının TROK olmadığı sonucuna varılır.

2. Kayıp Veri Ele Alma Yöntemleri: Araştırmacılar silme tekniklerinden Liste veya Durum Bazında Silme (listwiseorcasewisedatadeletion-LD or CD) ve Çiftler Bazında Silme (pairwise data deletion- PD) gibi kayıp değerler problemini ortadan kaldıracak yöntemleri kullanabilirler. Bunların dışında en sık kullanılan atama teknikleri: Kayıp Gözlem ile Tam Gözlemin Yer Değiştirmesi (Case Substitution), Yerine Ortalamayı Koyma (Mean Substitution), Regresyon Ataması (Regression Imputation), Hot Deck Ataması (Hot Deck Imputation), ColdDeck Ataması (Cold Deck Imputation), Beklenen Maksimizasyon Yaklaşımı (Expectation Maximization-EM-approach) ve Çoklu Atama (Multiple Imputation) olarak sıralanabilir.

1.4.1. Liste veya durum bazında silme (tam gözlemlerin kullanılması yöntemi/ listwise deletion-LD): Bu yöntemde sadece tam gözlemler dikkate alınır, kayıp gözlemler dikkate alınmaz. Kayıpsız gözlemler dikkate alındığından kayıp veri sayısının az olduğu durumlarda kullanılması tavsiye edilir. Çok kullanılan bir yöntem olmasının yanında kayıp veri yapısının tamamıyla rassal kayıp (TROK) olması gerekmektedir (Kalaycı, 2008). Sadece tam bilgi taşıyan veri gruplarıyla ilgilenildiği için "Tüm durum metodu" olarak da bilinir. ROK veya IE durumlarında kullanılması yanlı sonuç verecektir.

1.4.2. Kayıp gözlem yerine ortalamayı koyma (mean substitution): Kayıp veri yerine sıkça kullanılan bir strateji olup, kayıp değer içeren değişkenin ortalamasını kayıp değerin yerine koymaktır (Little ve Rubin, 2002). Böylece kişiye ilişkin herhangi bir bilgi bulunmadan, herhangi normal dağılımlı bir değişken için değerlerin en iyi tahminini ortalamalar oluşturacağından tam veriler için elde edilen ortalama, tam verilerle elde edilen ortalamaya eşit çıkacaktır.

1.4.3. Regresyon ataması (regression imputation): Regresyon yönteminin amacı, bir ya da daha fazla bağımsız değişken yardımıyla bağımlı değişken değerinin test edilmesidir. Regresyon atama yönteminde, bağımlı değişken kayıp gözlemli değişken; bağımsız değişkenler ise diğer değişkenlerdir. Bu yöntemin özellikle kayıp verinin sayısının orta düzeyde olduğu ve veride yaygın bir dağılım gösterdiği durumlarda kullanılması önerilmektedir. Bu yöntemin kullanılabilmesi için bağımlı değişkenle bağımsız değişkenler arasında ilişkinin oldukça yüksek olması gereklidir (Alpar, 2003; Kalaycı, 2008). Beklenti maksimizasyon yaklaşımı (expectationmaximization-EM-approach-BM): BM iki adımlı bir yöntemdir. İlk adımda başlangıç değerler kayıp veriye aktarılır; ikinci adımda da bu başlangıç değerleriyle oluşturulan beklentiler yükseltilir. Bu beklenti yükseltme döngüsü ayarlanan değerler daha önceden belirlenen farklı bir kriter üzerine yoğunlaşana kadar defalarca tekrarlanır. BM imputasyon yöntemi parametre değerleri için önyargısız standart hataları oluşturur (Cheema, 2012). Genel olarak BM algoritması,

- Tahmin edilmiş değerleri kayıp verilerin yerine koyar,
- Bu kayıp verileri kullanarak parametre tahmin eder,
- Daha uygun parametre buluncaya dek algoritmayı yineler.

BM algoritması; B-adımı, beklenen ve M-adımı en büyükleme olmak üzere iki adımdan oluşmaktadır. B-adımı: Gözlenemeyen verinin ya da kayıp verinin yerinin doldurulması problemidir. M-adımı: Tahmin edilen kayıp veri değerini kabul ederek oluşan tam veri modeli üzerinden, bilinen en çok olabilirlik tahmini hesaplamaktadır. M-adımı sonucunda ortaya çıkan tahminler, BM algoritmasının çıktısını oluşturmaktadır (Yazıcı, 2005).
Türkiye'de Oğuzlar (2001), ekonometri alanındaki araştırmalarında kayıp değer sorununa yönelik çalışma yapmıştır. Bal (2003) ise sağlık alanında, çok gruplu veri setlerinde eksik gözlem sorununun çözümlemesine yönelik çiftler bazında silme, yerine ortalama koyma, BM ve regresyon yöntemlerini karşılaştırmış, Yazıcı (2005) kayıp veri içeren problemlerde parametre tahmini hesaplamalarında BM algoritması ve uzantılarını tanıtan bir çalışma yapmıştır. Baygül (2007) sağlık alanında kayıp veri yönteminde etkin kullanılan yöntemleri karşılaştırmıştır. Satıcı ve Kadılar (2009) tarafından kayıp gözlem olması durumunda kitle ortalamasına ilişkin Sing ve Horn (2000), Sing ve Deo (2003) ve Rueda vd. (2005) tarafından sunulan tahmin ediciler incelenmiş ve bunlara alternatif iki yeni tahmin edici önerilmiştir. Yozgatlıgil vd. (2011) Türkiye'nin iki farklı iklim bölgesine ait yağış ve sıcaklık serilerinde kayıp veri tahmini için kullanılan yöntemleri karşılaştırmıştır. Çokluk ve Kayrı (2011) bir ölçme aracının yapı geçerliğinin test edilmesini incelemişlerdir.

T-testi ve ANOVA, araştırmalarda en sıklıkla kullanılan istatistiksel tekniklerdir. Ortalamalar arasındaki farkın anlamlılığını test etmek için t-testi ve çoklu gruplar için kullanılan ANOVA varyans analizinin başlıca tekniğidir. Yapılan çalışmalarda ortaya çıkan kayıp verileri ele alma yöntemleri, analizin anlamlılık düzeyini, bu da ortaya çıkabilecek hata türlerini (I. Tip hata, II. Tip hata) etkileyebilmektedir. Bu araştırmada, kayıp veri giderme yöntemlerinin, araştırmalardaki anlamlılık düzeyini nasıl etkilediği, hangi tür kayıp veri setleri için hangi yöntemlerin tercih edilmesi gerektiği ortaya konmaya çalışılmıştır. Ayrıca, araştırmalardaki kayıp verilerin giderilmesine yönelik kullanılan yöntemleri inceleyen çalışmaların sınırlı ve dar kapsamlı olması bu çalışmayı önemli kılan bir diğer noktadır.

2. AMAÇ

Bu çalışmada, normal dağılım gösteren farklı büyüklük ve farklı korelasyondaki tam veri setlerinde kullanılan silme ve atama yöntemlerinin t-testi ve ANOVA parametreleri, 1.tip hata üzerindeki etkisi araştırılmış; hangi yöntemin en etkili olduğu ortaya konmaya çalışılmıştır.

3. YÖNTEM

Araştırma iki ve daha çok sayıdaki değişkenin arasında birlikte değişiminin derecesini belirlemeyi amaçlayan betimsel türde tarama modelidir (Karasar, 2004). Bu çalışmada yapay verilerden yararlanılmıştır. Veri türetimi belirli koşullar göz önüne alınarak R-Studio paket programı yardımıyla elde edilmiştir. Bu araştırmada koşullu türetilmiş veri setleri;

- Çok değişkenli normal dağılıma sahip veri setleri,
- Farklı büyüklükteki veri setleri (n=50, n=100, n=200, n=400) ,

- Değişkenlerdeki kayıp veri sayısı farklı olan veri setleri (%5, %10, %20),
- Değişkenler arasındaki korelasyona göre yüksek (r<-0.70 veya r>0.70) ya da düşük korelasyonlu (-0.30< r < 0.30) veri setleri olarak ayrılmıştır.

Böylece bu araştırmada n=50, 100, 200, 400 birim ve her biri üç değişkenden oluşan toplamda 2250 adet koşullu türetilmiş veri kullanılmıştır. Rastgele türetilen tam veri setlerindeki düşük ve yüksek korelasyona sahip değişkenler 50, 100, 200, 400 birimlik toplam sekiz veri seti içinden sırasıyla %5, %10, %20 gözlem eksiltilerek 24 yeni veri seti oluşturulmuştur. Özetle analizler için toplamda 32 veri seti ile çalışılmıştır. Oluşturulmuş yapay veri periyotlarından yararlanılarak, uygulanılan yöntemlerin başarılarını karşılaştırmak amacıyla kayıp veri tamamlama yöntemlerinden yerine ortalamayı koyma, beklenti maksimizasyonu (BM), çoklu atama teknikleri ile silme yöntemlerinden liste/durum bazında silme teknikleri seçilmiştir.

Türetilen veri setlerindeki kayıp veri mekanizmasının TROK olup olmadığını belirlemek amacıyla veri setlerine Little'ın TROK testi uygulanmış ve sonuçlar Tablo 1.'de özetlenmiştir.

Tablo 1.
Türetilen veri setlerine ait Little'ın TROK (MCAR) testi sonuçları

Veri Setleri	Korelasyon Düzeyleri	Eksik Olma Durumu %	χ^2	sd	P
50 Birimlik Veri Seti	Düşük Korelasyon	5	3.029	6	0.805
		10	1.343	6	0.969
		20	3.465	6	0.749
	Yüksek Korelasyon	5	6.259	6	0.395
		10	6.561	6	0.363
		20	5.314	6	0.504
100 Birimlik Veri Seti	Düşük Korelasyon	5	9.490	6	0.148
		10	5.194	6	0.519
		20	2.361	6	0.884
	Yüksek Korelasyon	5	1.164	6	0.979
		10	1.446	6	0.963
		20	1.586	6	0.954
200 Birimlik Veri Seti	Düşük Korelasyon	5	7.193	6	0.303
		10	4.141	6	0.658
		20	3.527	6	0.74
	Yüksek Korelasyon	5	5.556	6	0.475
		10	5.002	6	0.544
		20	6.454	6	0.374
400 Birimlik Veri Seti	Düşük Korelasyon	5	3.481	6	0.747
		10	2.984	6	0.811
		20	3.600	6	0.731
	Yüksek Korelasyon	5	3.645	6	0.725
		10	1.964	6	0.923
		20	1.201	6	0.977

Tablo 1. incelendiğinde tüm veri setleri için p anlamlılık değeri p>0.05 olduğu için veri setlerinin TROK yapısına uygun olduğu görülmektedir.

4. BULGULAR VE YORUMLAR

Bu çalışmada kayıp veri ele alma yöntemlerinin farklı veri setlerinde t-testine ve ANOVA analizi parametreleri üzerine etkisi incelenmiştir. Bu amaçla öncelikle tam veriler üzerinden t-testi ve ANOVA yapılmış, daha sonra eksik veri setleri belirlenen yöntemlerle tamamlanarak veya silinerek, t-testi ve ANOVA analizleri tekrarlanmıştır. Ulaşılan bulgular ayrı analizler hâlinde tablolaştırılmıştır.

t- Testine Ait Bulgular ve Yorumlar: Normal dağılıma sahip, düşük ve yüksek korelasyonlu farklı oranda kayıp veri içeren, farklı büyüklükteki veri setlerine uygulanan t-testi sonuçlarına ait t ve p değerleri Tablo 2 ve Tablo 3'te verilmiştir. Öncelikle t-testine ait bulgular tablo 2 ve 3'te özetlenmiştir.

Tablo 2.
Düşük korelasyonlu veri setlerine ait t-testinin t ve p değerleri

			TAM		SİLME		ORTALAMA		REGRESYON		BM	
	N	%	t	p	t	P	t	p	t	p	t	p
Düşük	50	0	**0.498**	**0.620**								
		5			0.2688	0.7893	0.702	0.486	0.806	0.424	**0.642**	**0.524**
		10			0.2927	0.7714	0.049	0.961	**0.472**	**0.639**	0.052	0.959
		20			**-0.1823**	**0.8572**	-.085	0.933	0.024	0.981	-0.055	0.957
	100	0	**-0.089**	**0.929**								
		5			-0.6263	0.5328	-0.260	0.795	-0.528	0.598	**-0.229**	**0.820**
		10			0.3887	0.6987	0.192	0.848	**-0.153**	**0.879**	0.213	0.831
		20			0.2715	0.7874	0.030	0.976	-0.009	0.993	**-0.033**	**0.974**
	200	0	**1.158**	**0.248**								
		5			0.8082	0.4201	**1.318**	**0.189**	1.505	0.134	1.340	0.182
		10			**1.299**	**0.1962**	1.388	0.167	1.687	0.093	1.429	0.155
		20			0.9199	0.3604	**1.276**	**0.203**	0.941	0.348	1.376	0.170
	400	0	**-0.652**	**0.515**								
		5			-1.057	0.2911	**-0.747**	**0.456**	-0.927	0.354	-0.758	0.449
		10			-13.107	0.191	-0.901	0.368	**-0.606**	**0.545**	-0.895	0.372
		20			-13.637	0.1746	-1.160	0.247	**-0.617**	**0.538**	-1.182	0.238

Tablo 2 ve Tablo 3'te koyu olarak belirtilen değerler tam verilere en yakın değerler olduğundan seçilmesi uygun olacak yöntemi göstermektedir.

Tablo 2 incelendiğinde değişkenler arası düşük korelasyona sahip 50 ve 100 birimlik veri setlerinde regresyon yönteminin, 200 ve 400 birimlik veri setlerinde ise ortalama ve silme yöntemlerinin tam verilerden elde edilen t ve p değerlerine en yakın sonuçları verdiği gözlenmiştir. Kullanılan yöntemlerin hiç birinde, gerçekte anlamlı farklılık olmayan ($p>0.05$) durumlara yanlış olarak fark var ($p<0.05$) denilmesi ile oluşan 1. Tip hataya rastlanmamıştır. Ancak 200 birimlik düşük korelasyonlu %10 kayıp içeren veri setinde regresyon yönteminin sonucunda ortaya çıkan p değerinin ($p=0.093$) 1. Tip hataya yaklaştığı görülmektedir

Tablo 3 incelendiğinde ise değişkenler arası yüksek korelasyona sahip veri setlerinde BM yönteminin tam verilerden elde edilen t ve p değerlerine en yakın sonuçları verdiği gözlenmiştir. Kullanılan yöntemlerin hiç birisinde gerçekte anlamlı farklılık olmayan ($p>0.05$) durumlara yanlış olarak fark var ($p<0.05$) denilmesi ile oluşan 1. Tip hataya

rastlanmamıştır. Kullanılan yöntemlerin hiç birinde gerçekte anlamlı farklılık olmayan (p>0.05) durumlara yanlış olarak fark var (p<0.05) denilmesi ile oluşan 1. Tip hataya rastlanmamıştır. Ancak 100 birimlik yüksek korelasyonlu %20 kayıp içeren veri setinde yerine ortalama koyma (p=0.098) ve regresyon (p=0.078) yöntemlerinin 1. Tip hataya yaklaştığı görülmektedir.

Tablo 3.
Yüksek korelasyonlu veri setlerine ait t-testinin t ve p değerleri

			TAM		SİLME		ORTALAMA		REGRESYON		BM	
	N	%	t	p	t	p	t	p	t	p	T	p
Yüksek	50	0	**0,851**	**0,399**								
		5			0,517	0,608	-0,167	0,868	0,410	0,684	**0,720**	**0,475**
		10			0,755	0,455	**0,791**	**0,433**	0,991	0,327	1,601	0,116
		20			0,052	0,959	-0,210	0,835	**0,541**	**0,591**	1,391	0,170
	100	0	**-0,238**	**0,813**								
		5			-0,071	0,943	-0,120	0,905	-0,059	0,954	**-0,272**	**0,786**
		10			-0,342	0,733	-0,303	0,763	**-0,270**	**0,788**	-0,463	0,644
		20			-1,779	0,147	-1,670	0,098	-1,780	0,078	-1,883	0,063
	200	0	**0,780**	**0,436**								
		5			0,372	0,710	**0,768**	**0,443**	1,454	0,148	0,985	0,326
		10			0,668	0,505	-0,121	0,904	0,171	0,864	**0,843**	**0,400**
		20			**0,735**	**0,465**	-1,433	0,154	-0,289	0,773	0,118	0,906
	400	0	**-0,359**	**0,720**								
		5			-0,143	0,887	**-0,320**	**0,749**	-0,697	0,486	-0,261	0,794
		10			0,067	0,947	1,122	0,263	0,439	0,661	**0,336**	**0,737**
		20			0,143	0,887	-0,275	0,783	-0,506	0,613	**-0,397**	**0,692**

ANOVA Testine Ait Bulgular ve Yorumlar: Normal dağılıma sahip, düşük ve yüksek korelasyonlu farklı oranda kayıp veri içeren farklı büyüklükteki veri setlerine uygulanan ANOVA testi sonuçlarına ait F ve p değerleri Tablo 4 ve Tablo 5'te verilmiştir.

Tablo 4.
Düşük korelasyonlu veri setlerine ait ANOVA testinin F ve p değerleri

			TAM		SİLME		ORTALAMA		REGRESYON		BM	
	N	%	F	p	F	p	F	p	F	P	F	p
Düşük	50	0	0,411	0,665								
		5			**0,411**	**0,665**	0,512	0,603	0,508	0,605	0,544	0,584
		10			0,704	0,502	0,659	0,522	**0,544**	**0,584**	0,696	0,504
		20			**0,279**	**0,760**	0,770	0,469	0,247	0,782	0,890	0,418
	10 0	0	4,236	0,017								
		5			5,013	0,009	4,961	0,009	**4,641**	**0,012**	5,036	0,008
		10			**4,121**	**0,021**	0,194	0,824	2,671	0,074	2,533	0,085
		20			1,010	0,374	**2,069**	**0,132**	1,725	0,184	1,665	0,195
	20 0	0	0,305	0,738								
		5			0,399	0,672	0,400	0,671	**0,351**	**0,704**	0,388	0,679
		10			**0,185**	**0,831**	0,123	0,885	0,139	0,871	0,114	0,893
		20			**0,138**	**0,871**	0,127	0,881	0,129	0,879	0,900	0,914
	40 0	0	2,061	0,129								
		5			1,981	0,139	1,984	0,139	2,518	0,082	**2,026**	**0,133**
		10			**1,710**	**0,183**	1,442	0,238	1,023	0,361	1,483	0,228
		20			**2,055**	**0,132**	1,092	0,336	2,972	0,052	1,157	0,316

Düşük korelasyona sahip veri setlerinden elde edilen analiz sonuçları incelendiğinde ağırlıklı olarak silme yönteminin en yakın sonuçları verdiği gözlenmiştir.

Tablo 5.
Düşük korelasyonlu veri setlerine ait ANOVA testinin F ve p değerleri

					Silme		Ortalama		Regresyon		BM	
	N	%	F	p	F	p	F	p	F	P	F	p
Yüksek	50	0	**0,662**	**0,521**								
		5			0,758	0,474	0,778	0,465	0,572	0,568	**0,642**	**0,531**
		10			1,032	0,367	0,504	0,608	0,818	0,447	**0,693**	**0,505**
		20			0,017	0,983	0,005	0,995	0,999	0,376	**0,715**	**0,494**
	100	0	**0,495**	**0,611**								
		5			0,357	0,700	0,361	0,698	0,499	0,609	**0,492**	**0,613**
		10			0,132	0,876	0,194	0,824	0,613	0,544	**0,512**	**0,601**
		20			0,245	0,784	0,734	0,483	**0,501**	**0,607**	0,623	0,538
	200	0	**0,141**	**0,868**								
		5			0,398	0,672	0,399	0,672	**0,172**	**0,842**	0,189	0,828
		10			0,340	0,712	0,328	0,706	0,119	0,888	**0,148**	**0,863**
		20			1,687	0,192	0,558	0,573	0,213	0,808	**0,193**	**0,825**
	400	0	**0,746**	**0,509**								
		5			0,608	0,545	0,609	0,544	**0,724**	**0,485**	0,761	0,468
		10			0,596	0,552	0,779	0,459	0,615	0,541	**0,609**	**0,544**
		20			0,126	0,881	0,783	0,458	0,932	0,394	**0,650**	**0,522**

Tablo 5 incelendiğinde örneğin yüksek korelasyona sahip veri setlerinden elde edilen analiz sonuçlarına göre ağırlıklı olarak BM yönteminin en yakın sonuçları verdiği gözlenmektedir. Bu tür verilerde silme ve yerine ortalama koyma yöntemlerinin ANOVA analizinde gerçekçi sonuçlar vermediği görülmüştür. I. ve II. Tip hataya herhangi bir yöntemin yol açmadığı da bu veri grubunda ifade edilebilir.

5. TARTIŞMA VE SONUÇ

Bu çalışmada kayıp veri ele alma yöntemlerinin farklı veri setlerinde t-testine ve ANOVA analizi parametreleri üzerine etkisi incelenmiştir. Türetilen tüm veri setlerinin silme yöntemiyle veya yerine ortalama koyma, BM ve regresyon gibi atama yöntemleri sonucu elde edilen yeni veri setlerinin analizleri sonucunda, genel olarak BM ve regresyon atama yöntemlerinin ön plana çıktığı görülmüştür. Ayrıca düşük ve yüksek korelasyonlu veri setlerinde t-testinde silme yönteminin etkin bir yöntem olmadığı da görülmüştür. Düşük korelasyonlu veri setlerinde ortalama ve regresyon atama yöntemleri, yüksek korelasyonlu veri setlerinde kullanılan yöntemlerin araştırmamızda kullanılan örnek veri setleri için genelleme yapılabilecek bir farklılık oluşturmadığı görülmüştür. Bu durum Baygül (2005)'ün yaptığı çalışma ile benzerlik göstermektedir. Ancak bu durum farklı büyüklükteki farklı kayıp miktarına sahip veri setleri için farklılıkların ortaya çıkabileceğinin de ipuçlarını vermiştir. Farklı büyüklükteki farklı kayıp miktarına sahip veri setlerinde kullanılan yöntemler sonucu ANOVA testinin değerleri incelendiğinde, düşük korelasyonlu veri grubu için silme yönteminin ön plana çıktığı; yüksek korelasyonlu veri setlerinde BM ve regresyon yönteminin tam değerler

oldukça yakın sonuçlar verdiği görülmüştür. Böylece BM'nin etkin yöntem olduğu literatür ile benzerlik göstermektedir. Araştırmacılar silme yöntemi, yerine ortalama koyma yöntemi, regresyon yöntemi, BM yöntemi harici diğer yöntemleri kullanarak en etkili yöntem konusunda farklı çalışmalar yapabilirler.

Daha genel sonuçlar elde edilmesi açısından örneklemdeki veri setlerinin birim sayıları arttırılabilir. Araştırmacılar normal dağılıma sahip olmayan veri setleri üzerinde de çalışabilirler. Araştırmacıların farklı kayıp oranları üzerinde çalışmaları önerilebilir. Bu çalışmada kayıp veri olma durumu TROK yapısı olarak belirlenmiştir. Araştırmacılar ROK yapısına uygun veri setlerindeki kayıp sorununu gidermeye yönelik çalışmalar yapabilirler. Bu çalışmada farklı büyüklükteki farklı kayıp miktarı içeren veri setlerine t testi ve ANOVA analizleri uygulanmıştır. Araştırmacılar kullanılan silme ve atama yöntemlerinin t testi ve ANOVA'dan farklı istatistiksel yöntemler üzerindeki sonuçlarını inceleyebilirler. Araştırmacılar farklı faktör sayılarında, farklı yöntemler kullanarak en etkili yöntemi belirleyebilirler.

KAYNAKLAR

Afiffi A. And Elashoff R. M. (1966). Missing observations in multivariate statistics: ı. review of the literature, *Journal of the American Statistical Association*, *61*(315) 595-604.

Allison P. D. (2001). *Missing data, sage university papers series on quantitativ eapplications in the social sciences*, ThousandsOaks, CA, Sage.

Alpar, R. (2003). *Uygulamalı çok değişkenli istatistiksel yöntemlere giriş-1*, Nobel Kitabevi.

Bal C. (2003). *Çok gruplu veri setlerinde kayıp gözlem sorununun çözümlenmesi ve sağlık alanında bir uygulama,* Doktora Tezi,Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, Eskişehir.

Baygül A, (2007). *Kayıp veri analizinde sıklıkla kullanılan etkin yöntemlerin değerlendirilmesi*, Yüksek Lisans Tezi, İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Cheema, J. (2012). *Handling missing data in educational research using spss. Unpublished doctoral dissertation*. George Mason University, USA.

Çokluk Ö. , Kayrı M., (2011). Kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi, *Kuram ve Uygulamada Eğitim Bilimleri,* Kış; 289-309.

Dempster, A. P.,Laird, N. M, and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm, *JournalRoyal Statistical Soc*. *39*

Enders C. K. (2011). *Analyzing Longitudinal Data With Missing Values*, USA.

EvensandFiona H. (2003). *Detecting fishing underwater video using the EM algorithm*, Proceeding of the IEEE International Conference on Image Processing, Barcelona.

Gildea, L. And Hofmann, T. (1999). *Topic-Based language model susing EM*, (http://www.cs.brown.edu/people/th/papers/GildeaHofmannEUROSPEECH99.pdf , 31. 01. 2013, Erişildi).

Iturria, S.J. and Blangero, J. (2000). An EMalgorithmforobtainingmaximumlikelihoodestimatesınthemulti-phenotypevariancecomponents linkage model, *Ann. Hum. Genet.,64.*

Kalaycı Ş. (2008), *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*.
Karasar N. (2004), *Bilimsel araştırma yöntemleri*. Ankara : Nobel Yayıncılık.
Little R. J. A. (1998). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association 38,* 1198-1202.
Little R. J. A and Rubin D. R.(2002). *Statistical Analysis WithMissing Data*, Second Edition, Wiley, New York.
Oğuzlar A. (2001). *Alan araştırmalarında kayıp değer problemi ve çözüm önerileri*, V. Ulusal Ekonometri ve İstatistik Sempozyumu, Çukurova Üniversitesi İİBF Ekonometri Bölümü, Adana, 19-22 Eylül.
Rubin, D. R. (1976). Inference and missing data. *Biometrika*, *63* (3), 581-592.
Satıcı, E. ve Kadılar, C. (2009), Kayıp gözlem olması durumunda kitle ortalamasının tahmini, *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, *10*(2), 549-556.
Sezgin E. ve Çelik Y.(2013). *Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması*, Akademik Bilişim Konferansı, Akdeniz Üniversitesi, 23-25 Ocak 2013.
Tabachnick, B. G. and Fidel (2001). *L.S. Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn& Bacon.
Yazıcıoğlu, Y. ve Erdoğan, S. (2007). *Spss uygulamalı bilimsel araştırma yöntemleri*. Ankara: Detay Yayıncılık.
Yazıcı, F. (2005). *EM algoritması ve uzantıları*, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü,Ankara.
Yozgatligil, C., Purutcuoglu, V., Yazıcı, C. ve Batmaz, İ. (2011). *Validity of homogeneity tests for meteorological time series data: a simulation study.* "Proceedings of the 58th World Statistics Congress (ISI2011)".

EXTENDED ABSTRACT

Educational and psychological scientist have improved their ability to carry out quantitative analysis on large and complex data bases with the use of computers. Generally, scientists have ignored or have underestimated some kind of research problems by reason of missing data but with the help of improved technology and computers, this problem can be handled easily. The term missing data means that some type of interested information about the phenomena are missing.

The goal of the researcher is running the most precise analysis of the data for making acceptable and effective deductions about the population. Missing data is one of the most common problems in data analysis. The problem occurs as a result of various factors. Equipment errors, reluctant respondents, researcher goofs can be given as an example for these factors. Quantity and pattern of missing data determines it's seriousness for the research.

Researchers should also take into consideration the pattern of missing data as well as amount and source of missing data. Occasionally, pattern of missing data can be called as missing data mechanism. It is essential to remind here that the word "mechanism" is used as a technical term. It gives point to structural association with the missing data

and the observed and/or missing values of other variables in the data without emphasizing the hypothetical primary reason of these assoications. Missing completely at random (MCAR), missing at random (MAR), and not missing at random (MNAR) are three patterns of missingness. The data used in this study is simulated data. MCAR pattern can be produed by randomly discharding cases. But orther missing data patterns cannot be produced with simulated data. For these reasons MCAR data pattern was used as missing data pattern.

In this study various statistical methods aimed at overcoming missing data problem in multivariate data sets which involve missing observation. Our research model is in nature of basic research that defines the theoretical studies aimed at data generation. The primary source of data used for statistical anlysis performed in this study was a simulated data sets. Reason for using simulated data was that it **is** difficult to satisfy all of the assumptions under experimental conditions such as different sample sizes ranging from very small to very large with data missing at different rates in these samples. In this study n=50, 100, 200, 400 unit and total 250 conditional derivative data each of which consisting of three variables are used. Data derivation is obtained with the help of R-Studio package program by considering the particular conditions.

The variables in full data sets derived randomly in N=50, 100, 200, 400 unity are reduced incidentally at %5, %10, %20 rates respectively and new data sets to be used for missing data analyses are obtained. The data sets reduced under random conditions which involves%5, %10, %20 missing data has TROK (MCAR) structure. 24 new data sets are created by reducing %5, %10, %20 observation respectively from 8 data sets having low and high correlation 50, 100, 200, 400 unity. 32 data sets are studied in sum for our new analyses. For the purpose of comparing the performances of the methods applied by use of the created artificial data periods; mean replacement, expectation maximization (EM), multiple assignment techniques among missing data completion methods and list/occasional deletion techniques among deletion methods are chosen. T-test and ANOVA are applied to data that do not involve data and data sets involving missing data according to scenarios of missing at various rates. The actual parameters obtained and parameters obtained as a result of missing data methods are applied comparison analysis. Furthermore similarity and efficiency of the methods used are aimed to be determined from the results of t-test and ANOVA.

Missing data methods used in the study:
List/Occasional Deletion Method: The first method first comes in to mind in data sets involving missing value is ignoring the records that contain missing value. Thus the observations involving missing value shall be removed from data set and not included in to analysis.(Sezgin and Çelik, 2013).

Mean Imputation method: This is easy and fast method in which all missing cases are substituted by the mean of total sample. This method has some disadvantages practically. Usage for respondents at the extremes can couse misleading results. Rich and poor persons would not want to give their incomes in a telephone survey. If mean of the population is substituted for this maissing part, it would be spurious guess. Mean substitution doesn't change variable mean but it can be used only if the missing pattern is MCAR. But it has reducing effect on variance and causes biased and deflated errors.

Expectation Maximization (EM/ BM): EM methods are exist for randomly missing data and produces unbiased parameter estimates. EM assumes an assumption that is normal distribution before handling missing data and estimates missing data values on the likelihood under that distribution. EM is an iterative procedure and includec twp steps-expectation (E) and maximization (M) for each iteration. With E step, the conditional expectation of the parameter is calculated on missing data and with M step the parameters by maximizing the complete data likelihood are estimated.

Regression method: The purpose of regression method is testing the value of the dependent variable with the help of one or more independent variables. In regression assignment method dependent variable is missing observed variable and independent variables are other variables. This method is particularly recommended to be used where number of missing date is at intermediate level and shows a spread distribution. The relation between the independent variable and independent variables must be very high for use of this method. (Alpar, 2003; Kalaycı, 2008).

The differences between the methods used in the study result differentiated in the data sets with various sizes having different correlation and in different sizes. While regression and EM methods are useful in data sets with low (50 unit, 100 unit) unitsregression and mean replacement method has given more consistent results with full data in data sets with high (200 units, 400 units) units. Furthermore the deletion method is seen to be a efficient method in low correlated data sets.