T.C.
MERSİN ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ve TIBBİ BİLİŞİM ANABİLİM DALI

MİKRODİZİ VERİLERİNDE FARKLI YAPI VE SAYIDA ORTAYA ÇIKAN EKSİK VERİLERİN 1.TİP HATA ÜZERİNE ETKİSİNİN ARAŞTIRILMASI

Asena Ayça ÖZDEMİR

YÜKSEK LİSANS TEZİ

DANIŞMAN
Prof. Dr. Emine Arzu KANIK

MERSİN–2015

T.C.
MERSİN ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ve TIBBİ BİLİŞİM ANABİLİM DALI

MİKRODİZİ VERİLERİNDE FARKLI YAPI VE SAYIDA ORTAYA ÇIKAN EKSİK VERİLERİN 1.TİP HATA ÜZERİNE ETKİSİNİN ARAŞTIRILMASI

Asena Ayça ÖZDEMİR

YÜKSEK LİSANS TEZİ

DANIŞMAN
Prof. Dr. Emine Arzu KANIK

Tez No: 292

MERSİN – 2015

Mersin Üniversitesi Sağlık Bilimleri Enstitüsü

Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı Yüksek Lisans Programı çerçevesinde Prof. Dr. E. Arzu KANIK danışmanlığında Asena Ayça ÖZDEMİR tarafından hazırlanmış olan "Mikrodizi Verilerinde Farklı Yapı ve Sayıda Ortaya Çıkan Eksik Verilerin 1. Tip Hata Üzerine Etkisinin Araştırılması" başlıklı çalışma, jürimiz tarafından Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 05/01/2016

Prof. Dr. E. Arzu KANIK
Mersin Üniversitesi
Jüri Başkanı

Prof. Dr. Bahar TAŞDELEN
Mersin Üniversitesi
Jüri Üyesi

Yrd. Doç. Dr. İlker ÜNAL
Çukurova Üniversitesi
Jüri Üyesi

Bu tez, Enstitü Yönetim Kurulunun 05/01/2016 tarih ve 2016/11 sayılı kararı ile onaylanmıştır.

Doç. Dr. Fatma ATEŞ
Enstitü Müdürü

T.C. MERSİN ÜNİVERSİTESİ SAĞLIK BİLİMLERİ ENSTİTÜSÜ 1933

TEŞEKKÜR

Yüksek lisans eğitimim boyunca ve tez hazırlama sürecinde yardımlarını esirgemeyen başta Anabilim Dalı Başkanı ve danışman hocam, Sayın Prof. Dr. E.Arzu KANIK'a ve Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı ailesine her türlü katkıları, gösterdikleri sabır ve anlayış için teşekkürü bir borç bilirim.

Hayatımın her aşamasında olduğu gibi beni bu zorlu süreçte de yalnız bırakmayan, güç veren, eğitim hayatım boyunca maddi ve manevi desteğini esirgemeyen ve her daim yanımda olacağını bildiğim canım anneme teşekkür ederim.

Karşılaştığım her zorlukta yanımda olan ve desteğini her zaman hissettiren Ümit KARAKAŞ'a teşekkür ederim.

Mersin Üniversitesi Mühendislik Fakültesi Elektrik Elektronik Mühendisliği Araştırma Görevlisi Mustafa Berkan BİÇER'e tezime sağladığı katkıları ve yardımları için teşekkür ederim.

Asena Ayça ÖZDEMİR

Mersin, 2016

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELERDİZİNİ

Sayfa

SİMGELER ve KISALTMALAR DİZİNİ

cDNA :Komplementer Deoksiribonükleik Asit

df :Serbestlik derecesi

DNA :Deoksiribonükleik Asit

FDR :Yanlış Bulgu Oranı

GA :Güven Aralığı

GNO :Genel Not Ortalaması

HGP :İnsan Genom Projesi(Human Genome Project)

HUGO :İnsan Genom Organizasyonu

MAR :Rasgele Olarak Kayıp

MCAR :Tamamen Rasgele Olarak Kayıp

MN :Matematik Notu

MNAR :Rasgele Olmayan Kayıp

mRNA :Mesajcı ribonükleik asit

NHGRI :Birleşik Devletler Ulusal İnsan Genom Araştırmaları Enstitüsü

R :Y ile aynı boyutta ikili matris

ri :R'nin her bir değeri için gösterimi

SAM :Mikrodizi Anlamlılık Testi

SD :Standart sapma

SE :Standart hata

vi :Ortak vektör

Y :Gözlenen veriler

Ymis :Gözlemlenemeyen veriler

Yobs :Gözlemlenen veriler

B :Populasyon parametresi

φ :Eksik veri mekanizmasının altında yatan parametre

μ :Ortalama

ÖZET

Mikrodizi Verilerinde Farklı Yapı ve Sayıda Ortaya Çıkan Eksik Verilerin 1.Tip Hata Üzerine Etkisinin Araştırılması

Eş zamanlı olarak binlerce genin ekspresyon düzeylerini ölçmeye yarayan mikrodizi teknolojisi, son yıllarda önemli araçlardan biri olmuştur. Mikrodizi teknolojisinde arka plan pikselleri ve spot pikselleri arasındaki eşitsizlik, hibridizasyon sırasındaki hatalar, floresan yoğunluğu, slayt yetmezliği, slaytlarda bulunan toz ya da çizikler, görüntünün bozulması, çözünürlüğün yeterli olmaması, deneysel hatalar gibi pek çok sebep veri yapısında eksik/kayıp verilerin oluşmasına sebep olabilmektedir. Bu sebeplere göre eksik veri yapısı bazı setlerde %10 olabiliyorken bazı veri setlerinde bu oran %90'a kadar çıkabilmektedir.

Bu çalışmamızda Princeton Üniversitesi Onkoloji Bölümü Mikrodizi veri tabanında bulunan araştırmacının kullanımına açık olarak yer alan, Adenoma ve Carcinoma mikrodizi setlerinde, farklı oranlarda meydana gelen eksik gözlem durumunda iki bağımsız grup ortalamalarının karşılaştırılması sonucu oluşan 1. Tip hatalar yorumlanmıştır. Çalışmada Matlab R2015a programından yararlanarak %1, %5, %10, %15, %20, %30, %40, %50, %60, %70, %80 ve %90 oranlarında rasgele eksik veri yapısı oluşturulmuştur ve her oran için 1000 tekrar gerçekleştirilmiştir. Oluşturulan eksik veri setlerinin her biri için Student t testi, Benjamini Hochberg ve Benjamini Yekutieli prosedürleri ve mikrodizi veri setleri için önerilen mikrodizi anlamlılık analizi (significance analysis of microarray) uygulanmıştır.

Oluşturulan eksik veri oranlarına göre; eksiklik yüzdesi arttıkça 1. Tip hata oranlarında bozulma meydana gelmektedir. Gerçek veri setinde kullanılan dört yöntem birbiriyle karşılaştırıldığında beklenen %5 1. Tip hata oranına en yakın sonucu mikrodizi anlamlılık analizi vermiştir ve eksiklik oranlarından en az etkilenen testtir. Beklenen %5 1. Tip hata oranına ikinci olarak en yakın sonucu veren test Benjamini Yekutieli prosedürü olmuştur.Kullanılan bütün testler örneklem genişliğinden etkilenmektedir.

Anahtar kelimeler: Mikrodizi, mikrodizi veri analizi, eksik veri, 1. Tip hata

ABSTRACT

Researching Effect of Missing Values Occured Different Structures and Numbers onto Type I Error in Microarray Datas

Simultaneously, the means for measuring expression levels of thousands of genes microarray technology has been one of the major tools in recent years. There are many reasons such as inequality between background pixels and spot the pixels, error during hybridization, fluorescent intensity, slide failure, dust or scratches found on the slide, poor image, lack of adequate resolution may lead to missing or lost data for data structure in the microarray technology. For reasons like these, missing data structure in some data sets could 10%, this rate can reach up to 90%. In this study, in the Princeton University Department of Oncology Microarray database located as available to researchers, in the Adenoma and Carcinoma microarray sets, in case of missing observations seen at different rates resulting from the comparison of two independent group average Type 1 errors were interpreted. In proportion as 1%, 5%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% random missing data structure is formed and conducted 1000 repeated for each rate.

According to the missing value rate generated, increasing missing data rate cause to deterioration in theType I error rate. When comparising the four methods to each other in the real data, Significance Analysis of Microarray give the nearest results to expected 5% Type I error rate and the lowest effected from missing value rates. Benjamini Yekutieli procedure is the second test that give the nearest results to expected 5% Type I error rate. All of the test used effected from sample size.

Key words: Microarray, microarray analysis, missing value, type 1 error

1. GİRİŞ

1970'lerde San Jose'deki Almaden Araştırma laboratuvarlarında ilişkisel veri tabanının ilk prototibi olan R sistemi yapılmıştır. R sistemi, SQL gibi ilişkisel formlarda, veri yapısının temelini oluşturmuştur. Büyük, orta ve kişisel masaüstü bilgisayarların işlem güçlerinin gelişmesi ile birlikte veri tabanları verileri toplamak ve depolamak için en yaygın yöntemlerden biri haline geldi. Hatta bunların çoğalması verilerin depolanması etrafında bir disiplinin yaratılmasına yol açmıştır. Böylece düzgün bir şekilde birden fazla veri tabanından veri ilişkilendirmek ve yönetmek çok kolaylaşmıştır (1).

30 yıl önce bir kişinin gizliliğini koruması oldukça kolaydı. Kişisel bilgileri saklamak için az sayıda bilgisayar sistemi vardı. İnternet o kadar ilkel durumdaydı ki çoğu kişi varlığının farkında bile değildi ve sadece bir kaç bin kişinin sahip olduğu kişisel cep telefonları büyük bir.O zamandan günümüze mevcut olan her şey hızlı bir şekilde değişti. GPS konumlandırma bilgileri, cep telefonu görüşmeleri, kısa mesajlar, kredi kartı ile alışveriş, elektronik postalar, çevrimiçi sosyal ağlardaki görüşmeler ve hatta elektronik tıbbi kayıtlarımız olmak üzere yaşam akışımız sürekli sayısallaştırıldı ve arşivlendi. Yaşam ve biyomedikal bilimler büyük veri devriminden korunmaktan ziyade ona büyük katkılarda bulundu (2). Büyük veri çağı bugün tam gücüyle devam ediyor, çünkü dünya değişiyor. Bu değişimin içine biyologlar da katıldı. Yüksek verimli genetiğin gelişiyle bilim adamları, büyük veri setleri ile uğraşmaya başladı. Her geçen yıl, genlerin düzenlenmesi, genomların evrimi, insan vücut boşluklarında hangi mikroplar nerede yaşar, farklı kanser etkilerinin genetik oluşumu nedir sorularına kadar her şeyi araştırmak için büyük veri setleri oluşturdular. Sağlık sektörü, zaman içerisinde kayıtlar tutarak büyük miktarda veri üretti (3). Avrupa Moleküler Biyoloji Laboratuvarı'nın bir parçası ve dünyanın en büyük biyoloji-veri tabanlarından biri olan Hinxton'daki Avrupa Biyoinformatik Enstitüsü, şu anda genlerin, proteinlerin ve küçük moleküllerin verileri ve yedekleri olmak üzere 20 petabayt (1 petabayt 1015 byte) bilgi saklamaktadır (4).

Büyük örneklem genişliği ve yüksek boyutluluk birçok modern veri setlerini içerir. Bunlardan genetik alanda en çok göze çarpanı, her biri moleküllerin ifade

değerlerinin on binlercesini içeren ve 500.000'den fazlası kamuya açık bir şekilde yayınlanan mikrodizi veri setleridir(5).

Mikrodizi, binlerce genin hücre ve dokularda eş zamanlı olarak ifade düzeylerinin incelenmesine olanak sağlayan bir yöntemdir ve tek seferde bir organizmanın (bakteriler, mayalar, bitkiler ve insanlar) tüm genomunun ifadelerini inceleyebilir (6-8) . Genellikle yüksek kapasitede veri elde etmek için kullanılırlar ve şu anda 40,000 geni aynı anda analiz edebilecek boyutta mikrodiziler mevcuttur. Cam ya da silikon dizi yüzeylerine genler sabitlenerek ifadeleri ölçülür (8).Ölçülen yapının türüne göre DNA mikrodizi ya da ptotein mikrodizi olarak adlandırılır(6).

Mikrodizi teknolojisinde arka plan pikselleri ve spot pikselleri arasındaki eşitsizlik, hibridizasyon sırasındaki hatalar, floresan yoğunluğu, slayt yetmezliği, slaytlarda bulunan toz ya da çizikler, görüntünün bozulması, çözünürlüğün yeterli olmaması, deneysel hatalar gibi pek çok sebep veri yapısında eksik/kayıp verilerin oluşmasına sebep olmaktadır(9). Eksik verilerin oranı bazı veri setlerinde %10'a kadar oluşabilirken, bazı veri setlerinde bu oran %90'a ulaşabilmektedir(10).Mikrodizilerde eksik verilerin yapısı matematiksel olarak genellikle MCAR (tamamen rastgele kayıp) ya da MAR (rastgele kayıp) şeklinde ortaya çıkmaktadır(11).

Literatüre bakıldığında, mikrodizi veri setlerinde eksik verilere oldukça sık rastlandığı ve bu değerleri tahmin etme yöntemlerinin kullanıldığı çalışmalar oldukça fazladır. Çalışmalarda genellikle %1, %5, %10 ve %20'lik eksik veri yapıları oluşturulduktan sonra farklı yöntemlerle tahminde bulunmuşlardır. Yapılan çalışmalar eksik verilerin 1. tip hatayı ve gücü nasıl etkilediği yönünden oldukça kısıtlıdır.

Binlerce genin birlikte değerlendirildiği çalışmalarda hipotez testi çokluğundan kaynaklı sorunlar ortaya çıkmaktadır. Farklı olmaması beklenen genlerde dahi %5 1. tip hata yapma olasılığından dolayı 10.000 gen içerisinde 500 genin anlamlı bulunma ihtimali vardır. Hipotez testi çokluğundan kaynaklı olan 1. Tip hata oranındaki şişkinliği önlemek için Bonferroni tabanlı ailesel hata oranı (Family wise error rate, FWER) düzeltme yöntemlerinin kullanılması önerilir. Ancak bu yöntemlerin kullanılabilmesi için bazı varsayımların sağlanması gerekmektedir. Bu durumda FWER yerine yanlış bulgu oranı (False discovery rate, FDR) olarak adlandırılan kontrol prosedürlerinin kullanılması önerilmektedir(12).

Bu çalışmanın amacı, mikrodizi veri setlerinde farklı yapı ve sayıda oluşan eksik gözlemlerin çalışmanın sonucu ve elde edilen 1. Tip hata değerlerini nasıl değiştirdiğini ortaya koymaktır.

2. GENEL BİLGİLER

2.1.Eksik Veri Mekanizmaları

Yapılan pek çok araştırmada eksik veri durumu söz konusu olabilmektedir. Verilerin neden eksik olabileceğiyle ilgili pek çok sebep olabilir. Bu nedenler matematiksel olarak eksik veri mekanizmalarıyla açıklanabilir (13).

Eksik verilerin varlığı, yapısı ve sayısı araştırmaların sonucunu etkileyebilmektedir. Elde edilen bu sonuçların doğruluğu ve güvenilirliği tartışmaya açık bir konudur (14).

Veri seti içerisindeki eksik verinin oranı %1'den daha az ise önemsiz sayılabilir, %1-5 arasında ise idare edilebilir ve kontrol altına alınabilir. %5-15 arasında ise başa çıkmak için gelişmiş yöntemlerin kullanılması önerilir. Eğer %15'in üzerinde ise sonuçları çok ciddi şekilde etkileyebilir(15).

Eksik verilerin varlığı pek çok problemi de beraberinde getirebilir. İlk olarak, istatistiksel işlemler otomatik olarak eksik verileri eleyebilir. Bunun sonunda, analiz yapmak için yeterli veriye ulaşılamayabilir. İkinci olarak analiz devam edebilir, fakat girilen veri sayısı az miktarda olduğu için sonuçlar istatistiksel olarak anlamlı çıkmayabilir. Üçüncü olarak, analiz edilecek olgular, tüm olgularda rasgele değilse yanıltıcı sonuçlar verebilir. Eksik veri sayısı çok az dahi olsa analiz sonuçlarının yanlış çıkmasına ya da yanlılığa yol açabilmektedir (15).

Eksik verileri düzeltmek, tahmin etmek ve başa çıkmak için literatürde sınıflandırma yöntemleri, regresyon analizleri, yapay sinir ağları, kümeleme yöntemleri gibi pek çok metot veri atamak ya da tahmin etmek için geliştirilmiştir(15).

Little ve Rubin'e göre 3 farklı eksik veri yapısı vardır. Bunlar aşağıdaki yapılardan oluşmaktadır (16).

1) Tamamen rasgele olarak kayıp (MissingCompletely at Random, MCAR)
2) Rasgele olarak kayıp (Missing at Random, MAR)
3) Rasgele olmayan kayıp (Missing Not at Random, MNAR)

2.1.1.Tamamen Rasgele Olarak Kayıp

Tamamen rasgele eksik veri mekanizmasında verilerin kayıp olma durumu gözlenen ve ya diğer eksik verilerden bağımsız olarak gerçekleşmiştir(17).

MCAR varsayımının karşılanması durumunda kayıp veri mekanizması ihmal edilebilirdir (16).

Bir anket çalışmasında kişilere giden ankette tesadüf olarak bazı soruların eksik olması ya da cevaplayan kişilerin soruyu görmemesi MCAR modeline örnek olabilir.

2.1.2.Rasgele Olarak Kayıp

Rasgele olarak kayıp veri yapısı MCAR'a göre daha zayıf bir varsayımdır. Bir değişkendeki kayıp verilerin olasılığının, diğer bütün değişkenler kontrol altına alınması koşuluyla, bu değişkenin kendi değeri ile ilişkisiz olduğunu ifade etmektedir (16).

Bir anket çalışmasında bireylerin soruların ya da cevapların içeriğinden bağımsız olarak, başka herhangi bir sebeple(acelesi olduğundan, okumaktan sıkıldığından vb.) soruları atlaması MAR yapısına örnek olabilir.

2.1.3.Rasgele Olmayan Kayıp

Diğer bir deyişle göz ardı edilemez kayıp olarak adlandırılan rasgele olmayan kayıp yönteminde bir gözlemin eksikliği doğrudan toplanan veya talep edilen veriler ile ilgilidir. Veri kümesi içinde diğer değişkenlerden tahmin edilemez (18).

Anketteki bir sorunun yanlış sorulmasından dolayı doğru cevabın çözülemediği durumlardaki kayıp durumu MNAR varsayımına örnektir(14).

Çizelge 2.1: Eksik veri modelleri

Eksik veri genel modeli	$P(y_i, r_i \mid \boldsymbol{v_i}, \beta, \varphi) = P(y_i \mid \boldsymbol{v_i}, \beta).P(r_i \mid y_i, \boldsymbol{v_i}, \varphi)$
Tamamen rasgele olarak kayıp modeli	$P(r_i \mid y_i, \boldsymbol{v_i}, \varphi) = P(r_i \mid \boldsymbol{v_i}, \varphi)$
Rasgele olarak kayıp modeli	$P(r_i \mid y_i, \boldsymbol{v_i}, \varphi) = P(r_i \mid (y_i)_{\boldsymbol{obs}}, \boldsymbol{v_i}, \varphi)$

Bu üç eksik veri yönteminden yola çıkarak, Y gözlenen tüm verileri (tüm y_i'ler) temsil etmek üzere; Y_{obs} gözlemlenen veriler ve Y_{mis} gözlemlenemeyen (eksik olan) veriler olsun. Tüm veri setindeki değerler $Y=(Y_{obs}, Y_{mis})$ olacaktır.

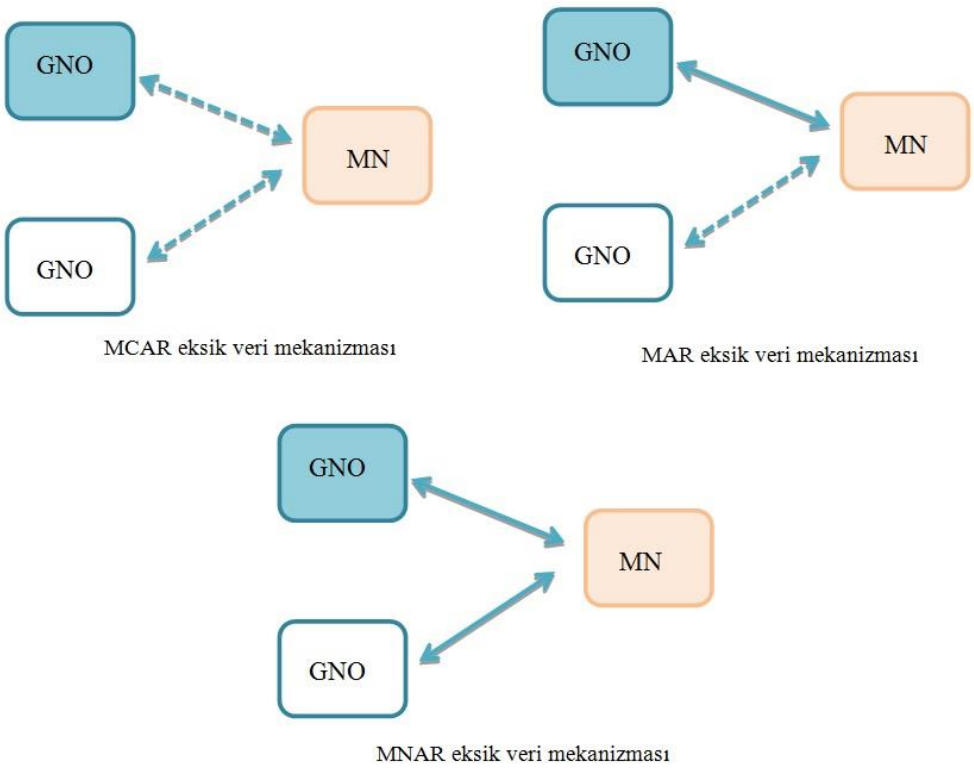
R, Y ile aynı boyutta ikili matris olsun ve R '0' değerini aldığında buna karşılık gelen değer Y_{obs}, R '1' değerini aldığında ise Y_{mis} olacaktır. Her bir değer i ile gösterilirse R'nin her bir değer için gösterimi r_i olsun. r_i ortak bir olasılık dağılımına sahip rasgele değişkenler kümesi olarak y_i gözlem verilerini gözlemleme olasılığını vermektedir. β popülasyon parametresi ve v_i ortak vektör olmak üzere genel model Çizelge 2.1'de verilmiştir. Burada φ eksik veri mekanizmasının altında yatan parametre olup, $P(r_i|y_i, v_i,\varphi)$ bütün gözlem verilerini (gözlenen ve eksik verileri) veren r_i'nin olasılık dağılımını, bağımsız değişkenlerin ortak vektörünü, farklı tasarım değişkenleri, kovaryantı ve eksik veri mekanizmasının altında yatan parametreyi vermektedir (13).
Eksik gözlem verileri MCAR olduğunda r_i olasılık dağılımı y_i değerlerinin gözlenen veya eksik değerler olmasına bağlı değildir. Fakat bağımsız değişkenlerle, farklı tasarım değişkenleriyle ya da kovaryantla bağlantılı olabilir. Buna göre MCAR modeli tüm y_i'ler Çizelge2.1'deverilmiştir. Burada Y_{obs}'deki sadece gözlenen verileri kullanarak, β popülasyon parametreleri hakkında geçerli çıkarım elde edildiği anlamına gelir. Gözlem verilerinin eksik olmasına sebep olan altta yatan mekanizma popülasyon parametrelerinin tahmininde göz ardı edilebilir(13).

Gözlem verileri MAR yapısına sahip olduğu durumda r_i olasılık dağılımı eksik veriye değil, yalnızca gözlenen gözlem değerlerine bağlıdır. Bu durumda MAR modeli tüm $(y_i)_{mis}$'ler için Şekil 2.1'deverilmiştir.Buna göre β parametresi hakkındaki geçerli çıkarım; analiz metodu gözlemler ve eksik veriler arasındaki korelasyon için izin verdiği sürece, r_i ve olasılık dağılımı için bir model olmadan da elde edilebilir. Yani uygun bir kestirim yöntemi (örneğin maksimum olabilirlik kestirimi) kullanılması durumunda eksik verilere neden olan mekanizma model parametrelerinin tahmininde göz ardı edilebilir (13).

MNAR yapısı için, r_i olasılık dağılımı tüm veri setindeki y_i ($(yi)_{obs}$ ve $(yi)_{miss}$'lerin hepsi) değerlerine bağlıdır ve basitleştirilemez. β parametresi hakkında sadece r_i ve y_i'nin ortak dağılımının kullanıldığı eksik veri genel modelinden yararlanarak geçerli çıkarım yapılabilir. Yani model parametrelerinin yansız

tahminlerine ulaşmak için popülasyon modeli eksik veriler için ekstra bir model içermelidir (13).

Üç eksik veri yapısına bir örnek olarak, MN bir öğrencinin eksik göstergesi olarak tanımlanan matematik notu olsun. GNO ise genel not ortalamasını temsil etsin. MCAR yapıda eksik gözlem olduğunda aralarındaki ilişki modeldeki gibi olacaktır (GNO mavi gözlenen, GNO beyaz gözlenemeyen). Buna göre öğrencinin matematik notu gözlenen ya da gözlenemeyen GNO'suna bağlı değildir. MAR yapıda eksik gözlem olduğunda öğrencinin matematik notu sadece gözlenen GNO'suna bağlı olacaktır. MNAR yapıda eksik gözlem olduğunda ise öğrencinin matematik notu hem gözlenen hem de gözlenemeyen genel not ortalamasına bağlı olacaktır (19).



Şekil 2.1: Eksik veri mekanizmaları

2.2.Gen, Genom Ve Gen Ekspresyonu

Kalıtımın işlevsel bir birimi olan gen DNA ve RNA'nın kimyasal yapı taşları olan nükleotitlerin doğrusal dizisidir. Daha kavramsal olarak yaklaşırsak gen, kopyalanabilen, ifade edilebilen, mutasyona uğrayabilen ve bilgiyi depolayan bir birimdir (20). Organizmadaki genlerin tümü organizmanın genomunu oluşturur. Her hücre aynı gen dizilimine sahip olmasına rağmen, ifade edilen genler özelleşen hücre tipine göre farklılık gösterebilir. Örneğin bir pankreatik β hücresi insülin üretirken, bir saç kökü hücresi keratin üretir (21). Gen ekpresyonu; fenotipi etkileyen gen ürünlerinin sentezlenmesi sonucu oluşan genetik aktivitenin ifadesidir ve sürekli değişkenlik gösterir (22).

İnsanda 29.000-36.000 arası gen bulunmaktadır ve bir gen ortalama 3.000 nükleotidden ve insan genomu 3.164.700.000 nükleotidden oluşmaktadır. İnsanlarda nükleotid dizilerinin %99'u aynıdır, yani fark geriye kalan %1'lik kısımdan oluşur (23).

2.3.Mikrodizi

Bilgisayar teknolojisinin ve moleküler biyolojinin hızla gelişmesi sonucunda en dikkat çekici yöntemlerden biri olarak mikrodizi teknolojisi ortaya çıkmıştır ve bir organizmanın genlerinin genom ifade seviyelerini izlemek için kullanılabilecek vazgeçilmez araçlardan biri haline gelmiştir(24,25).

Mikrodizi teknolojisi en geniş tanımıyla, hücre ve dokulardaki gen ekspresyonlarının incelenmesinde kullanılan yeni ve güçlü bir teknolojidir (8). Binlerce genin hücre ve dokularda eşzamanlı olarak ifade düzeylerinin incelenmesine olanak sağlayan bir yöntemdir (6). Mikrodizi tek seferde bir organizmanın tüm genomunun ifadelerini inceleyebilir (7).

Şu anda 40,000 geni aynı anda analiz edebilecek boyutta mikrodiziler mevcuttur. Cam ya da silikon bir yüzey üzerine sabitlenen genler, belirli işlemlerden geçtikten sonra spotlarda ifade düzeylerini yansıtırlar. Bu spotlardaki renkler her bir genin kendini ifade etme düzeyini gösterir. Daha sonra bu renkler bilgisayar çözümlemesiyle sayısal değişkenlere dönüştürülerek analiz için uygun hale getirilir (8).

Mikrodizi teknolojisiyle ilgili yapılan çalışmalar arasında iki önemli proje yer almaktadır. Bunlar İnsan Genom Projesi ve HapMap projesidir.

İnsan Genom Projesi (HGP), insan genomunun ifade edilebilen/açık (ökromatik) bölümünün dizilenmesini amaçlayan, uluslar arası işbirlikli bir girişimdir. Dördü Birleşik Devletlerde biri Birleşik Krallıkta olmak üzere beş büyük merkezin liderliğinde, Fransa, Almanya, Japonya ve Çin'den grupların katılımıyla 1990'da başlatılmıştır. Birleşik Devletler Ulusal İnsan Genom Araştırmaları Enstitüsü (NHGRI) merkez büro olarak görev yapar. Uluslararası bir örgüt olan İnsan Genom Organizasyonu (HUGO), insan genom projesinin pek çok yönüyle ilgilenir. HGP; gen ve genom işlevleri (işlevsel genomik), tüm transkripsiyon süreci (transkriptom), tüm insan proteinlerinin analizi (proteom), diğer organizmalarla insan genomunun kıyaslanması (karşılaştırmalı genomik), çok büyük miktardaki veriyi değerlendirebilmek için yeni yöntemlerin geliştirilmesi (biyoinformatik) ve epigenetik işlevler (epigenom) ile ilgilenir (26).

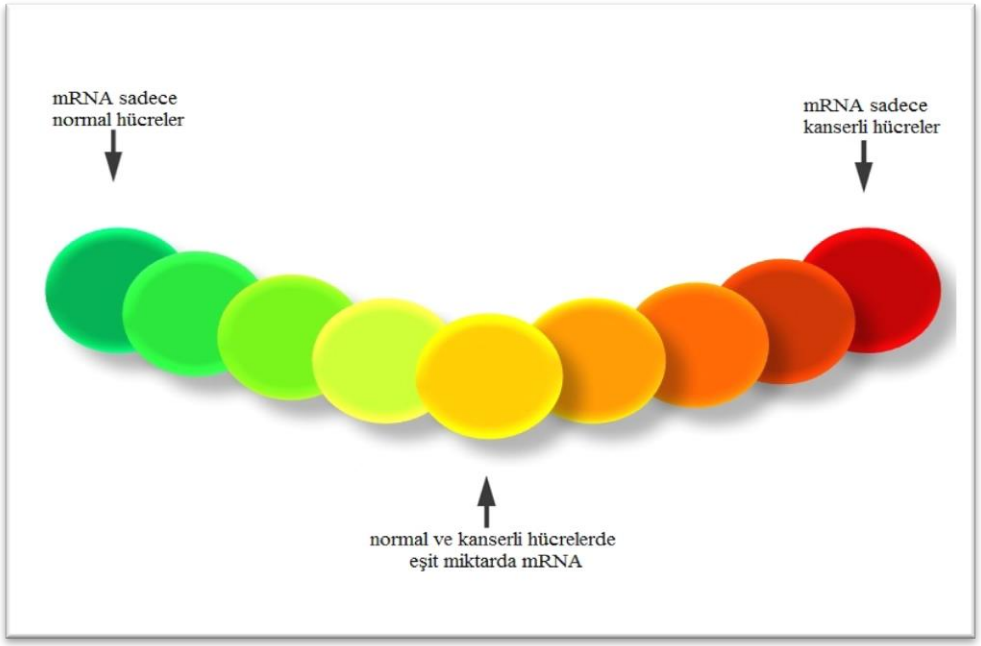
Uluslararası HapMap projesi, insanlarda farklılık ve benzerlik gösteren genetik katalogları tanımlayabilmek için oluşturulmuş pek çok ülkenin katıldığı bir projedir. HapMap projesini kullanarak araştırmacılar genlerin sağlık, hastalık, ilaçlara ve çevresel faktörlere karşı bireysel yanıtlar üzerindeki etkilerini bulabileceklerdir. Bu projenin amacı; genetik çeşitliliğin görüldüğü kromozomal bölgeleri tanımlayabilmek için bireysel farkların genetik sekanslarını karşılaştırmaktır (27). Proje, hastalıkların tanısı, tedavisi ve önlenmesine yönelik yeni metotların önünü açan, diğer araştırmacıların belirli hastalıkların riskini içeren genetik varyantları bağlayabileceği bilgileri sağlamak üzere tasarlanmıştır(28).

Yüzeye sabitlenen yapının türüne göre DNA mikrodizi ve protein mikrodizi olarak ikiye ayrılır(6).

Protein mikrodizi teknolojisinde geliştirilen farklı iki yöntem mevcuttur. Birinci yöntem tek bir protein türüne özgün olarak geliştirilen sistemler (antikor bazlı) , ikinci yöntem ise tüm proteinleri eş zamanlı çalışmaya alarak gerçekleştirilir. İkinci yöntemde bu proteinlere ait bilgiler, uygun programlar aracılığı ile işlenerek elde edilen bilgiler farklı hastalık tablolarında ne gibi değişiklikler gösterdiği incelenir (6).

DNA mikrodizi teknolojisi, genlerin ve gen polimorfizmlerinin araştırılması, gen ifade analizleri, mutasyon analizleri, evrimsel çalışmalar, sekans analizleri, terapötik

ajanların geliştirilmesi, genlerin sınıflandırılması gibi birçok alanda kullanılabilmektedir(6). DNA mikrodizilerde ilk olarak cam, silikon ya da naylondan oluşan katı bir yüzey üzerine nükleit asitler bağlanarak çipler oluşturulur. mRNA'lar cDNA'ya çevirildikten sonra radyoaktif veya floresan belirteçler(Cy3, Cy5) ile boyanır. Bu floresan işaretleyiciler farklı dalga boylarında ışık yayarlar ve uygun lazerle uyarılma sonrasında Cy3 yeşil renk verirken Cy5 kırmızı renk olarak görülür. Hasta ve sağlıklı bireyler olmak üzere iki farklı kaynakta sağlıklı bireyler yeşil ile boyanırken, hasta bireyleri kırmızı ile boyamak mümkündür. Mikrodizi spotlarına ait renklendirme Şekil 2.2'deki gibidir. Boyama işleminden sonra hibridizasyon/melezleme işlemi gerçekleştirilir ve reaksiyondaki yanlış sonuç verebilecek dizinleri ortadan kaldırabilmek ve doğru değerlendirebilmek için yıkama işlemi gerçekleştirilir. Temizlenmiş dizideki hibridizasyonu görünür hale getirebilmek ve değerlendirebilmek için etiketler uyarılır. Mikrodizi üzerindeki radyoaktif sinyaller ya da floresen sinyalleri toplayan kimyasal işlemcilerle bağlantılı olan bileşenleri bir tarama cihazı ile okunur. Tarayıcı ile bağlantılı bir yazılım programı spotların yoğunluğunun belirleyerek anlamlı veriler haline getirilir. Örnek içerisindeki Cy3 ile işaretli olan dizinler, hedef probdaki cDNA’ya hibridize oldu ise o prob yeşil renk, Cy5 ile işaretli olan dizinler hibridize oldu ise o prob kırmızı renk, hem Cy3 hem de Cy5 işaretli dizinler eşit miktarda hibridize olduysa prob sarı renk yayacaktır. Eğer gen ifade edilememişya da okunamamış ise prob siyah renkte gözükecektir (6,8).



Şekil 2.2: Mikrodizi spot renkleri

H1, H2, H3, H4 hasta grubunu oluşturan bireyler, K1, K2, K3, K4 kontrol grubunu oluşturan bireyler, GEN1,..., GENn ekspresyonları ölçülen genler ve Y_{ij} 'ler genlerin ekspresyon düzeylerini belirtmek üzere bir mikrodizi verisi şekildeki gibi satır veri yapısı şeklinde olmaktadır. Mikrodizi verilerinde genel amaç, her bir gen için hasta ve kontrol grubu arasında gen ekspresyonları bakımından fark olup olmadığını araştırmaktır.

Çizelge 2.2: Mikrodizi veri yapısı

	HASTA GRUBU				KONTROL GRUBU			
	H1	H2	H3	H4	K1	K2	K3	K4
GEN1	Y11	Y12	Y13	Y14	Y15	Y16	Y17	Y18
GEN2	Y21	Y22	Y23	Y24	Y25	Y26	Y27	Y28
GEN3	Y31	Y32	Y33	Y34	Y35	Y36	Y37	Y38
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
GENn	Yn1	Yn2	Yn3	Yn4	Yn5	Yn6	Yn7	Yn8

2.3.1.Mikrodizi Veri Normalizasyonu

Mikrodizi cihazından elde edilen ham verileri biyolojik ya da deneysel varyasyonlara sahip olacağından, verileri hiçbir düzeltme yapmadan yorumlamak ve analiz etmek yanlıştır. Bu yüzden verilere normalizasyon işlemlerinin uygulanması gerekir. Normalizasyon gen ekspresyon analizinin ilk aşaması olup, melezlemedeki hataları ortadan kaldırmak için kullanılan bir yöntemdir. Mikrodizi normalizasyon üç amaçla oluşturulur (29).

1. Farklı örnekler arasındaki biyolojik varyasyonu korumak
2. Deneysel varyasyonu en aza indirmek
3. Farklı deneylerin karşılaştırabilmesine olanak sağlamak

Normalizasyon aşaması için pek çok yöntem kullanılmaktadır. Bunlardan birisi $2^{-\Delta\Delta CT}$normalizasyon yöntemidir. $2^{-\Delta\Delta CT}$normalizasyon yönteminde A hasta grubu, A' hasta referans grubu, B kontrol grubu B' kontrol referans grubu olmak üzere, cihazdan elde edilen CT değerleri için ΔCT_H = CT(A)-CT(A') ve ΔCT_K = CT(B)-CT(B') değerleri hesaplanır. Buradan yola çıkarak, $\Delta\Delta CT = \Delta CT_H$-Ortalama($\Delta CT_K$) değerleri hesaplanır. Son olarak $2^{-\Delta\Delta CT}$ değerleri hesaplanarak normalizasyon işlemi tamamlanır (30).

Bir diğer kullanılan normalizasyon yöntemi ise $\log_2$ dönüşümüdür. Her bir gen için, R kırmızı spot yoğunluğu, G yeşil spot yoğunluğu olmak üzere normalizasyon fonksiyonu denklem 2.1'deki gibidir.

Denklem 2.1:$\log_2$normalizasyonu

$$M = log_2 R - log_2 G$$

Burada M=0 ise ekspresyon seviyeleri eşit, M=1 ise ekspresyon seviyeleri arasında iki kat değişim, M=2 ise ekspresyon örnekleri arasında 4 kat değişim olduğunu ifade eder (31).

2.3.2.Mikrodizi Verilerinde Eksik Yapı Mekanizması

Atama yöntemlerinin çoğu ifade veri matrisinde tamamen rasgele kayıp varsayımı altında geliştirilmiş ve onaylanmıştır. Ancak bu varsayım her zaman pratikte mümkün değildir. Farklı ölçümler (diziler) sıklıkla değişken deneysel koşullar altında yürütülmektedir, dolayısıyla diğer faktörlerin yanı sıra melezleme, ortam ya da zaman kavramları da farklılıklara yol açabilir.

Olguların bazı değişkenleri hipotezden bağımsız, teknik sorunlar nedeniyle ölçülemez. Bu tür yapılarda veri MCAR eksik veri yapısına sahiptir. Ölçümler bazı özel durumlar (eksik veri, diğer gözlenen hasta özelliklerine koşullu olarak bağlı olabilir) için güvenilir ya da ulaşılabilir olmadığı durumda ise MAR eksik veri yapısına sahiptir (11).

2.3.3.Mikrodizilerde Eksik Veri Sebepleri

Mikrodizi deneylerden elde edilen verilerde, büyük matrisler şeklinde ifade edilmekte ve sık sık bazı değerler eksik olarak karşımıza çıkmaktadır(32).

Gen ekspresyonu, veri kümelerinde yetersiz çözünürlük, görüntü bozulması, slayt üzerindeki toz ve ya çizikler ya da laboratuvar işlemi sırasında deneysel hatalar gibi pek çok sebeple eksik veriler görülür(33).

Arka plandaki pikseller ile spottaki piksellerin karşılaştırılması sonrasında, eğer spot pikselinin fraksiyonu arka plandaki pikselin medyanından daha büyük ve belirli bir eşik değerinden daha az ise bu spotta tanımlanan gen ifadesi eksik olarak tanımlanacaktır. İfadelerdeki eksik değerler için bir başka neden melezleme sırasında teknik hatalardır(9).

Mikrodizi verileri %10'a kadar eksik veri içerebilirken, bu oran %90'a kadar ulaşabilmektedir(10).Mikrodizilerdeki eksik yapılar için pek çok analiz yöntemi geliştirilmiştir. Eksik yapıları atama yöntemleriyle tamamlayıp tam bir veri seti ile dizinin analizini yapmak mümkündür. Alternatif olarak genler ya da dizinler eksik yapı olmayana kadar kaldırılabilir (34).

Literatürde "Mikrodizi Eksik Veri" ile ilgili çalışmalar Türkçe ve İngilizce kaynaklarda incelenmiştir. Karşımıza yaygın bir şekilde "mikrodizi eksik veri ataması" konusundan oluşan çalışmalar çıkmıştır.

Ally Rogers ve arkadaşlarının yaptığı çalışmada GAW18 verisinde sıralama ve atama arasındaki uyumu değerlendirmişleridir. Çalışmada 240,456 tek nükleotit polimorfizmi ve 959 birey, yani 230,597,304 olası genotip bulunmaktadır ve 5,000,000'den fazla genotipin eksik olduğu söylenmektedir. Eksik verilerin 1. tip hatayı ve gücü olumsuz yönde etkilediği tartışmada belirtilmiştir (35).

Q. Shang ve arkadaşları, mikrodizide bulunan eksik veriler sebebiyle kümeleme analizi algoritmalarının yetersiz kalacağını düşünmektedir. Eksik gözlemlerin bulundukları yerlere atama yapılmasını önermektedir(36).

Chia-ChunChiu ve arkadaşlarının 2013 yılında yaptıkları çalışmada 13 farklı gerçek mikrodizi veri setinden %1, %5, %10, %15 ve %20'lik rasgele eksik veri yapısı oluşturulmuş ve k-en yakın komşuluk, yinelemeli k-en yakın komşuluk, ardışık k-en yakın komşuluk, en küçük kareler uyarlamalı, yerel en küçük kareler, yinelemeli en küçük kareler, sıralı en küçük kareler, tekil değer ayrıştırma ve Bayes temel bileşenler olmak üzere dokuz farklı atama yöntemiyle eksik gözelere atamalar yapılmıştır. Her bir algoritma 110 defa tekrarlanmıştır ve sonuç olarak yerel en küçük kareler yönteminin en uygun atama yöntemi olduğu sonucuna ulaşmışlardır(37).

O. Troyanskaya ve arkadaşlarının 2001 yılında yapmış oldukları çalışmada Saccharomycescerevisiae isimli DNA mikrodizi verisinde, tekil değer ayrıştırma, k-en yakın komşuluk ve satır ortalaması ataması olan üç farklı atama yöntemini değerlendirmişlerdir. %1-20 arasında eksik veri üzerinde bu üç metodu uygulamışlardır. K-en yakın komşuluk atamasının diğerlerinde göre daha güçlü ve hassas bir yöntem olduğu sonucuna ulaşmışlardır(32).

Lı´gia P. Bra´s ve Jose´ C. Menezes'in 2007 yılında yapmış oldukları çalışmada Saccharomycescerevisiae veri setinde hücre döngüsünden düzenlenen bir çalışma dan TS1 ve TS2 iki veri seti, cDNA zaman serisi mikrodizi verisi ve insan kanser hücrelerine ait olan NTZ verisi olmak üzere dört farklı veri seti kullanılmıştır. Çalışmada %1, %5, %10 ve eşit olmayan eksik veri yapıları oluşturularak K-en yakın komşuluk, ardışık K-en yakın komşuluk ve tekrarlamalı K-en yakın komşuluk atama yöntemleri karşılaştırılmıştır. Tekrarlamalı K-en yakın komşuluk yöntemi yüksek oranlarda eksik varlığında küme tabanlı yöntemlerin tahmin yeteneğini artırabilir ve farklı şekilde eksprese genin saptanmasında daha az zararlı bir etkiye sahip olduğu sonucuna ulaşılmıştır(38).

Erdal Çoşgun ve Ergün Karaağaoğlu'nun ele aldıkları derlemede, mikrodizi verilerinde randomforest sınıflandırma yönteminin eksik veri analizinde çok etkili bir yöntem olduğu ve doğru sınıflama oranının eksik veriler olsa da devam edeceğine değinmişlerdir(39).

Alexandre G de Brevernve arkadaşlarının yapmış oldukları çalışmada veri setinde eksik gözlemlerin olmasının, klasik kümeleme yöntemleriyle elde edilen gen kümelerini bozabileceğine değinmişlerdir. Bununla birlikte K-en yakın komşuluk yaklaşımının eksik ifade edilen genlerin değerlerini tahmin etmek için düşük hata düzeyli bir yöntem olduğuna değinmişlerdir (40).

Muhammad Shoaib B. Sehga ve arkadaşlarının yapmış oldukları çalışmada kollateral eksik veri tahmin etme yöntemi, en küçük kare tahmin yöntemi, temel bileşenler analizi tahmin yöntemi ve K-en yakın komşuluk yöntemini karşılaştırmışlardır. Üç ayrı zaman serisi olmayan yumurtalık kanseri temelli mikrodizi veri setleri ve maya sporlama temeline dayanan zaman serisi verisi kullanmışlardır. Her yöntemin kantitatif 0.01- 0.2 arasında rastgele eksik değer olasılıklarını içeren normalleştirilmiş karekök ortalama hata ölçüsü kullanılarak analiz edilmiştir. Sonuçlar her iki tip veri seti için, kollateral eksik veri tahmin etme yönteminin diğer yöntemlere göre daha üstün ve eksik değerleri tahmin etmede daha güvenilir sonuçlar verdiği yönünde bulunmuştur(41).

3.GEREÇ VE YÖNTEM

Çalışmada, Princeton Üniversitesi Gen Ekspresyon Projesi'nde yer alan açık erişimli mikrodizi veri tabanında yer alan ADENOMA ve CARCINOMA veri setleri kullanılmıştır. Bu setler Notterman ve arkadaşlarının 2001 yılında CancerResearch dergisinde yayınlanan "Transcriptional Gene ExpressionProfiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays" isimli çalışmada kullanılmıştır. Veri setine http://genomics-pubs.princeton.edu/oncology/ adresinden ulaşılabilir. Bu iki veri setinin alınmasının sebebi, verilere uygun formatlarda (excel, txt) kolaylıkla ulaşılabilmesi, normalleştirilmiş verilerin bulunması, gen sayıları yaklaşık olarak aynı, birey sayıları birbirinden farklı olması ve iki grup karşılaştırması olarak Student t testinin kullanılmasıdır.

Carcinoma veri setinde 18 hasta ve 18 sağlıklı birey olmak üzere toplamda 36 birey ve bu bireylere ait 7457 gen bulunmaktadır. Adenoma veri seti, 4 hasta ve bunlara karşılık 4 sağlıklı normal bireye ait 7086 genden oluşmaktadır. Hasta bireylerle sağlıklı bireyler arasında gen ekspresyon düzeyleri arasındaki farkı bulabilmek için iki bağımsız grup ortalama karşılaştırma testlerinden biri olan Student t testi ve mikrodizi veri setleri için geliştirilmiş Mikrodizi Anlamlılık Analizi (Significance Analysis of Microarray (SAM)) uygulanmıştır. Yanlış bulgu oranlarına göre yapılan değerlendirmelerde ise Benjamini Hochberg ve Benjamini Yekutieli prosedürleri kullanılmıştır.

İlk olarak veri setinde 0.01, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 oranlarında grup ayırımı yapmadan, tüm veri seti içerisinde tamamen rasgele yapıda (MCAR), her bir oran için 1000 tekrar olacak şekilde eksik gözeler oluşturulmuştur. 12 adet eksik gözlem yapısı ve 1000 tekrarlı olacak şekilde hazırlanan bu düzenekte Adenoma ve Carcinoma veri setleri için toplamda 24.000 veri seti ile çalışılmıştır. Her iki veri seti için bağımsız iki grup karşılaştırmasında sırasıyla Student t testi ve SAM analizi yapılmıştır. Yanlış bulgu oranlarını hesaplamak için Benjamini Hochberg ve Benjamini Yekutieli prosedürleri uygulanmıştır. Student t testi sonucunda elde edilen p değerleri $p \leq 0.05$ ve $p > 0.05$ olarak değerlendirilmiştir. Elde edilen p değerlerinin ortalamaları alınmıştır. p değerleri $p \leq 0.05$ ve $p > 0.05$ olacak şekilde

sayılmış ve bu sayılara karşılık yüzdeler çizelgelerde verilerek tartışılmıştır. SAM analizi için gereken delta değeri Multi Experiment Viewer 4.9.0 programında her iki veri seti için ayrı ayrı hesaplanmıştır. Carcinoma Veri seti için delta 2.14, Adenoma veri seti için delta 6.22 olarak bulunmuştur ve SAM analizi sonuçlarında tanımlanabilen genlerin sayısı elde edilerek yorumlanmıştır. Benjamini Hochberg ve Benjamini Yekutieli prosedürlerinde elde edilen FDR değerleri ve anlamlı bulunan ve bulunmayan gen sayılarına ait bulgular verilerek tartışılmıştır. Bütün simülasyon aşamaları ve test sonuçları MATLAB R2015a programında yapılmıştır.

3.1.İki Bağımsız Grup Ortalamasının Karşılaştırılması

İstatistiksel teknikler arasında en yaygın kullanılan yöntemlerden biri iki bağımsız grup ortalamasının karşılaştırılmasıdır.

$\{X_1, X_2, X_3, \dots, X_{n1}\}ve\{Y_1, Y_2, Y_3, \dots, Y_{n2}\}$birbirinden bağımsız iki grup olmak üzere, eğer iki örneklem birbirinden farklı merkezlere sahip popülasyondan çekildiyse bu iki grubun ortalamalarının birbirinden farklı olması beklenir. Buna göre $\{X_1, X_2, X_3, \dots, X_{n1}\}ve\{Y_1, Y_2, Y_3, \dots, Y_{n2}\}$ gruplarının ortalamaları denklem 3.1 ve denklem 3.2'deki gibi hesaplanır (42).

Denklem 3.1 X grubunun ortalaması

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$$

Denklem 3.2 Y grubunun ortalaması

$$\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

3.1.1.Varsayımları

1.Parametrik testlerden biri olduğundan normal dağılım gösteren veri setlerinde kullanılır.
2.Gruplarda varyansların homojen olması gerekmektedir.
3.Grup sayısı iki ve gruplar birbirinden bağımsız olmalıdır.
4.Değişkenler, sürekli yapıda olmalıdır (42).

3.1.2.Hipotezleri

μ_X,$\{X_1, X_2, X_3, \dots, X_{n1}\}$ grubunun ortalaması ve $\mu_Y$ise $\{Y_1, Y_2, Y_3, \dots, Y_{n2}\}$ grubunun ortalaması olmak üzere, bağımsız iki grup arasındaki farkın sıfıra eşit olup olmadığı hipotezine dayanarak yokluk hipotezi denklem 3.3'de verilmiştir.(42).

Denklem 3.3. X ve Y bağımsız gruplara ait yokluk hipotezi

$$H_0: \mu_X - \mu_Y = 0$$

Buna karşılık geliştirilen alternatif hipotez çeşitleri tek yönlü ve çift yönlü olarak değişmekle birlikte denklem 3.4, denklem 3.5ve denklem 3.6'de verilmiştir.

Denklem 3.4.X ve Y bağımsız gruplarına çift yönlü alternatif hipotez

$$H_a: \mu_X - \mu_Y \neq 0$$

Denklem 3.5.X ve Y bağımsız gruplarına tek yönlü alternatif hipotez1

$$H_a: \mu_X - \mu_Y > 0$$

Denklem 3.6.X ve Y bağımsız gruplarına tek yönlü alternatif hipotez2

$$H_a: \mu_X - \mu_Y < 0$$

3.1.3.Formüller

Hipotezleri test etmek için kullanılan yöntemler örnek genişliğine göre değişmektedir(43).

3.1.3.1.Geniş Örneklemler İçin Anlamlılık Testi

Örneklem genişliğinin 30'dan büyük olduğu durumlarda kullanılır. Standart yokluk hipotezimiz $H_0: \mu_X - \mu_Y = \mu_0$ (yaniμ_0=0) olmak üzere standart normal dağılım gösteren test istatistiği denklem 3.7'de verilmiştir.

Denklem 3.7. Z test istatistiği

$$Z_0 = \frac{\bar{x} - \bar{y} - \mu_0}{SE(\bar{x} - \bar{y})} \sim N(0,1)$$

Denklem 3.8. X ve Y'ye ait ortak standart hata

$$SE(\bar{x} - \bar{y}) = \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(x_i - \bar{x})^2}{n_1 - 1} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(y_i - \bar{y})^2}{n_2 - 1}}$$

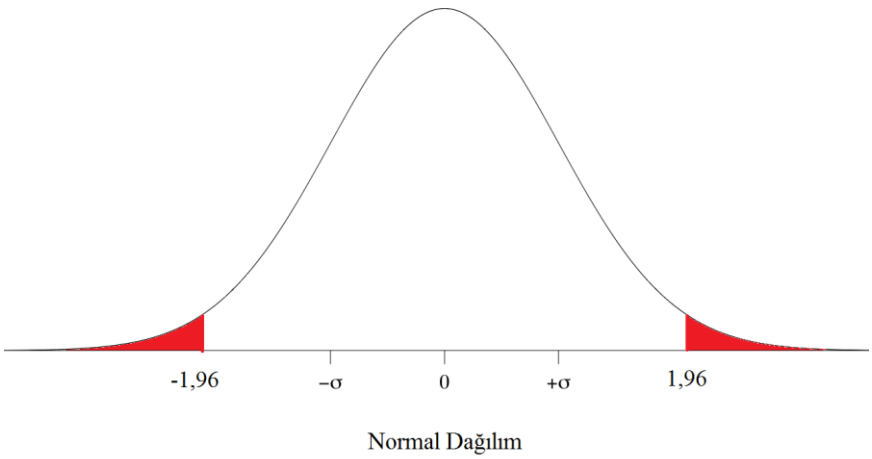
Denklemde standart hatayı açarak yeniden yazacak olursak denklem 3.9'daki gibi olacaktır.

Denklem 3.9. Z test istatistiği (açık)

$$Z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(x_i-\bar{x})^2}{n_1-1} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(y_i-\bar{y})^2}{n_2-1}}}$$

Elde edilen Z_0değeri Ek 1'de gösterilen Z tablosundaki karşılık gelen değer ile karşılaştırılır. Buna göre eğer $|Z_0| \geq Z$ tablo değeri ise H_0 hipotezi reddedilir. Yani iki grup ortalaması birbirine eşit değildir ve gruplar arası istatistiksel olarak önemli kabul edilebilecek bir fark vardır. $|Z_0| < Z$ tablo değeri ise H_0 hipotezi kabul edilir. Sonucunda iki grup ortalamaları arasında istatistiksel açıdan önemli kabul edilebilecek bir fark yoktur.

Normal dağılım eğrisine göre yorumlayacak olursak; -1,96'dan küçük ve 1,96'dan büyük alanlara (kırmızı alan) düşen Z_0 değeri sonucu H_0 hipotezi reddedilir, yani grupların ortalamaları arasında istatistiksel olarak önemli kabul edilebilecek bir fark vardır. Z_0 değeri dağılımın içine düşerse H_0 hipotezi kabul edilir. Yani grupların ortalamaları arasında istatistiksel olarak önemli kabul edilebilecek bir fark yoktur (44).



Şekil 3.1: Normal dağılım eğrisi

3.1.3.1.1.Geniş Örneklemler İçin Güven Aralığı

$\mu_X - \mu_Y$için%(1-α) güven aralığı denklem 3.10'da verilmiştir.

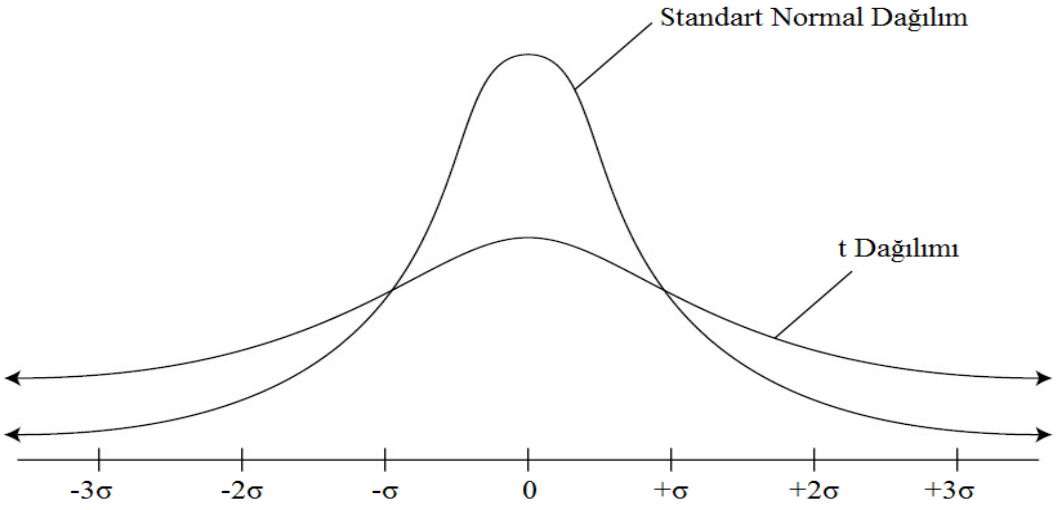
Denklem 3.10. Geniş örneklemler için güven aralığı

$$GA(\alpha): \bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}} SE(\bar{x} - \bar{y}) = \bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(x_i - \bar{x})^2}{n_1 - 1} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(y_i - \bar{y})^2}{n_2 - 1}}$$

$z_{\frac{\alpha}{2}}$;standart normal dağılımdaki $\frac{\alpha}{2}$ için kritik değerdir.

3.1.3.2.Küçük Örneklemler İçin Anlamlılık Testi

Örneklem genişliğinin 30'dan küçük olduğu durumlarda kullanılır. Küçük örnek istatistiklerinin gösterdikleri dağılım normal dağılım eğrisi gibi simetriktir. Dağılımın şekli standart normal dağılıma göre daha yayvan bir şekle dönüşür. Çeşitli örnek büyüklükleri ve ihtimal seviyeleri için ayrı ayrı hesaplanmış olan t-tablosu kullanılır(43).



Şekil 3.2: Standart normal dağılım ve t dağılımı

Standart yokluk hipotezimiz $H_0: \mu_X - \mu_Y = \mu_0$ (yaniμ_0=0) olmak üzere test istatistiği denklem 3.11'deki gibidir.

Denklem 3.11. T test istatistiği

$$T_0 = \frac{\bar{x} - \bar{y} - \mu_0}{SE(\bar{x} - \bar{y})} \sim T_{(df)}$$

Test istatistiği yaklaşık olarak aynı serbestlik dereceli dağılımı gösterir. Serbestik derecesini hesaplarken denklem 3.12'den yararlanılır.

Denklem 3.12. T testi serbestlik derecesi

$$df = \frac{(SE^2(\bar{x}) + SE^2(y))^2}{\frac{SE^4(\bar{x})}{n_1 - 1} + \frac{SE^4(\bar{y})}{n_2 - 1}} \approx n_1 - n_2 - 2$$

T dağılımına ait ortak standart hata denklem 3.13'deki gibidir.

Denklem 3.13. T dağılımı ortak standart hata

$$SE(\bar{x} - \bar{y}) = \sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{(x_i - \bar{x})^2}{n_1 - 1} + \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{(y_i - \bar{y})^2}{n_2 - 1}}$$

Denklem 3.14. T test istatistiği(açık)

$$t_0 = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{(x_i - \bar{x})^2}{n_1 - 1} + \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{(y_i - \bar{y})^2}{n_2 - 1}}}$$

Elde edilen $t_0$değeri Ek 2'de gösterilen t tablosundaki karşılık gelen değer ile karşılaştırılır. Tek yönlü hipotez ve çift yönlü hipoteze göre tabloda seçilen değerler değişir. Çift yönlü hipotezler için α=0.025 değerlerine bakılırken, tek yönlü hipotezler içinα=0.05 değerlerine bakılır. Bu değerlerle eşleşen $(n_1 + n_2 - 2)$ serbsetlik derecesine karşılık gelen t tablo değeri t_0 değeri ile karşılaştırılır. Buna göre eğer $|t_0| \geq t$ tablo değeri ise H_0 hipotezi reddedilir. Yani iki grup ortalaması birbirine eşit değildir ve gruplar arası istatistiksel olarak önemli kabul edilebilecek bir fark vardır. $|t_0| < t$tablo değeri ise H_0 hipotezi kabul edilir. Sonucunda iki grup ortalamaları arasında istatistiksel açıdan önemli kabul edilebilecek bir fark yoktur (44).

3.1.3.2.1.Küçük Örneklemler İçin Güven Aralığı

$\mu_X - \mu_Y$için%(1-α) güven aralığı denklem 3.15'deki gibidir.

Denklem 3.15. Küçük örneklemler için güven aralığı

$$GA(\alpha): \bar{x} - \bar{y} \pm t_{df,\frac{\alpha}{2}} SE(\bar{x} - \bar{y}) = \bar{x} - \bar{y} \pm t_{df,\frac{\alpha}{2}} \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(x_i - \bar{x})^2}{n_1 - 1} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{(y_i - \bar{y})^2}{n_2 - 1}}$$

$t_{df,\frac{\alpha}{2}}$; student t dağılımdaki $\frac{\alpha}{2}$ için kritik değerdir.

3.2.Mikrodizi Anlamlılık Analizi

Mikrodizi Anlamlılık Analizi, mikrodizi veri setleri için farklı koşullardaki gen ekspresyonları arasındaki farkı tanımlamak ve bu değişimlerin anlamlılığını test etmek için geliştirilmiş bir yöntemdir. Literatürde Significance Analysis of Microarray (SAM) adı ile geçer. SAM analizi, t-testinin geliştirilmiş bir versiyonudur(45).

A ve B birbirinden bağımsız iki grup olmaktave i her bir geni göstermek üzere, $\bar{x}_A(i)$ hasta grubuna aitekspresyon ortalamaları, $\bar{x}_B(i)$kontrol grubuna aitekspresyon ortalamaları, $s(i)$ iki örneklem t-testi paydasındaki standart sapma ve $s_0$tüm genlerde sabit olan ve tüm yoğunluk seviyelerinin genleri arasında varyasyon yaratabilmek amacıyla eklenmiş bir terim olmak üzere SAM analizine ait hesaplama denklem 3.16'daki gibidir (46).

Denklem 3.16. Mikrodizi Anlamlılık Analizi

$$d(i) = \frac{\bar{x}_A(i) - \bar{x}_B(i)}{s(i) + s_0}$$

Denklem 3.17. Standart sapma ise

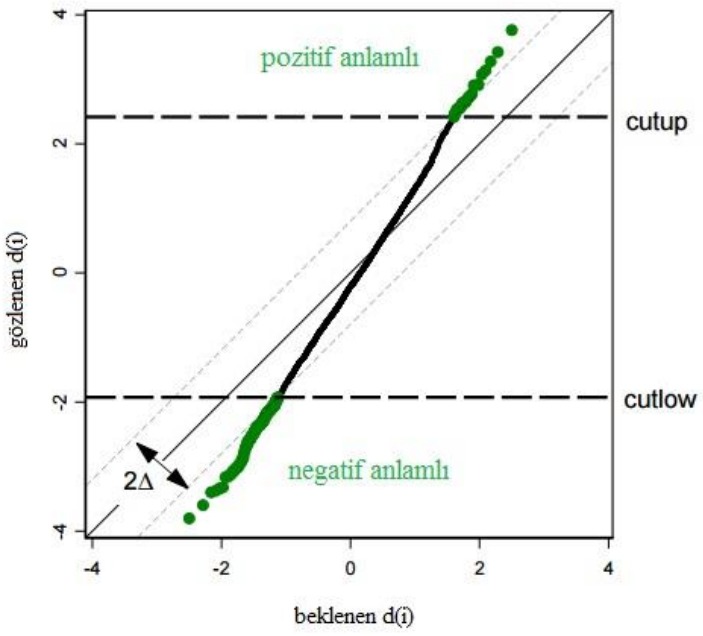
$$s(i) = \sqrt{\left(\frac{\frac{1}{n_A} + \frac{1}{n_B}}{n_A + n_B - 2}\right)\left\{\sum_{k=1}^{n_A}[x_k(i) - \bar{x}_A(i)]^2 + \sum_{k=n_A+1}^{n_A+n_B}[x_k(i) - \bar{x}_B(i)]^2\right\}}$$

3.2.1.Mikrodizi Anlamlılık Testi Aşamaları

1) i=1,...,m olmak üzere her bir gen için $d_{(i)}$ ekspresyon skoru hesaplanır. Bu değer genin gözlenen d-değeri olarak adlandırılır. $d_{(1)} \leq . . . \leq d_{(m)}$ olacak şekilde sıralanır.
2) Birinci grup ve ikinci grup arasındaki genler B permutasyon sayısı olmak üzere, olası bütün B'ler kadar rasgele karıştırılır. Karıştırılan bu yeni gruplarda, orijinal

grup ile eşit eleman sayısına sahiptir. Her bir b permutasyonu için, i=1,...,m olmak üzere $d_i{}^b$ ekspresyon skoru oluşturulur. Beklenen d-değeri, $\bar{d}_\iota = \sum_b d_i{}^b / B_i$ şeklinde hesaplanır.

3) d_i gözlenen değere karşı $\bar{d}_\iota$ beklenen değerin grafiği (SAM plot) Şekil 3.3'deki gibi çizilir.
4) $\Delta>0$ olmak üzere araştırmacı tarafından belirlenen sabit bir eşik değeridir. $d_i - \bar{d}_\iota \geq \Delta$ olacak şekilde merkezin sağında $(\bar{d}_{i1}, d_{i1})$ noktası bulunur ve $d_{i1} = cut_{up}(\Delta)$ belirlenir. $d_i \geq cut_{up}(\Delta)$ ise i. gen pozitif anlamlı olarak adlandırılır. Benzer şekilde $d_i - \bar{d}_\iota \leq -\Delta$ olacak şekilde merkezin solunda $(\bar{d}_{i2}, d_{i2})$ noktası bulunur ve $d_{i2} = cut_{low}(\Delta)$ belirlenir. $d_i \leq cut_{low}(\Delta)$ ise i. gen negatif anlamlı olarak adlandırılır.
5) $FDR = \frac{(\frac{1}{B}) \sum_b \#\{d_i{}^b \notin (cut_{low}(\Delta), cut_{up}(\Delta))\}}{\#\{d_i \notin (cut_{low}(\Delta), cut_{up}(\Delta))\}}$ hesabı ile yanlış bulgu oranı (FDR) tahmin edilir.
6) Δ eşik değerine göre 4. ve 5. adımlar pek çok kez tekrar edilir. Tanımlanan genlerein sayısı arasındaki en iyi dengeyi sağlayan Δ değeri seçilir ve FDR tahmin edilir(47,48).



Şekil 3.3: SAM plot

3.3. Yanlış Bulgu Oranı

Yanlış bulgu oranı (FDR), yanlışlıkla reddedilen hipotezlerin beklenen oranı olarak ifade edilmektedir. FDR tüm hipotezler doğru olduğunda ailesel hata oranı (FWER) değerine eşit olmaktadır. Hipotezlerden en az bir tanesinin doğru olmaması durumunda FDR değeri FWER değerinden daha küçük olmakta, dolayısıyla da istatistiksel gücü artırdığından FWER yerine FDR'nin kullanılması daha çok arzu edilmektedir (49).

Çizelge 3.1: Hipotez Testi Karar Çizelgesi

	Yokluk Hipotezi Doğru (H_0)	Alternatif Hipotez Doğru (H_1)	Toplam
Önemli (Red)	V	S	R
Önemsiz (Kabul)	U	T	m-R
Toplam	m_0	$m - m_0$	m

Çizelge 3.1'dem toplam hipotez testi sayısı,m_0 yokluk hipotezi sayısı, $m - m_0$ alternatif hipotez sayısı, V; yanlış bulgular olarak adlandırılan yanlış pozitiflerin sayısı (I. Tip Hata), S; doğru bulgular olarak adlandırılan doğru pozitiflerin sayısı, T yanlış negatiflerin sayısı (II. Tip Hata), U doğru negatiflerin sayısı, R bulgu olarak adlandırılan reddedilen yokluk hipotezinin sayısı ve Q bulgular arasındaki yanlış bulguların oranı olmak üzere, yanlış bulgu oranı denklem 3.21'deki gibi hesaplanır (49).

Denklem 3.21. Yanlış bulgu oranı

$$Q = \frac{V}{R}$$

$$FDR = E(Q) = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$$

3.3.1. Benjamini Hochberg ve Benjamini Yekutieli Yanlış Bulgu Oranı

Benjamini Hochberg en çok kullanılan FDR yöntemi yöntemidir. Bu yöntem, istatistiksel testler birbirinden bağımsız olduğu durumda, güçlü ve en az 'yanlış negatif' vermesinden dolayı, yanlış bulgu oranını kontrol etmek için en uygun yöntemdir.

Benjamini Hochberg yönteminde p değerleri $p_1 \leq \cdots \leq p_m$ şeklinde sıralanır. Bir k değişkeni k = max{i : $p_{(i)} \leq$ (i/m) q} şeklinde hesaplanır (q=0,05). Buna göre; $H_0{}^{(1)}, \ldots, H_0{}^{k}$ hipotezleri reddedilir. Diğer bir deyişle Hochberg yönteminde p değerleri en küçükten en büyüğe doğru sıralanır. En büyük p değeri hiçbir değişiklik yapmadan bırakılır. En büyük ikinci p değeri, gen listesindeki toplam gen sayısı ile çarpılarak sıra numaralarına bölünür. Buradan ilk düzeltilmiş p değeri aşağıdaki gibi hesaplanır.

Düzeltilmiş p değeri= p değeri x (n/n-1)

İkinci düzeltilmiş p değeri aşağıdaki gibi hesaplanır.

Düzeltilmiş p değeri= p değeri x (n/n-2)

Böylece n-(n-1). gene kadar devam edilir. Belirlenen bir anlamlılık düzeyi (0,05) ile karşılaştırma yapılarak, bu değerden küçük olanlar anlamlı ve büyük olanlar anlamsız kabul edilir. Buradan yanlış pozitif oranı (YPO) yani 1. Tip hata oranıdenklem 3.22 ve yanlış negatif oranı (YNO)denklem 3.23'deki gibi hesaplanır (50).

Denklem 3.22. Yanlış pozitif oranı

$$YPO = \frac{\text{Düzeltilmiş p değeri } \leq \ 0{,}05}{n - \ \text{Düzeltilmiş p değeri } \leq \ 0{,}05 \text{ sayısı}}$$

Denklem 3.23. Yanlış negatif oranı

$$YNO = \frac{\text{Düzeltilmiş p değeri} > \ 0{,}05}{n - \ \text{Düzeltilmiş p değeri} > \ 0{,}05 \text{ sayısı}}$$

Benjamini Yekutieli prosedürü ise genetik yapısı gereği, kromozomların içinde pozitif regresyon bağımlılığı bulunduğundan dolayı Benjamini ve Yekutelli tarafından önerilmiştir (51).

Benjamini Yekutieli prosedüründe p değerleri küçükten büyüğe doğru $p_1 \leq \cdots \leq p_m$ şeklinde sıralanır. Bir k değeri k = max{i : $p_{(i)} \leq (i/m)\ \tilde{q}$} şeklinde belirlenir. Burada $\tilde{q} = \sum_{i=1}^{m} \frac{q}{\frac{1}{i}}$ şeklinde hesaplanmaktadır.Buna göre; ${H_0}^{(1)}, \ldots, {H_0}^{k}$ hipotezleri reddedilir (51).

4. BULGULAR

Adenoma ve Carcinoma veri setlerine ait tanımlayıcı bilgilerÇizelge 4.1'de verilmiştir. Çizelge 4.1 incelendiğinde, Adenoma veri setinde Student t testi sonuçlarına göre 7086 toplam genden 715 genin hasta ve kontrol gruplarında istatistik açıdan farklılık gösterdiği($p \leq 0,05$), geriye kalan 6371 tanesinin ise istatistik açıdan farklılık göstermediği bulunmuştur ($p > 0.05$). İstatistiksel olarak anlamlı çıkan genlerin, p değerlerine ait ortalama ve standart sapması 0,02126±0,01487, istatistiksel olarak anlamlı bulunmayan genlerin p değerlerine ait ortalama ve standart sapması0,466232±0,267336 olarak bulunmuştur.

Carcinoma veri setinde bulunan 7457 gen için hasta ve kontrol grubunun karşılaştırılması için Student t testi kullanılmıştır. İstatistiksel olarak anlamlı çıkan genlerin sayısı 1780 iken anlamlı bulunmayan genlerin sayısı 5677 olarak hesaplanmıştır. p değerlerine ait ortalama ve standart sapma değerleri sırasıyla 0,01313±0,01498 ve 0,46374±0,28159 olarak bulunmuştur.

Çizelge 4.1: Adenoma ve Carcinoma veri setlerine ait tanımlayıcı bilgiler

Veri Seti	Hasta-Kontrol Sayısı	Toplam Gen Sayısı	Anlamlı Gen Sayısı	Anlamlı Olmayan Gen Sayısı	Anlamlı Genlere Ait p değerleri Ortalaması	Anlamlı Olmayan Genlere Ait p değerleri Ortalaması
ADENOMA	4-4	7086	715	6371	0,02126±0,01487	0,46623±0,26734
CARCINOMA	18-18	7457	1780	5677	0,01313±0,01498	0,46374±0,28159

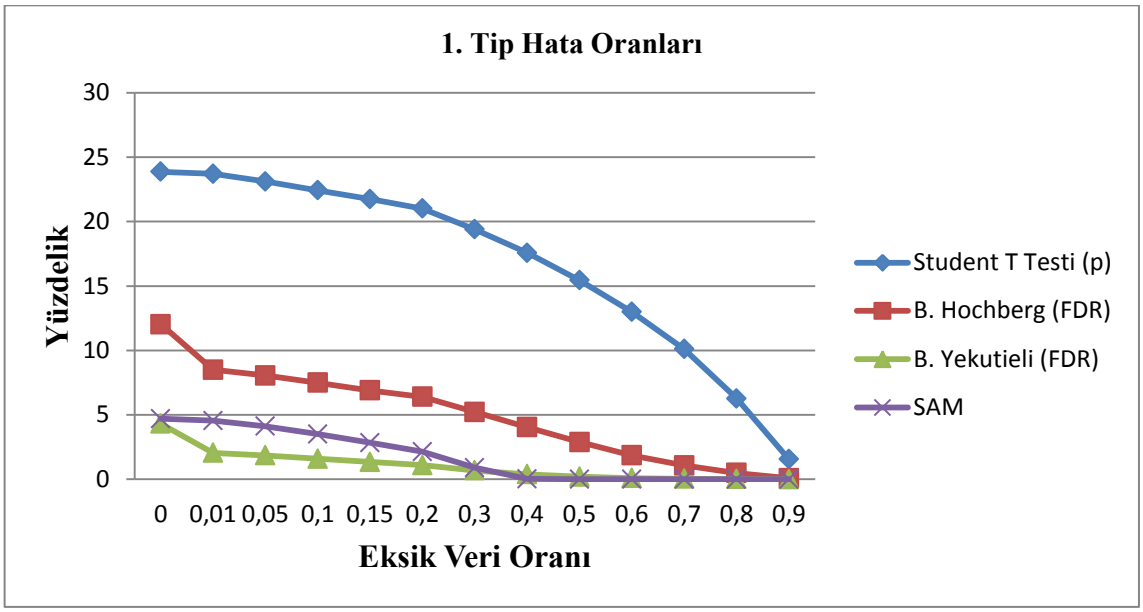
Ayrıca Benjamin Yekutieli prosedürünün varsayımının sağlanıp sağlanmadığının kontrolü için her iki veri setine ait Pearson korelasyon katsayılarına bakılmıştır. Adenoma veri seti için 0.80 'in üzerinde korelasyona sahip gen çifti sayısı 523,538 ve Carcinoma veri seti için %80'in üzerinde korelasyona sahip gen çifti sayısı 4628 olarak bulunmuştur.

Çizelge 4.2’de Carcinoma veri setinde eksik veri oranlarına karşılık gelen Student t testi, Benjamini Hochberg, Benjamini Yekutieli yöntemlerinde ait 1000 tane verinin p ve FDR değerlerine ait ortalama ve standart sapmaları verilmiştir. Ayrıca, sözü edilen yöntemler ve SAM analizi sonucunda anlamlı bulunan 1780 gendeki kayıp veri oranlarına göre ortalama ve standart sapmaları ve 1. Tip hata oranları verilmiştir. Şekil 4.1’de de grafiksel olarak sunulmuştur.

Anlamlı genlere ait Student t testi sonucu elde edilen p değerleri ortalaması %0'lık eksik veri yapısında 0,01313±0,01498 olarak ve 1780 adet anlamlı olan gen ve 1. tip hatası %23,87 olarak bulunmuştur. Eksiklik yüzdesi arttıkça anlamlı bulunan genlerin sayısında azalma meydana gelmektedir ve 1. tip hataları azalmaktadır.%1'lik eksik veri durumda ise 1. tip hata düzeyinde azalma meydana gelmiştir ve anlamlı genlerin p değerlerinin ortalaması 0,01311±0,00014 olmuştur. Eksiklik yüzdeleri arttıkça 1. tip hata düzeyleri azalmaktadır. Benjamini Hochberg prosedüründe, FDR değerlerinin sayısı gerçek veride 897 ve FDR değerlerine ait ortalama, standart sapma 0,01630±0,01540 olarak bulunmuştur. Eksiklik oranı arttıkça FDR değerlerinin ortalamaları yükselmekte, sayılarının oranı azalmakta ve 1. Tip hata değerleri azalmaktadır. Benjamini Yekutieli prosedüründe eksik veri olmadığı durumda anlamlı genlerin sayısı 321 olarak bulunmuştur ve eksiklik oranı arttıkça bu sayı azalmaktadır. %90 eksiklik oranına ulaşıldığında neredeyse hiç anlamlı gen kalmamaktadır. 1. Tip hata oranları gerçek veride %4,31 olarak hesaplanmıştır. %1’lik eksik veri durumda bu oran yarısı kadar azalmıştır ve %90’lara ulaşıldığında %0,01’den daha küçük bir değere sahiptir. Düzeltilmemiş veri setine Benjamini Hochberg prosedüründen daha yakın sonuçlar göstermektedir. Yekutieli'de de eksiklik yüzdesi arttıkça FDR ortalamalarında %80'e kadar artış gözlenmektedir, fakat %80 ve %90 eksik gözlem durumunda azalışa geçmekte ve %90'lık eksik veri durumunda en küçük değere sahip olmaktadır. SAM analizi sonuçlarına göre gerçek veride anlamlı gen sayısı 348 ve 1. Tip hata oranı %4,7 olarak bulunmuştur. Eksiklik oranı arttıkça 1. Tip hatada azalmalar olmuştur ve %40 eksik veri durumundan sonra Şekil 4.1’de de gösterildiği gibi hiçbir hesaplama yapılamamıştır.

Çizelge4.2: Carcinoma anlamlı bulunan genlere ait bulgular

Carcinoma	Ortalama ± Standart Sapma			Ortalama ± Standart Sapma (Anlamlı Gen Sayısı)				1. Tip Hata Oranı			
Eksik Veri Oranı	Student T Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)	Student T Testi	B.Hochberg	B.Yekutieli	SAM	Student t Testi	B.Hochberg	B.Yekutieli	SAM
%0	0,01313±0,01498	0,01630±0,01540	0,01430±0,01420	1780	897	321	348	23,87	12,03	4,31	4,70
%1	0,01311±0,00014	0,01641±0,00022	0,01831±0,00048	1767,6±7,7	632,3±4,7	152,6±2,7	339,0±4,7	23,70	8,50	2,05	4,55
%5	0,01317±0,00022	0,01679±0,00037	0,01915±0,00084	1723,5±13,2	600,2±8,2	137,9±4,6	306,12±8,2	23,11	8,05	1,85	4,11
%10	0,01342±0,00027	0,01722±0,00044	0,02011±0,00101	1672,4±16,1	558,6±9,8	118,9±5,4	261,6±9,5	22,43	7,50	1,60	3,51
%15	0,01379±0,00028	0,01776±0,00050	0,02107±0,00123	1621,0±18,1	517,4±10,4	99,8±5,9	212,1±10,8	21,74	6,90	1,34	2,84
%20	0,01422±0,00029	0,01832±0,00056	0,02204±0,00141	1566,7±19,0	474,5±11,5	82,2±5,6	159,5±14,2	21,01	6,40	1,10	2,14
%30	0,01517±0,00034	0,01975±0,00064	0,02359±0,00191	1446,4±20,8	389,3±11,4	51,4±5,5	67,1±10,6	19,40	5,22	0,69	0,90
%40	0,01636±0,00037	0,02143±0,00080	0,02482±0,00250	1309,9±22,2	301,2±11,9	28,9±4,5	2,4±5,5	17,57	4,04	0,39	0,03
%50	0,01771±0,00040	0,02331±0,00091	0,02568±0,00377	1151,0±22,6	215,1±11,2	14,7±3,6	0	15,44	2,88	0,20	-
%60	0,01929±0,00047	0,02506±0,00121	0,02605±0,00581	968,70±22,3	138,8±10,0	6,9±2,5	0	12,99	1,86	0,09	-
%70	0,02106±0,00053	0,02653±0,00160	0,02484±0,01029	753,1±23,6	79,7±8,1	3,1±1,7	0	10,1	1,07	0,04	-
%80	0,02275±0,00072	0,02757±0,00234	0,01752±0,01584	467,3±19,3	36,7±6,0	1,1±1,0	0	6,27	0,49	0,02	-
%90	0,02466±0,00137	0,02935±0,00608	0,00307±0,00995	116,2±10,7	6,2±2,5	0,1±0,3	0	1,56	0,08	<0,01	-

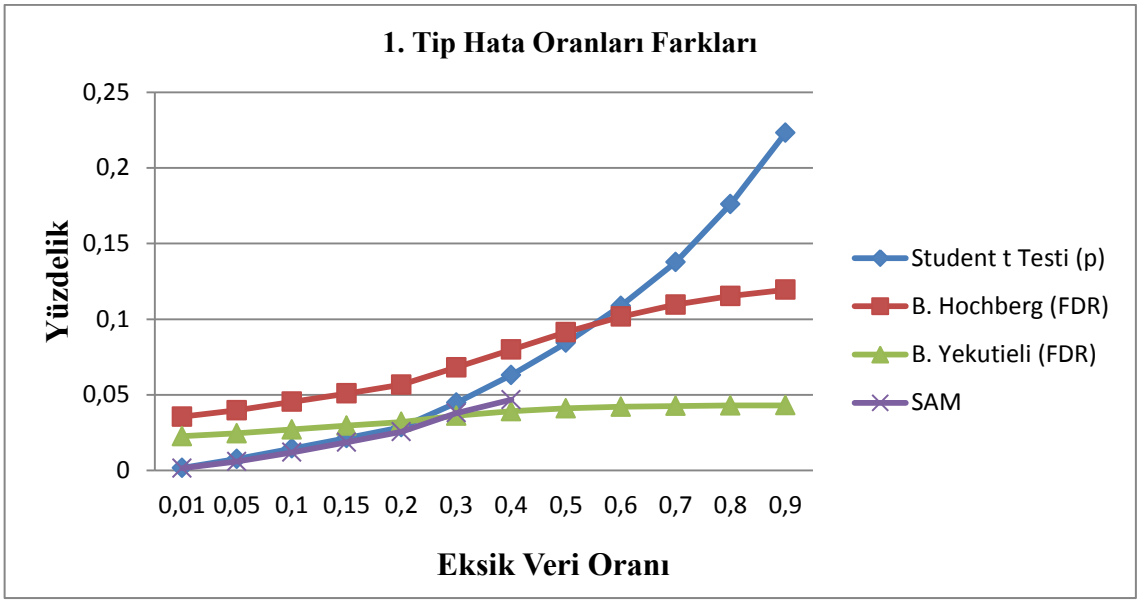


Şekil 4.1: Carcinoma 1.Tip hata oranları

Carcinoma verisinde eksik veri olmadığı gerçek veri seti ile eksik verilerin olduğu durumlardaki 1. Tip hatalara ait farklar alınarak Çizelge 4.3'teki değerler elde edilmiştir. Ayrıca Şekil 4.2'de de bu bilgiler grafiksel olarak sunulmuştur. Şekil 4.2 incelendiğinde %20 eksik veri yapısının kırılma noktası olduğu görülmektedir. Özellikle Student t testinde bu fark daha açık bir şekilde ortaya çıkmaktadır. Hochberg prosedüründe t testine göre daha yavaş bir artış vardır. Gerçek veri seti ile t testinin farkları diğer prosedürlere göre daha az iken Yekutieli ile bu fark %20 eksiklikte eşitlenmektedir, Hochberg ile ise fark %55 eksikliğe geldiğinde eşitlenmektedir. Yekutieli prosedüründe diğer üç teste göre kırılmaların daha az olduğu görülmektedir. SAM analizi %1 ve %5'te t testi ile yakın iken bu fark eksiklik oranı arttıkça artmaktadır.

Çizelge 4.3: Carcinoma eksik veriler ile gerçek veriarasındaki 1. Tip hata oranları farkı

Eksik Veri Oranı	Student t Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)	SAM
%1	0,00166	0,0355	0,02258	0,0015
%5	0,00757	0,03981	0,02455	0,0059
%10	0,01443	0,04538	0,02711	0,0119
%15	0,02132	0,0509	0,02966	0,0186
%20	0,0286	0,05666	0,03203	0,0256
%30	0,04473	0,06809	0,03616	0,038
%40	0,06303	0,0799	0,03917	0,0467
%50	0,08435	0,09145	0,04108	-
%60	0,1088	0,10167	0,04212	-
%70	0,1377	0,1096	0,04263	-
%80	0,17603	0,11537	0,0429	-
%90	0,22312	0,11946	0,04303	-

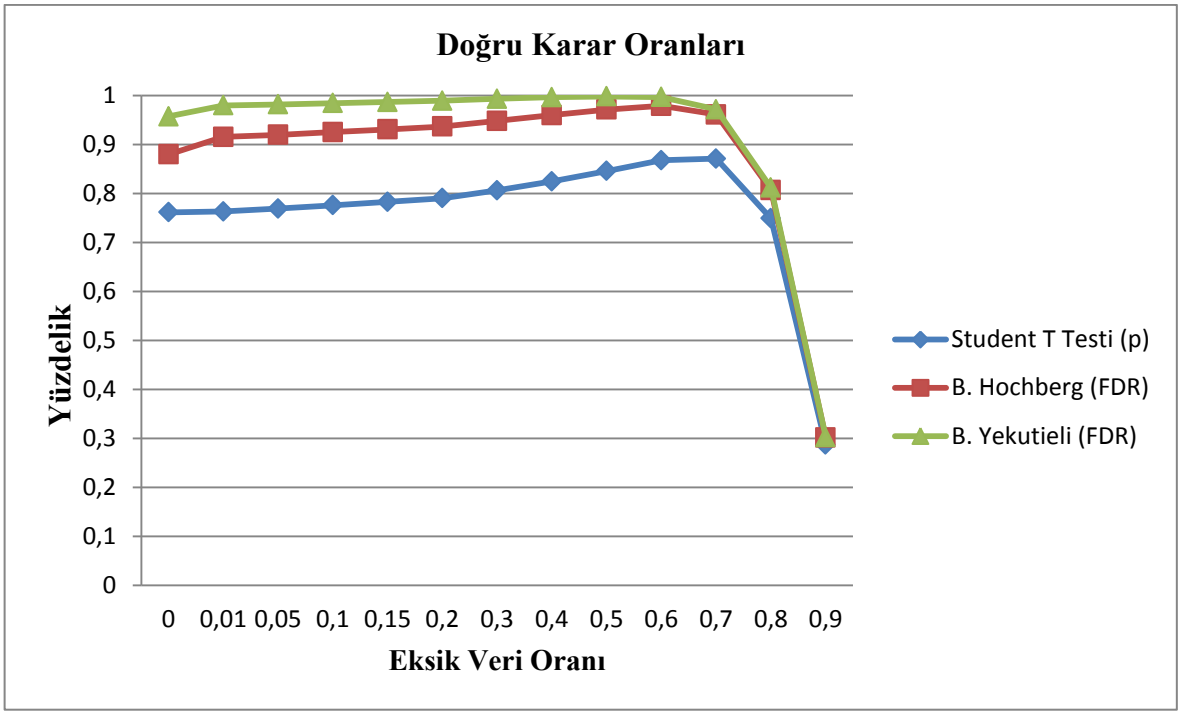


Şekil 4.2: Carcinoma eksik veriler ile gerçek veri arasındaki 1. Tip hata oranları farkı

Çizelge 4.4 ve Şekil4.3'te, t testi sonuçlarına göre %0 eksik veri durumda p değerlerinin ortalaması 0,46374±0,28159 olarak bulunmuştur ve %1, %5 ve %10 eksik veri durumlarında gerçek veri sonucuna göre azalma meydana gelmektedir. Ancak %15'den %90 eksik veri durumuna kadar p değerleri ortalamalarında artış gözlenmektedir. %0'dan %70'e kadar doğru karar oranlarında artış gözlenmektedir. %80 ve %90 eksik veri ciddi bir azalış yaşanmaktadır. Benjamini Hochberg ve Yekutieli prosedürlerinde de eksiklik yüzdesi arttıkça FDR değerlerinin ortalamalarının arttığı ve %80 eksik veri yapısına kadar, doğru karar oranları da bir artış olduğu görülmektedir. %80 ve %90 eksik veri durumunda hesaplanan doğru karar oranları birbirine çok yakındır.

Çizelge 4.4: Carcinoma anlamlı bulunmayan genlere ait bulgular

Carcinoma	Ortalama±Standart Sapma			Ortalama±Standart Sapma (Anlamlı Olmayan Gen Sayısı)			Doğru Karar Oranı		
Eksik Veri Oranı	Student T Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)	Student T Testi	B.Hochberg	B.Yekutieli	Student t Testi	B.Hochberg	B.Yekutieli
%0	0,46374±0,28159	0,59640±0,28020	5,22400±2,94870	5677	6560	7136	76,13	87,97	95,70
%1	0,46357±0,00085	0,67281±0,00060	11,89955±0,00910	5689,4±7,7	6824,7±4,7	7304,4±2,7	76,30	91,52	97,95
%5	0,46336±0,00170	0,67400±0,00117	11,95200±0,01898	5733,5±13,2	6856,8±8,2	7319,1±4,6	76,90	91,95	98,15
%10	0,46342±0,00220	0,67563±0,00151	12,01943±0,02498	5784,6±16,1	6898,4±9,8	7338,1±5,4	77,57	92,51	98,40
%15	0,46376±0,00253	0,67746±0,00186	12,09173±0,03096	5836,0±18,1	6939,5±10,4	7357,2±5,9	78,26	93,06	98,66
%20	0,46391±0,00273	0,67943±0,00199	12,17009±0,03487	5890,3±18,9	6982,5±11,5	7374,8±5,6	78,99	93,64	98,90
%30	0,46491±0,00291	0,68452±0,00224	12,35931±0,04026	6010,6±20,8	7067,7±11,4	7405,6±5,5	80,60	94,78	99,31
%40	0,46589±0,00316	0,69098±0,00241	12,58918±0,04475	6147,0±22,2	7155,8±11,9	7428,1±4,5	82,43	95,96	99,61
%50	0,46744±0,00327	0,69984±0,00262	12,87356±0,04651	6304,90±22,6	7240,9±11,3	7441,2±3,8	84,55	97,10	99,79
%60	0,47024±0,00318	0,71212±0,00273	13,21756±0,04768	6468,7±22,8	7298,6±10,8	7430,4±5,2	86,75	97,88	99,64
%70	0,47605±0,00328	0,72927±0,00267	13,63507±0,05080	6494,1±27,2	7167,0±15,6	7244,1±14,2	87,09	96,11	97,15
%80	0,48555±0,00366	0,75153±0,00300	14,11237±0,05762	5585,0±33,7	6016,2±28,9	6051,2±28,6	74,90	80,68	81,15
%90	0,49947±0,00579	0,77738±0,00474	14,64827±0,09350	2137,4±30,1	2248,90±29,0	2253,5±29,5	28,66	30,16	30,22

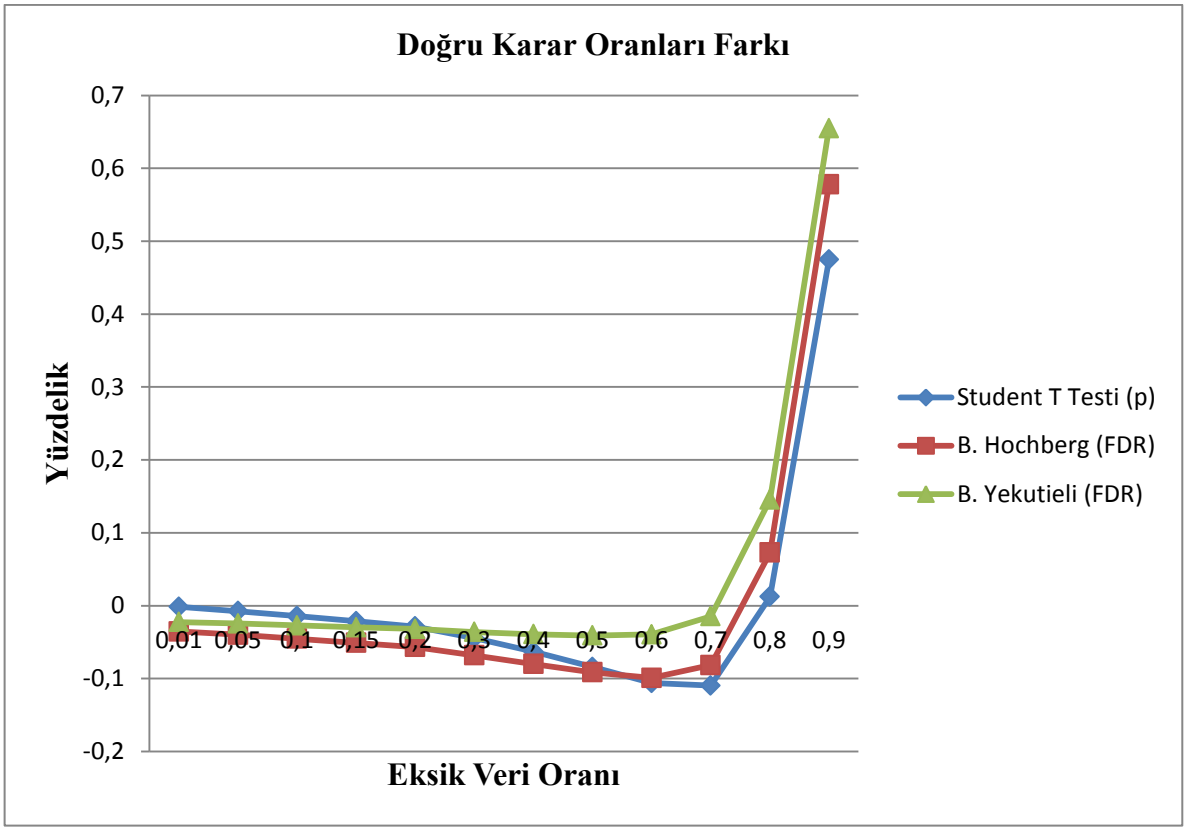


Şekil 4.3: Carcinoma doğru karar oranı

Çizelge 4.5 ve Şekil 4.4'te doğru karar oranlarının, gerçek veri seti ile olan farklarına göre %80 eksik veri durumuna kadar Student t testi ve Hochberg prosedürü birbirine yakın oranlarda farka sahiptir. Ancak Yekutieli prosedürü %70 eksiklik durumuna kadar eksiklik yüzdelerinden en az etkilenen prosedürdür.

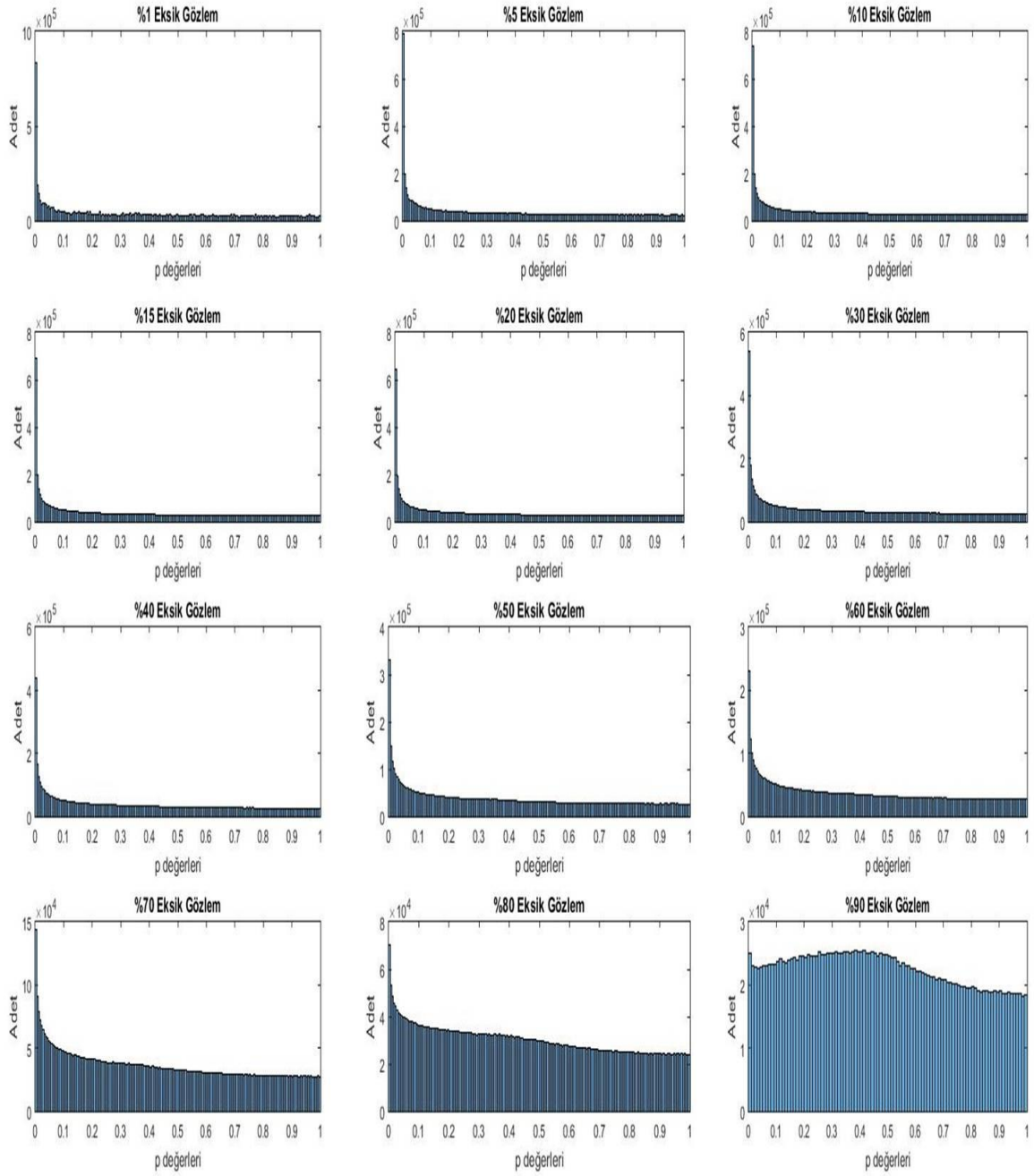
Çizelge 4.5: Carcinoma eksik veriler ile gerçek veri arasındaki doğru karar oranları farkı

Eksik Veri Oranı	**Student T Testi (p)**	**B.Hochberg (FDR)**	**B.Yekutieli (FDR)**
%1	-0,00166	-0,0355	-0,02253
%5	-0,00757	-0,03981	-0,0245
%10	-0,01443	-0,04538	-0,02706
%15	-0,02132	-0,0509	-0,02961
%20	-0,0286	-0,05666	-0,03198
%30	-0,04473	-0,06809	-0,03611
%40	-0,06303	-0,0799	-0,03912
%50	-0,0842	-0,09131	-0,04089
%60	-0,10617	-0,09905	-0,03944
%70	-0,10957	-0,08141	-0,01445
%80	0,01234	0,07292	0,14552
%90	0,47467	0,57813	0,65480



Şekil 4.4: Carcinoma eksik veriler ile gerçek veri arasındaki doğru karar oranları farkı

Carcinoma veri setinde Student t testi sonucu hesaplanan bütün p değerlerine ait dağılımda, eksiklik oranının artması ile hesaplanabilen testlerin sayılarında azalma, dolayısıyla p değerlerinin sayılarında azalma olduğu ve %60 eksik veri durumundan sonra dağılımın şeklinin değiştiği Şekil 4.5'da görülmektedir.

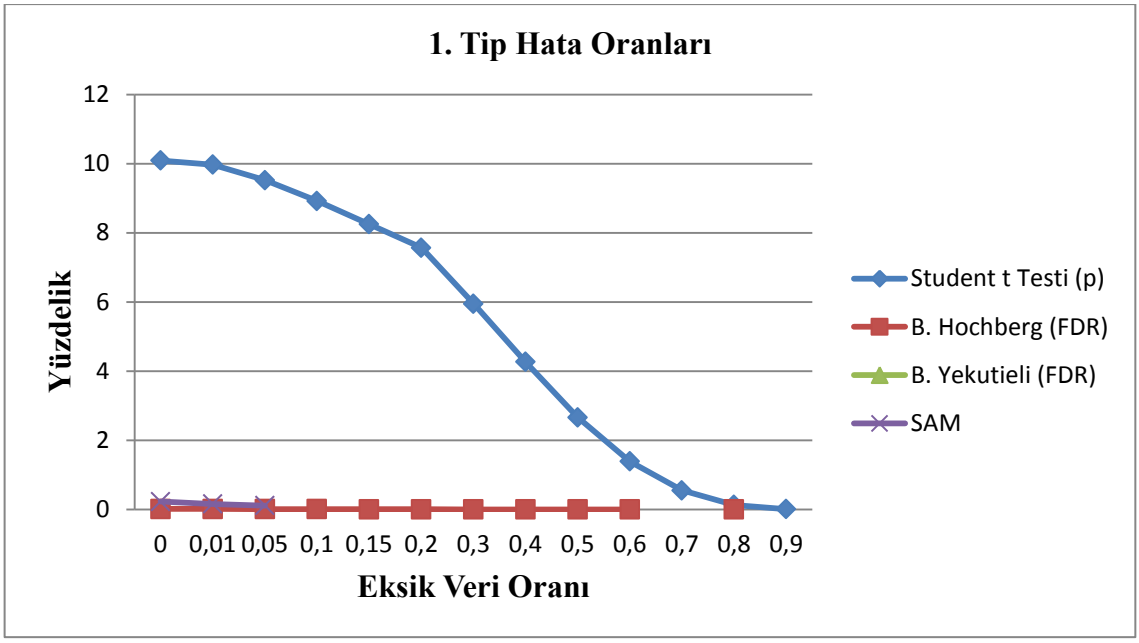


Şekil 4.5: Carcinoma Student t testine ait p değerlerinin dağılımı

Çizelge 4.6 ve Şekil 4.6'da gösterildiği gibi Student t testi sonuçlarına göre, gerçek veride 715 gen, 1. tip hatası %10,09olarak bulunmuştur. Eksiklik yüzdesi arttıkça p değerlerinin ortalamaları da artmaktadır ancak 1. tip hata oranı azalmaktadır. %90 eksik yapıda p değerlerinin ortalamasında düşme gözlenmiştir. Bunun sebebi anlamlı bulunan genlerde p değerleri 0,05'e yakın olanlarının örnek genişliklerinin azalması ile anlamsız hale dönüşmesidir. Benjamini Hochberg prosedürü, gerçek veride sadece tek bir geni anlamlı bulmuştur ve eksiklik oranı arttıkça 1. tip hatası 1000 deneme içerisinde oldukça düşüktür. %70 ve %90 eksik veri yapısında hiç anlamlı gen bulunmamıştır. Yekutieli prosedürü ise gerçek veri setinde hiç anlamlı gen bulamamıştır. Bu sebepten 1. tip hatalar hesaplanamamaktadır. SAM analizinde ise gerçek veri setinde 16 anlamlı gen bulunmuştur ve 1. Tip hatası 0,226 olarak hesaplanmıştır. SAM analizi %1 ve %5 eksiklik oranlarında da 1. Tip hatayı hesaplayabilirken, %10 ve üstü eksiklik durumlarında hesaplama yapamamaktadır.

Çizelge 4.6: Adenoma anlamlı bulunan genlere ait bulgular

Adenoma	Ortalama±Standart Sapma			Ortalama±Standart Sapma (Anlamlı Gen Sayısı)				1. Tip Hata Oranı			
Eksik Veri Oranı	Student t Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)	Student t Testi	B.Hochberg	B.Yekutieli	SAM	Student t Testi	B.Hochberg	B.Yekutieli	SAM
%0	0,02126±0,01487	0,011	-	715	1	0	16	10,09	0,014	-	0,226
%1	0,02132±0,00016	0,04488±0,01222	-	706,8±5,0	0,9±0,2	0	11,1±14,3	9,97	0,013	-	0,157
%5	0,02160±0,00033	0,03104±0,02305	-	674,3±10,4	0,7±0,5	0	7,8±16,6	9,52	0,009	-	0,111
%10	0,02190±0,00042	0,02218±0,02396	-	632,2±13,7	0,5±0,5	0	0	8,92	0,007	-	-
%15	0,02217±0,00050	0,01373±0,02172	-	584,9±14,7	0,30±0,5	0	0	8,25	0,004	-	-
%20	0,02244±0,00057	0,01058±0,01982	-	536,1±16,3	0,2±0,4	0	0	7,57	0,003	-	-
%30	0,02301±0,00066	0,00460±0,01399	-	421,8±16,1	0,1±0,3	0	0	5,95	0,001	-	-
%40	0,02355±0,00083	0,00161±0,00846	-	302,5±14,9	0,1±0,2	0	0	4,27	0,001	-	-
%50	0,02405±0,00109	0,00076±0,00591	-	188,5±13,1	0,01±0,1	0	0	2,66	<0,001	-	-
%60	0,02473±0,00142	0,00017±0,00271	-	98,50±9,4	0,005±0,07	0	0	1,39	<0,001	-	-
%70	0,02513±0,00233	-	-	39,0±5,9	0	0	0	0,55	-	-	-
%80	0,02583±0,00494	0,00005±0,00152	-	9,3±3,0	0,001±0,03	0	0	0,13	<0,001	-	-
%90	0,01288±0,01619	-	-	0,6±0,8	0	0	0	0,01	-	-	-

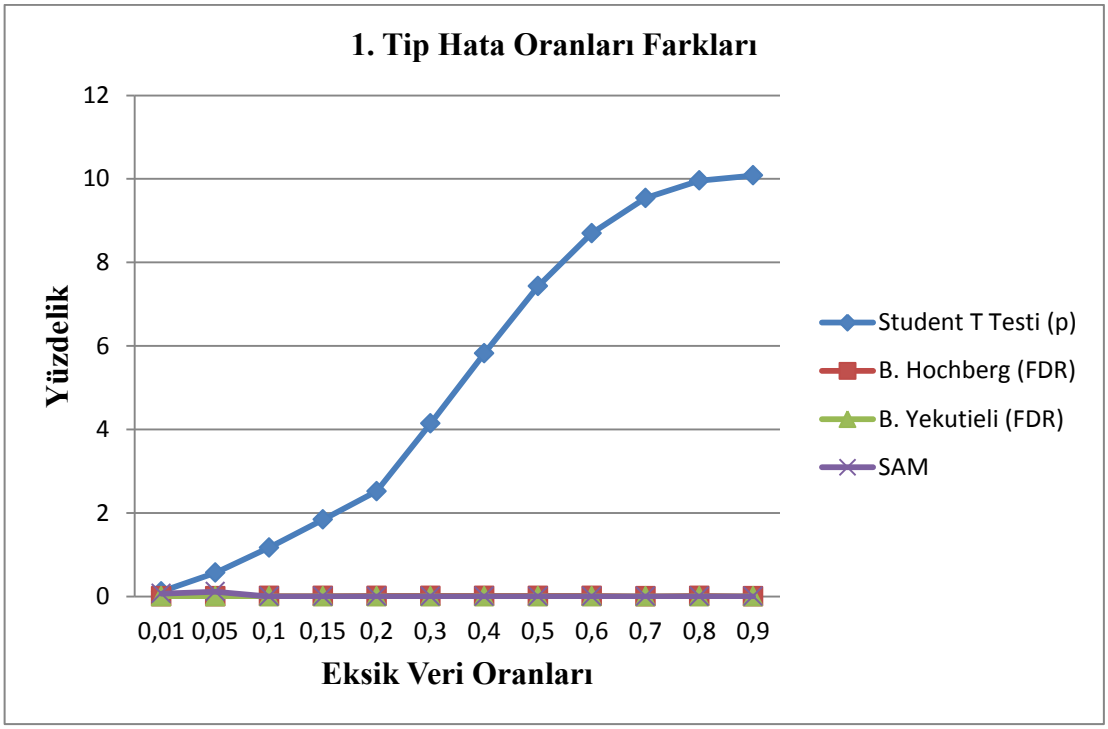


Şekil 4.6: Adenoma 1. Tip hata oranları

Çizelge 4.7 ve Şekil 4.7'de, Adenoma anlamlı bulunmayan genlere ait yüzdelerin gerçek veri setinde anlamlı bulunmayan genlerin yüzdesi ile olan farkında Student t testine göre farkların sürekli şekilde açıldığı görülmektedir. %20'ye kadar yavaş bir şekilde artarken, %20'den sonraki farkta daha büyük kırılma meydana gelmektedir.

Çizelge 4.7: Adenoma eksik veriler ile gerçek veri arasındaki 1. tip hata oranları farkı

Eksik Veri Oranı	**Student T Testi (p)**	**B.Hochberg (FDR)**	**B.Yekutieli (FDR)**	**SAM**
%1	0,12	0,001	-	0,069
%5	0,57	0,005	-	0,115
%10	1,17	0,007	-	-
%15	1,84	0,01	-	-
%20	2,52	0,011	-	-
%30	4,14	0,013	-	-
%40	5,82	0,013	-	-
%50	7,43	0,0139	-	-
%60	8,7	0,0139	-	-
%70	9,54	-	-	-
%80	9,96	0,01399	-	-
%90	10,08	-	-	-

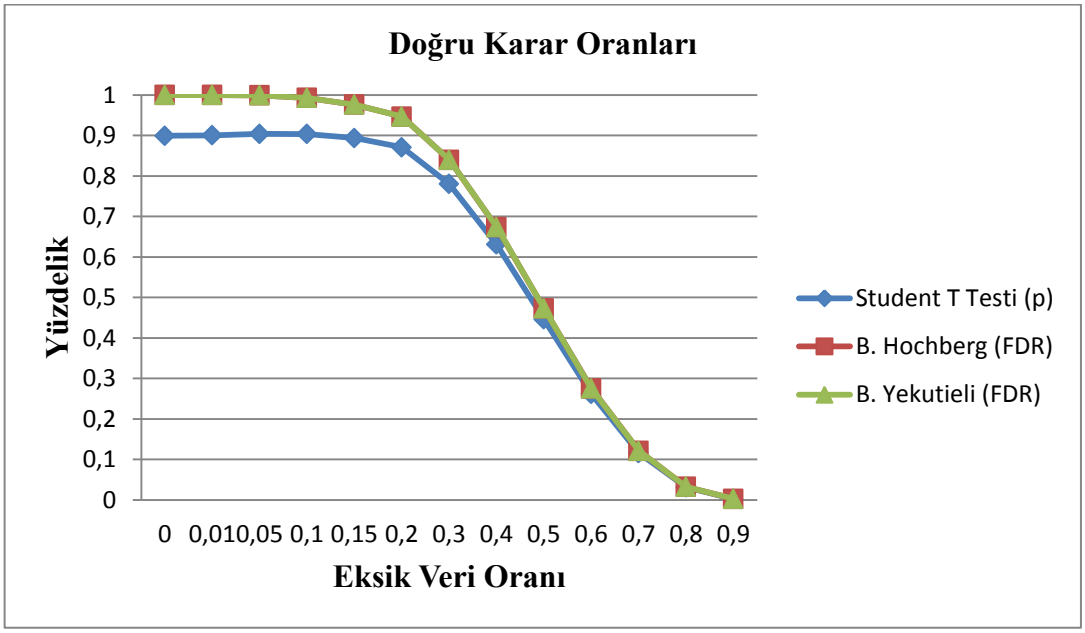


Şekil 4.7: Adenoma eksik veriler ile gerçek veri arasındaki 1. tip hata oranları farkı

Çizelge 4.8'e göre adenoma veri setinde anlamlı bulunmayan genlere ait Student t testi p değerlerinin ortalamasının ve Benjamini Hochberg ve Yekutieli prosedürlerinde de FDR değerlerinin her eksik oran için arttığı görülmektedir. Çizelge 4.9 ve Şekil 4.9'da gösterildiği gibi eksiklik oranına göre doğru karar oranlarına bakıldığında Student t testi için, %1, %5 ve %10'luk eksik veri durumunda gerçek veriye göre arttığı ve %20'den sonra ise azaldığı görülmektedir. %1'den sonraki eksiklik oranları için Benjamini Hochberg ve Yekutielli test sonuçlarında, anlamlı bulunan genlerin sayılarında azalma görülmekteir. Şekil 4.8'de Benjamini Hochberg ve Yekutielli prosedürlerinin birbirleri ile çok yakın sonuçlar verdiği görülmektedir. %40'lık eksik veri durumunda üç yöntemin doğru karar oranları bakımından çok yaklaştığı görülmektedir.

Çizelge 4.8: Adenoma anlamlı bulunmayan genlere ait bulgular

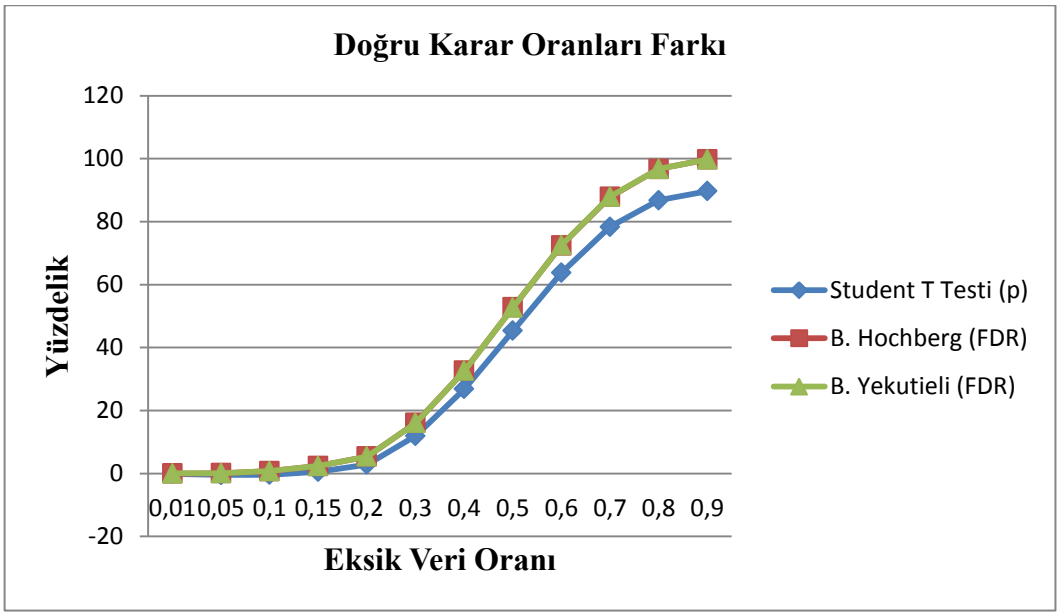
Adenoma	Ortalama±Standart Sapma			Ortalama±Standart Sapma (Anlamlı Olmayan Gen Sayısı)			Doğru Karar Oranı		
Eksik Veri Oranı	Student T Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)	Student t Testi	B.Hochberg	B.Yekutieli	Student t Testi	B.Hochberg	B.Yekutieli
%0	0,46623±0,26733	0,75160±0,17140	7,09650±1,62090	6371	7085	7086	89,91	99,99	100
%1	0,46677±0,00070	1,32050±0,00038	24,86920±0,00722	6379,2±5,0	7085,0±0,3	7085,90±0,2	90,03	99,99	99,99
%5	0,46879±0,00145	1,32426±0,00084	24,94262±0,01584	6405,0±10,8	7078,6±2,6	7079,1±2,6	90,39	99,90	99,90
%10	0,47164±0,00202	1,32892±0,00111	25,03053±0,02163	6401,3±15,1	7032,9±7,0	7033,9±7,2	90,34	99,25	99,27
%15	0,47436±0,00219	1,33344±0,00131	25,11527±0,02561	6332,1±18,4	6916,7±11,7	6917,8±11,9	89,36	97,61	97,63
%20	0,47730±0,00245	1,33763±0,00149	25,19689±0,02814	6169,7±22,7	6705,6±16,6	6705,5±17,0	87,07	94,63	94,63
%30	0,48240±0,00301	1,34557±0,00191	25,34350±0,03519	5527,9±28,8	5949,7±25,1	5950,5±24,4	78,01	83,96	83,98
%40	0,48685±0,00348	1,35231±0,00223	25,47569±0,04370	4470,5±30,7	4773,00±27,9	4774,2±28,2	63,09	67,36	67,38
%50	0,49075±0,00451	1,35851±0,00282	25,58970±0,05369	3160,4±30,2	3348,8±28,5	3350,7±29,3	44,60	47,26	47,29
%60	0,49441±0,00597	1,36387±0,00393	25,68712±0,07179	1853,90±27,3	1952,5±26,3	1950,0±26,5	26,16	27,55	27,52
%70	0,49754±0,00869	1,36847±0,00573	25,77037±0,10893	820,1±21,4	859,2±21,7	859,1±21,9	11,57	12,13	12,12
%80	0,49911±0,01787	1,37241±0,01095	25,85540±0,20062	221,9±13,2	231,2±13,3	231,6±13,8	3,13	3,26	3,27
%90	0,50411±0,06067	1,37768±0,03712	25,91409±0,71774	18,7±4,2	19,3±4,3	19,6±4,5	0,26	0,27	0,28



Şekil 4.8: Adenoma doğru karar oranları

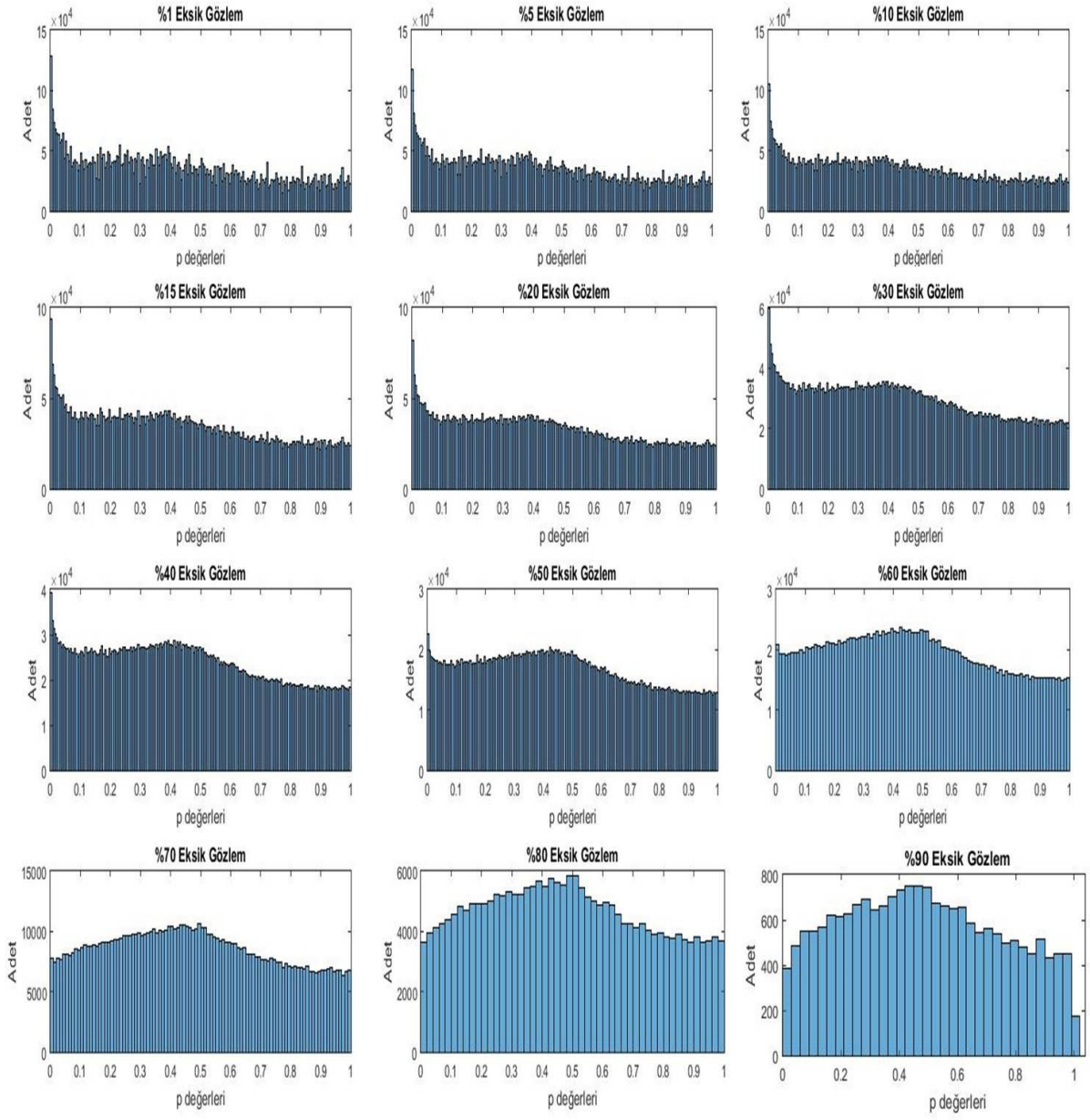
Çizelge 4.9: Adenomaeksik veriler ile gerçek veri arasındaki doğru karar oranları farkı

Eksik Veri Oranı	Student T Testi (p)	B.Hochberg (FDR)	B.Yekutieli (FDR)
%1	-0,12	0	0,01
%5	-0,48	0,09	0,1
%10	-0,43	0,74	0,73
%15	0,55	2,38	2,37
%20	2,84	5,36	5,37
%30	11,9	16,03	16,02
%40	26,82	32,63	32,62
%50	45,31	52,73	52,71
%60	63,75	72,44	72,48
%70	78,34	87,86	87,88
%80	86,78	96,73	96,73
%90	89,65	99,72	99,72



Şekil 4.9: Adenoma eksik veriler ile gerçek veri arasındaki doğru karar oranları farkı

Adenoma veri setinde Student t testi sonucu hesaplanan bütün p değerlerine ait dağılımda eksiklik oranının artması ile hesaplanan p değerlerinin sayılarında azalma olduğu ve %15 eksik veri durumundan sonra dağılımın şeklinin değiştiği Şekil 4.10'da görülmektedir.



Şekil 4.10: Adenoma Student t testine ait p değerlerinin dağılımı

5. TARTIŞMA

Son yıllarda tıp alanında, genetik faktörlerin hastalıkların üzerine ilişkisini incelemek amacıyla yapılan çalışmalarda artış gözlenmektedir. Bu amaç doğrultusunda mikrodizi teknolojisi geliştirilmiştir.

Literatür incelendiğinde Ally R. ve arkadaşlarının yapmış olduğu mikrodizi çalışmasında eksik verilerin varlığının 1. tip hata oranlarında ve testin gücünde etkisinin olduğunu açıkça belirtmiştir (35). Literatürdeki diğer çalışmalara bakıldığında eksik verilerin ataması için uygulanan ve geliştirilen yöntemlere değinilmiştir. Bu yöntemler genellikle %20'ye kadar olan eksiklikleri içerdiğinden, 1. tip hatası yüksek ve testin gücücün oldukça azaldığı durumlar için uygulamalar yetersiz kalmaktadır.

Bu nedenle bu çalışmada, mikrodizi veri setlerinde çok sık rastlanan ve farklı oranlarda ortaya çıkan eksik veri durumunun, testin sonucunu ve 1. Tip hataları nasıl etkilediği, istatistiksel yöntemlerden hangisinin daha uygun sonuçlar verdiği incelenmiştir.

Mikrodizi veri setleri, aynı bireyden çok sayıda genin kullanılması ve dolayısıyla ifade düzeylerinin birbirleriyle doğal ilişkili olduğu bilinmektedir. Bu sebepten Daniel Yekutieli and Yoav Benjamini yapmış oldukları çalışmada genler arası bağımlılık düşüncesine dayanarak Benjamini Yekutieli prosedürünün kullanılmasını önermektelerdir (48). Çizelge 4.2'ye bakıldığında literatür ile paralel olarak Benjamini Yekutieli prosedürü gerçek veri setinde Student t testi ve Benjamini Hochberg prosedürlerine göre beklenen 1. Tip hata oranını çok daha küçük hesaplamış ve beklenen %51. Tip hata oranın daha yakın sonuç verdiği görülmüştür.

Yongchao Ge ve arkadaşlarının yapmış olduğu çalışmada, tipik bir mikrodizi deneyi, aynı zamanda genlerin binlerce ifadelerini ölçtüğünden dolayı, çoklu test sorunu ile karşı karşıya kalınmasından bahsetmiştir. Çoklu test sorununda 1. Tip hata oranlarının arttığı ve öncelikle 1. Tip hata oranının belirlenmesi, sonrasında bu oranların kontrol altına alınması gerektiğinden bahsetmiştir. Çalışmalarında FDR değerlerinin kullanımının FWER gibi katı bir değerden daha avantajlı olduğundan bahsetmişledir (52).Literatürde de bahsedildiği gibi iki grup karşılaştırmalarında birbirinden bağımsız yapılan test istatistikleri sonucunda $(1-(1-p)^c$ denkleminde, c deneme sayısı, p 1. Tip hata olmak üzere hesaplanan yeni 1. Tip hata oranlarında şişme meydana

gelmektedir(53). D. Ovla ve arkadaşlarının yapmış oldukları SNP analizlerinde çokluk nedeniyle artan 1. Tip hatanın düzeltilmesine yönelik çalışmada Benjamini Yekutieli prosedürünün en uygun yöntem olduğu sonucuna ulaşmışlardır (12).Literatür ile bağlantılı olarak çalışmamızda, Carcinoma veri setinde eksik gözlemin olmadığı gerçek veri durumunda Student t testine göre 1. Tip hata oranı 23,87 olarak ve Adenoma veri setinde ise 1. Tip hata oranı 10,09 olarak hesaplanmıştır. 1. Tip hatayı kontrol altına almak için geliştirilen Benjamini Hochberg ve Benjamini Yekutieli prosedürlerine gerçek veri seti üzerinde bakıldığında, Yekutieli prosedürü beklenen %5 1. Tip hata seviyesini daha fazla korumaktadır. Benjamini Hochberg prosedürü bu aşamada yetersiz kalmaktadır. Mikrodizi veri setleri için geliştirilmiş olan mikrodizi anlamlılık testi 1. Tip hata oranını %4,7 olarak hesaplamıştır ve diğer bütün yöntemlere göre beklenen %5 1. Tip hata seviyesine en yakın sonucu vermektedir. Student t testi eksik veri durumu olmamasına rağmen her iki veri seti için de 1. Tip hataları diğer yöntemlere göre oldukça büyük hesaplamaktadır. Student t testini sonuçlarını kontrol altına almak için geliştirilen Benjamini Yekutieli ve Hochberg prosedürleri ve Student t testinin mikrodiziler için geliştirilmiş bir versiyonu olan mikrodizi anlamlılık testi, bu yanılgıya rağmen literatürde karşılaştığımız sonuçlara göre oldukça az kullanılmaktadır.

Literatürde mikrodizi çalışmaları, genellikle birey sayısının az, gen sayısının çok olduğu veri setleri ile karşımıza çıkmaktadır. Ve bu tür çalışmalarda asıl amaç hasta ve kontrol gruplarını ilgili genler bakımından karşılaştırmaktır. Fakat gen sayısı ile örneklem genişliği arasındaki ($n<p$) dengesizlikten kaynaklı olarak genlerin karşılaştırılmasında kullanılan Hotelling T^2 testi kullanılamamaktadır. n_1 hasta grubuna ait birey sayısı, n_2 kontrol grubuna ait sayısı ve m karşılaştırılmaların yapılacağı genlerin sayısı olmak üzere serbestlik derecesi (n_1+n_2-m) olarak hesaplanmaktadır. Carcinoma verisi için hesaplanan serbestlik derecesi 18+18-7457=-7421ve Adenoma veri seti için hesaplanan serbestlik derecesi 4+4-7086=-7078 olarak hesaplanacaktır. Dolayısıyla F istatistiğine ait serbestlik derecesinin negatif çıkmasından kaynaklı olarak klasik Hotelling T^2 testi kullanılamamıştır.

Bizim çalışmamızda, Carcinoma verisi nispeten Adenoma verisine göre daha büyük örneklem büyüklüğüne sahiptir. Adenoma verisine karşılık daha büyük örnek genişliklerinde çalışılan Carcinoma verisi kontrol amaçlı alınmıştır. Her iki veri seti

(küçük ve büyük örnek genişliği) eksik veri durumu, kullanılan test türü ve 1. Tip hata bakımından değerlendirilmiştir.

Her iki veri setinde anlamlı bulunan genlere ait bulgular karşılaştırıldığında (Çizelge 4.2 ve Çizelge 4.6) Carcinoma veri setinde Benjamini Yekutieli prosedürüne göre 321 anlamlı gen hesaplanabilirken, Adenoma veri setinde hiçbir anlamlı gen hesaplayamamıştır. Bu sonuçtan Benjamini Yekutieli prosedürünün örneklem büyüklüğünden etkilendiği düşünülmektedir. Student t testine göre hesaplanan 1. Tip hata seviyesi küçülmesine rağmen hala beklenen %5 1. Tip hata oranından oldukça uzaktır. Benjamini Hochberg prosedürü 1. Tip hataları gerçek veri setinde %12,03 olarak hesaplarken, küçük veri setinde %0,014 olarak hesaplamıştır. İki veri seti karşılaştırıldığında Hochberg prosedürü 1. Tip hata oranlarını büyük veri setinin %90 eksik veri oranında, küçük veri setinin gerçek veri oranına göre daha büyük hesaplamıştır. Student t testi sonuçlarına göre büyük örneklem genişliğine sahip veri setinde hesaplanan p değerleri ortalamaları eksiklik yapısı arttıkça sürekli bir artış içerisindedir. Ancak küçük örnek genişliğine sahip veri setinde p değerlerinin ortalaması %90 eksik veri yapısına kadar sürekli artmakta ancak %90'da veri yapısındaki aşırı bozulmadan, hesaplanamayan test istatistiklerinden ve anlamlı bulunurken anlamsız hale dönüşen p değerlerinin sayısı oldukça fazla olduğundan dolayı p değerlerinin ortalaması düşmüştür. Carcinoma veri setinde SAM analizi %50 eksik veri yapısına kadar hesaplama yapabilirken, Adenoma verisinde %10 eksik veri yapısına kadar hesaplama yapabilmektedir. Bu durumda SAM analizinin eksik veriden daha az etkilendiğinin göstergesi olarak düşünülmektedir. Carcinoma veri setinde eksik verinin olmadığı durumda Benjamini Yekutieli prosedürü ve SAM analizinin %5'lik1. Tip hata oranını koruduğu gözlenmiştir. Eksik veri oranı %1 olduğunda sadece SAM analizi %5'lik seviyeyi korumaktadır ve %20 eksik veri yapısında iken Benjamini Yekutieli prosedürünün %1 eksik veri durumunda gösterdiği 1. Tip hata seviyesine yakın sonuçlar vermektedir. Eksik veri oranı artırıldığında tüm yöntemlerin 1.Tip hata seviyesini koruyamadığı gözlenmiştir. Adenoma veri setinde %5'lik seviyenin hiçbir durumda korunamadığı görülmüştür. Bu durumda veri setindeki örneklem büyüklüğünün küçük olmasından kaynaklandığı düşünülmektedir.

Her iki veri seti gerçek veri ile eksik veriler arasındaki 1.tip hata oranları farkı bakımından karşılaştırıldığında (Çizelge 4.3 ve Çizelge 4.7, Şekil 4.2 ve Şekil 4.7)Eksik

veri oranı %20 ve üzerine çıktığında tüm yöntemler 1. Tip hata bakımından etkilenmekle birlikte, Student t testinin en çok etkilenen test olduğu görülmektedir. Student t testi sonucunda hesaplanan p değerlerinde ait dağılımların bulunduğu Şekil 4.10 incelendiğinde eksik veri oranının %15 ve üzerinde olması durumunda p değerlerine ait dağılımların oldukça bozulduğu görülmektedir. Acuña ve arkadaşlarının yapmış olduğu çalışmada eksik veri oranın %15'in üzerinde olmasının çalışmanın sonuçlarını ciddi derecede etkileyebileceğini söylemiştir. Bu sonuçlar, Acuña ve arkadaşlarının çalışması ile paralellik göstermektedir(15).

Her iki veri setinde anlamlı bulunmayan genlere ait bulgular karşılaştırıldığında(Çizelge 4.4 ve Çizelge 4.8) Carcinoma veri setinde eksik veri oranın %90 seviyesine çıktığında üç prosedürde doğru karar verme oranından çok ciddi etkilenmekte ve yaklaşık %30 doğru karar verme oranına sahip olmaktadır. Eksik veri oranının %90'nın altında olduğunda yaklaşık olarak % 80 ve üzerinde doğru karar verme oranına sahip olduğu görülmektedir. Adenoma veri setinde ise % 40 ve üzerinde eksik veri oranı var olduğunda üç prosedürün de çok ciddi etkilendiği gözlenmiştir. Bu durumda veri setindeki örneklem büyüklüğünden ve kayıp veri oranından üç yöntemin de etkilendiğinin göstergesi olarak düşünülmektedir. Carcinoma veri setinde anlamlı bulunan genlerin anlamsız hale dönüşmesinden dolayı %70 eksik veri oranına kadar hesaplanabilen test sonuçları her üç prosedüre göre artış göstermektedir. Ancak %70'den sonra hesaplanamayan gen sayıları azalmaktadır. Adenoma veri setine göre ise %5 eksik veri yapısına kadar hesaplanabilen gen sayılarında artış, sonrasında ise tekrar azalma gözlenmiştir. Özellikle %20 eksik veri oranı ve sonrasında hesaplanamayan gen sayıları oldukça hızlı azalmaya başlamıştır. Bu bulgular doğrultusunda küçük örnek genişliğine sahip mikrodizi verilerinin eksik veri oranından daha fazla etkilendiği düşünülebilir.

6. SONUÇ ve ÖNERİLER

Mikrodizi verilerinde Student t testi ile yapılan çok sayıdaki karşılaştırmanın p değerlerinde şişme meydana getirdiği ve %5 olarak beklenen 1. tip hata düzeyini yüksek hesapladığı gözlenmiştir. Carcinoma veri seti için Student t testine göre hesaplanan 1. tip hata değeri %23,87 iken, t testinin mikrodiziler için geliştirilmiş bir versiyonu olan mikrodizi anlamlılık testi (SAM) 1. tip hata oranını %4,7 olarak hesaplamıştır. %5 1. Tip hata düzeyine en yakın SAM testi olarak bulunmuştur. SAM analizini takiben Benjamini Yekutieli prosedürüne göre gerçek veride 1. Tip hata oranı %4,31 olarak hesaplanmıştır. Benjamini Hoshberg prosedüründe ise %12,03 olarak hesaplanan 1. Tip hata değeri Student t testine kıyasla daha düşüktür, fakat Yekutieli prosedürüne ve SAM analizine göre daha yüksek ve %5 beklenen 1. Tip hatadan oldukça büyük olarak hesaplanmıştır. Eksiklik oranlarının var olduğu durumlarda veri yapısı bozulmaktadır. Bu nedenle tüm prosedürlerde 1. tip hataların azalmasına rağmen, mikrodizi veri setlerinde eksik verinin varlığının test sonuçlarını yanlış değerlendirmeye neden olduğu görülmüştür. SAM analizinde %40 eksik veri yapısından sonra hiçbir hesaplama yapmaması yanlış sonuç verme durumuna karşı bu yöntemin koruyucu olduğunun bir göstergesidir ve %50 ve üzerinde eksik veri yapısına sahip setlerde SAM yönteminin kullanılması önerilmemektedir. SAM analizinde %5'lik eksik veri yapısından sonra hesaplanan 1. Tip hata değerleri eksiklikten daha fazla etkilendiğinden %5'ten daha fazla eksik veri yapısına sahip setlerde kullanılması önerilmemektedir. Yekutieli prosedüründe ise %1'lik eksiklik oranında SAM'e göre daha fazla etkilendiği ve beklenen 1. Tip hatadan uzaklaştığı ve dolayısıyla eksik veri durumunda kullanılması hatalı sonuçlar verecektir.

Carcinoma versinden daha küçük örnek genişliğine sahip Adenoma veri setinde Benjamini Yekutieli prosedürüne göre 1. Tip hata oranları hesaplanamamıştır. Benjamini Yekutieli prosedürü örnek genişliğinden daha çok etkilenmektedir. SAM analizi de dahil olmak üzere diğer üç prosedürün de beklenen 1.Tip hata seviyesini koruyamadığı görülmüştür. Ayrıca SAM analizi eksik veri oranın %10 ve üzerinde olduğu durumlarda 1. Tip hata oranını hesaplayamamıştır.

Mikrodizi çalışmasında eksik verilerin varlığının 1. tip hata oranlarında ve testin gücündeetkisinin olduğu bilinmektedir. Yöntemler genellikle %20'ye kadar olan

eksiklikleri içerdiğinden, 1. tip hatası yüksek ve testin gücücün oldukça azaldığı durumlar için uygulamalar yetersiz kalmaktadır. %20'ye kadar olan atama yöntemlerinin literatürde karşılaştırılması bulunduğundan, atama yöntemlerinden hangilerinin daha iyi olacağına dair çalışmalar çalışmamızda kullanılan %1, %5, %10, %15, %20, %30, %40, %50, %60, %70, %80 ve %90 oranları için uygulanacak ve eksiklik yapısının artmasıyla bu atama yöntemlerinin nasıl etkileneceğini tanımlayan bir çalışma yapılacaktır.

Çalışmada bütün eksiklik oranlarındaki 1000'e denemede, her bir gene ait gücün hesaplanmaması, eksiklik yapısının, genler (satırlar) ve bireyler (sütunlar) olarak ayrılmadan bütün veri seti için rasgele olarak üretilmesi, hastanemizde mikrodizi teknolojisinin bulunmamamsı ve SAM analizi sonuçlarında sadece anlmalı genlere ulaşılabilmesi çalışmanın kısıtlılıklarını oluşturmaktadır.

7. KAYNAKLAR

1) **Zikopoulos P, Eaton C, Deroos D, Deutsch T, Lapis G.***Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data.*Cenveo Publisher Services, **2012**

2) **Schadt EE**, The Changing Privacy Landscape in the Era of Big Data. *Molecular Systems Biology*, **2012**; 8:612

3) **Wullianallur R, Wullianallur V**. Big Data Analytics in Healthcare: Promise and Potential.*Raghupathi and Raghupathi Health Information Science and Systems*, **2014**; 2:3

4) **Marx V**, The Big Challenges Of Big Data. *Nature*, **2013**; 498: 255-260

5) **Jianqing F, Fang H, Han J**. Challenges of Big Data Analysis.*National Science Review*, **2014;**1: 293–314

6) **Bal SH, Budak F**, Mikroarray Teknolojisi.*Uludağ Üniversitesi Tıp Fakültesi Dergisi* **2012;** 38 (3): 227-233

7) **Babur Ö,** Causality Analysis In Biological Networks. Doktora Tezi, Bilkent Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü, Ankara, **2010**

8) **Yoltaş A, Karaboz İ**, DNA Mikroarray Teknolojisi ve Uygulama Alanları.*Elektronik Mikrobiyoloji Dergisi*, **2010**; 8(1): 01-19

9) **Hourani M,El Emary IMM**. Microarray Missing Values Imputation Methods: Critical Analysis Review. *ComSIS,***2009**; 6:2

10) **Liew AWC, HongYan NFL**. Missing Valueimputation For Gene Expression Data: Computational Technique Store Cover Missing Data From Available Information. *Briefings In Bioinformatics*, **2010**; 6:16

11) **Aittokallio T.**Dealing With Missing Valuesin Large-Scale Studies: Microarraydata Imputation And Beyond.*Briefings In Bioinformatics*, **2009**; 1:12

12) **Ovla HD, Taşdelen B, Öztornacı RO.***SNP Analizlerinde Çokluk Nedeniyle Artan Tip I Hata Oranının Düzeltilmesine Yönelik Yaklaşımların İncelenmesi*, XVII. Ulusal Biyoistatistik Kongresi, **2015**, Girne, KKTC

13) **Johansson AM**, Methodology for Handling Missing Data in Nonlinear Mixed Effects Modelling. Upsala Üniversitesi, **2014**

14) **Köse İA, Öztemur B**; Kayıp Veri Ele Alma Yöntemlerinin T-Testi Ve Anova Parametreleri Üzerine Etkisinin İncelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*,**2014**; 14:1

15) **Acuña E, Rodriguez C**, The Treatment of Missing Values and its Effect on Classifier Accuracy,http://link.springer.com/chapter/10.1007/978-3-642-17103-1_60. Erişim Tarihi 15.09.2015

16) **Demir E, Parlak B**, Türkiye'de Eğitim Araştırmalarında Kayıp Veri Sorunu. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, **2012**, 3(1), 230-241

17) **Yılmaz H**, Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi Ve Sağlık Alanında Bir Uygulama. Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir, **2014**

18) **Bennett DA**, How can I deal with missing data in my study.*Australian And New Zealand Journal Of Public Health*, 2001; 25(5)

19) **Cheng-Han J**, A Probability Based Framework for Testing the Missing Data Mechanism. UCLA Electronic Thesis and Dissertations, **2013**

20) **Klug WS, Cummings MR**, *Genetik Kavramlar*. 6. Baskı, Ankara: Palme Yayıncılık, **2003**.

21) **Reece RJ**, *Analysis of Genes and Genomes*. England: John Wiley & Sons ltd, **2004.**

22) **King RC, Stansfield WD, Mulligan PK,***A Dictionary of Genetic.*7. Baskı, Oxford University Press, **2006**

23) **Erdemir F, Uysal G**,Genetik, Genomik Bilimi ve Hemşirelik.*Dokuz Eylül Üniversitesi Hemşirelik Yüksekokulu Elektronik* Dergisi,**2010;** 3(2),96-101

24) **Savlı H**, Dizilim Teknolojisi Çiplerden Sonsuzluğa. *Türkiye Klinikleri***2004**;24(5):534-40

25) **Babu MM**, An Introduction to Microarray Data Analysis. http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf. Erişim Tarihi 05.09.2015

26) **Passarge E**, *Renkli Genetik Atlası*. 3. Baskı, İstanbul: Nobel Tıp Kitabevleri ltd, **2009**.

27) International HapMap Project. http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en. Erişim Tarihi 15.09.2015

28) International HapMap Project. http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html.en. Erişim Tarihi 12.10.2015

29)http://compbio.pbworks.com/w/page/16252906/Microarray%20Normalization%20and%20Expression%20Index#DyeBias Erişim Tarihi23.11.2015

30)**Rulli S.**Introduction To Real-Time Quantitative PCR (qPCR). SABiosciences,*A QIAGEN Company*

31)**Smith GK, Spped T**.Normalization of cDNA Microarray Data. *Methods* **2003**;31: 265-273.

32) **Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, AltmanRB.** Missing Value Estimation Methods For DNA Microarrays.*Bioinformatics*,**2001***;* 17(6)

33) **Kim H, Golub GH, Park H**. Missing Value Estimation For DNA Microarray Gene Expression Data: Local Least Squares Imputation.*Bioinformatics*,2005; 21(2) 187–198.

34) **Bù TH, Dysvik B, Jonassen I**. Lsimpute: Accurate Estimation Of Missing Values in Microarray Data With Least Squares Methods.*Nucleic Acids Research*, **2004**; 32:3

35) **Rogers A, Beck A, Tintle NL**. *Evaluating the Concordance Between Sequencing, Imputation and Microarray Genotype Calls in the GAW18 Data*. Genetic Analysis Workshop 18 Stevenson, WA, USA. October **2012**; 13-17

36) **Sheng Q, Moreau Y, Smet FD, Marchal K, Moor BD**. Cluster analysis of microarray data. ftp://ftp.esat.kuleuven.be/sista/sheng/reports/cluster_long.pdf. Erişim Tarihi 30.10.2015

37) **Chiu CC, Chan SY, Wang CC, Wu WS.**Missing Value Imputation For Microarray Data: A Comprehensive Comparison Study and a Web Tool.*BMC Systems Biology* **2013;** 7: 6-12

38) **Bras LP, Menezes JC.**Improving Cluster Based Missing Value Estimation of Microarray Data.*Biomoleculer Enginnering*, **2007**; 24:273-282

39) **Coşgun E, Karaağaoğlu E.**Veri Madencili¤İ Yöntemleriyle Mikrodizilim Gen İfade Analizi.*Hacettepe Tıp Dergisi* 2011; 42:180-189

40) **Brevern AG, Hazout S, Malpertuy A**. Influence Of Microarrays Experiments Missing Values On The Stability Of Gene Groups By Hierarchical Clustering. *BMC Bioinformatics***2004;** 5:114

41) **Sehgal MSB, Gondal I, Dooley LS**. Collateral Missing Value Imputation: A New Robust Missing Value Estimation Algorithm For Microarray Data. *Bioinformatics,* **2005;**21:10

42) **Stephenson WR**. Two Independent Samples. http://www.public.iastate.edu/~wrstephe/stat496/twosample.pdf Erişim Tarihi 15.10.2015

43) Probability and Statistics Ebook. http://wiki.stat.ucla.edu/socr/index.php/AP_ Statistics_Curriculum_2007_Infer_2Means_Indep. Erişim Tarihi 16.10.2015

44) **Dawson B, Trapp R**. *Basic and Clinical Biostatistic*. 4. Edition, The McGraw-Hill Companies, **2004**

45) http://compbio.uthsc.edu/microarray/lecture2.html Erişim tarihi 15.10.2015

46) **Tusher VG, Tibshirani R, Chu G**. Significance Analysis of Microarrays Applied to The Ionizing Radiation Response.*PNAS* **2001**; 98:9

47) **Schwender H, Krause A, Ickstadt K**. Comparison of the Empirical Bayes and the Significance Analysis of Microarrays.*Universitätsbibliothek Dortmund*, **2003**

48) http://csbl.bmb.uga.edu/mirrors/JLU/DragonStar2014/download/pc/SAM.ppt Erişim Tarihi 15.10.2015

49) **Benjamini Y, Yekutieli D.**The Control Of The False Discovery Rate In Multiple Testing Under Dependency.*The Annals of Statistics***2001**; 29(4): 1165-1188

50) **Rajablı F.** Aile Temelli İlişkilendirme Çalışmalarında Sınırlı Örneklem Boyutu İçin İstatistiksel Sinyal İşleme Algoritmalarının Kullanılması. Doktora Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, **2011.**

51) **Borg K.**False Discovery Rate and q Value.*Technische University a Dortmund***2012**.

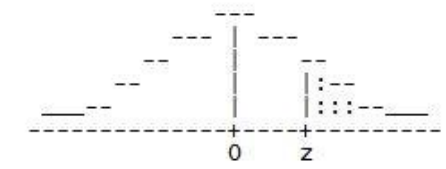
52) **Stevens J.**Applied Multivariate Statistics fort the Social Science. *Lawrence Erlbaum Associates*, **1986.**

53)**Ge Y, Sealfon SC, Speed TP.** Multiple testing and its applications to microarrays.*Statistical Methods in Medical Research***2009**, 18(6) 543-563.

EKLER

Ek 1: Normal Dağılım Tablosu

NORMAL DAĞILIM EĞRİSİ ALTINDA KALAN ALAN Z-TABLOSU

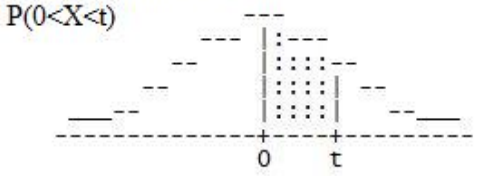


z	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Ek 2: T Dağılımı Tablosu

T-DAĞILIMI T-TABLOSU

P(0<X<t)



DF	t(0.4)	t(0.45)	t(0.475)	t(0.49)	t(0.495)	t(0.4975)
1	3.078	6.314	12.706	31.821	63.657	127.321
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
11	1.363	1.796	2.201	2.718	3.106	3.497
12	1.356	1.782	2.179	2.681	3.055	3.428
13	1.350	1.771	2.160	2.650	3.012	3.372
14	1.345	1.761	2.145	2.624	2.977	3.326
15	1.341	1.753	2.131	2.602	2.947	3.286
16	1.337	1.746	2.120	2.583	2.921	3.252
17	1.333	1.740	2.110	2.567	2.898	3.222
18	1.330	1.734	2.101	2.552	2.878	3.197
19	1.328	1.729	2.093	2.539	2.861	3.174
20	1.325	1.725	2.086	2.528	2.845	3.153
21	1.323	1.721	2.080	2.518	2.831	3.135
22	1.321	1.717	2.074	2.508	2.819	3.119
23	1.319	1.714	2.069	2.500	2.807	3.104
24	1.318	1.711	2.064	2.492	2.797	3.091
25	1.316	1.708	2.060	2.485	2.787	3.078
26	1.315	1.706	2.056	2.479	2.779	3.067
27	1.314	1.703	2.052	2.473	2.771	3.057
28	1.313	1.701	2.048	2.467	2.763	3.047
29	1.311	1.699	2.045	2.462	2.756	3.038
30	1.310	1.697	2.042	2.457	2.750	3.030
31	1.309	1.696	2.040	2.453	2.744	3.022
32	1.309	1.694	2.037	2.449	2.738	3.015
33	1.308	1.692	2.035	2.445	2.733	3.008
34	1.307	1.691	2.032	2.441	2.728	3.002
35	1.306	1.690	2.030	2.438	2.724	2.996
40	1.303	1.684	2.021	2.423	2.704	2.971
50	1.299	1.676	2.009	2.403	2.678	2.937
60	1.296	1.671	2.000	2.390	2.660	2.915
70	1.294	1.667	1.994	2.381	2.648	2.899
80	1.292	1.664	1.990	2.374	2.639	2.887
90	1.291	1.662	1.987	2.368	2.632	2.878
100	1.290	1.660	1.984	2.364	2.626	2.871
110	1.289	1.659	1.982	2.361	2.621	2.865
120	1.289	1.658	1.980	2.358	2.617	2.860
Z	1.282	1.645	1.960	2.326	2.576	2.807

ÖZGEÇMİŞ

10 Nisan 1988 tarihinde Mersin'de doğdu. İlköğretim ve lise eğitimini Mersin'de tamamladıktan sonra lisans eğitimini Adana Çukurova Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü'nde 2012 yılında tamamladı. 2013 yılında Mersin Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı'nda lisansüstü eğitimine başladı.