Sosyal Bilimler R Platformu

Burak AYDIN, James ALGINA, Walter LEITE, Hakan ATILGAN

2017

Contents

Chapter 1

Kapak

1.1 Tanıtım

Bu materyal İngilizce olarak hazırlanıp Türkçeye çevirilmiştir. Bu *platform* sosyal bilimler alanında çalışan ve nicel veri analizlerinin teoriden ziyade uygulama aşamasına ilgi gösteren araştırmacılar için oluşturulmuştur. Bütün istatistiksel prosedürler R (R Core Team, 2016b) ile yürütülmüş, gerçek veri kullanımına özen gösterilmiştir. Bu materyale *platform* denilmesinin üç sebebi vardır, (a) katkıya açıktır,(b) dinamik bir içeriğe sahiptir, (c) bilgisayar anakartı gibi kullanılabilir, R ile oluşturulmuş herhangi bir üst düzey çıktı platforma eklenebilir. Bu materyal Bookdown (Xie, 2016) ile inşa edilmiştir, Bookdown ise R Markdown (Allaire et al., 2016) üzerine inşa edilmiştir. Materyalin hazırlanma aşamasında R Studio (RStudio Team, 2016) kullanılmıştır.

1.1.1 Atıf

Bu materyalin uygun atıfı:

Aydın, B., Algina, J., Leite, W. L., & Atılgan, H. (2018). *Sosyal Bilimler İçin R' a Giriş.* Ankara: ANI Yayıncılık.

1.1.2 Neden Bookdown?

Bookdown ile görsel zenginliği mevcut materyaller oluşturulabilir. Shiny uygulamaları, basit olmayan grafikler gibi R'ın araştırmacılara sunduğu ve sunacağı teknolojiler bu materyale kolayca eklenebilir. Bookdown ile aynı materyal PDF, HTML veya EBOOK olarak farklı şekilde kolayca okunabilir. Bookdown ile yazılan bir kitap Git Hub içerisinde depolanabilir ve en önemlisi katkı sağlamak isteyen araştırmacıların kolaylık sağlar. Kısacası Bookdown yeni nesil kitap yazma araçlarından biridir.

Figure 1.1:

1.1.3 İçerik

Platformun bu versiyonunda yer alan konular;

- Windows için R kurulumu
- R'a giRiş
- Veri Setleri
- Betimsel analizler ve hipotez testi
- t-test
- Varyans analizine giriş
- Korelasyon
- Çoklu regresyona giriş

Kitabın içeriği bir ders kitabından ziyade yardımcı materyal olarak hazırlanmıştır. İçerik az ve öz ele alınmıştır. Eğer açıklamaların yetersiz olduğunu düşünüyorsanız ve katkı sağlamak istiyorsanız isminize Teşekkürler 2.2 kısmında yer verilecektir.

Chapter 2

Önyüz

Bu kitabın yazarları olarak elimizde olan notları bir araya getirmek istedik. Akademik özenle oluşturduğumuz bu içerik her zaman açık kaynak olarak kalacaktır. Katkılarınız, istekleriniz dikkate alınacak ve içeriğe dahil edilecektir. Bu içeriğin hangi motivasyonlarla hazırlandığını merak eden okuyucular kitabın sonuna eklediğimiz [1] proje önerisini okuyabilirler.

2.1 Yazarlar

Akademik çalışmalarımız araştırma tasarıları (research design) ve nicel veri analizine yoğunlaşmıştır. Monte Carlo simulasyonları mutlaka çalışma takvimimizde yer alır. Diğer bir ortak noktamız çok düzeyli modeller üzerine yaptığımız çalışmalardır.

2.1.1 Burak Aydın, Ph.D.

Eğitimde araştırma ve değerlendirme metotları doktora ve istatistik doktora yandal derecesi mevcuttur. 2010 yılından bu yana R kullanıcısıdır. Yapısal eşitlik modelleri, çok düzeyli modeller ve eğilim puanları üzerine akademik çalışmaları vardır. Detaylı bilgi için Kişisel or Kurumsal

2.1.2 James Algina, Ph.D.

Klasik ve Modern Test Teorisi kitabının yazarıdır. Amerikan Eğitim Araştırmaları Birliği ve Amerikan Psikoloji Birliği (APA) üyesidir. 100'den fazla saygın akademik çalışması mevcuttur.Detaylı bilgi için UF Anita Zucker Center

2.1.3 Walter L. Leite, Ph.D.

Florida Üniversitesi Eğitim Fakültesinde Doçenttir. Gerçek deneysel ve yarı deneysel çalışmalardan edinilen verilerin analizinde saygın çalışmaları mevcuttur. Detaylı bilgi için UF College of Education

[1] sadece Türkçe versiyonunda yer alır

2.1.4 Hakan Atılgan, Ph.D.

Ege Üniversitesi Eğitim Fakültesinde Ölçme ve Değerlendirme doçentidir. Yapısal eşitlik modelleri, varyans analizine dayalı güvenirlik belirleme yöntemleri ve psikometri üzerine çalışmaktadır. Yaklaşık 15 yıldır lisansüstü düzeyde istatistik dersleri vermektedir.Detaylı bilgi için Ege Üniversitesi

2.2 Teşekkür

Burada katkı sağlayanların isimlerine yer verilecektir.

2.3 Data

Veri kullanıma açıktır, Dünya Bankası ve İŞKUR tarafından toplanmıştır. Türkiye'de yaşayan, İŞKUR'dan mesleki eğitim talebinde bulunmuş bireyleri temsil eder. 5902 kişi içerir. Materyal içerisinde kullanılan veriye *dataWBT* (data WorldBank Türkiye) ismi verilmiştir ulaşmak için Bölüm 6.1.4 içerisinde yer alan basamaklar takip edilebilir.

dataWBT içerisinde yer alan değişkenler;

1. Id: katılımcı kimliği
2. Program: Mesleki eğitim aldı mı? 1=evet, 2=Hayır
3. Cinsiyet: Erkek, Kadın
4. Kurs: 51 farklı kurs, muhasebeden garsonluğa.
5. Şehir: Katılımcıların yaşadığı şehir
6. Eğitim: en yüksek diploma
7. Babanın eğitim durumu
8. Annenin eğitim durumu
9. Soru 1-6: Toplumsal cinsiyet algısı (4lü likert). Yüksek puan cinsiyet ayrımına işaret eder.
10. Yüksek öğretim durumu: 0=Lise veya altı diplomalı , 1= Yüksek öğretim diplomalı 11: Yaş : 2010 yılında katılımcı yaşları 12: Hane geliri: TL olarak hane geliri 13: Hanede yaşayan kişi sayısı 14: Kişi başı yıllık gelir (hane geliri/hanede yaşayan kişi sayısı) 15: Toplumsal cinsiyet algısı genel puanı: 2,3,4,5 ve 6. soruların ortalaması. 16: Gelir kaynakları (12 farklı kaynak)

Toplumsal cinsiyet algısı puanları ve kullanılışı hakkında detaylı bilgi için Gök ve Aydın (basımda) incelenebilir.

Dünya Bankası Tarafından Katılımcılara Sorulan Sorular

1. Evlendikten sonra hem kadın hem erkek hane gelirine katkıda bulunmalıdır (çift-gelirlilik).
2. Üniversite eğitimi kızlardan ziyade erkekler için önemlidir.
3. Maddi durum zorlamadığı sürece evlı bir kadın evinin dışında çalışmamalıdır.
4. Eşinin çalışması bir erkek için onur kırıcıdır.
5. Bir kadın düşüncelerini evde söyleyebilir fakat dışarda asla
6. Bir kadın eşinin sözünden çıkmamalıdır.

2.4 Finansal Destek

Bu materyalin hazırlanması Recep Tayyip Erdoğan Üniversitesi BAP birimi tarafından desteklenmiştir. BAP-53005-601

Öncesinde, bu proje TUBİTAK tarafından Şubat 2016 tarihinde ret edilmiştir.ID 1059B191501734.

Chapter 3

R'ın popülerliği

R programının kullanım sıklığı sürekli artmaktadır. Tippmann (2015) Scopus veri tabanında taranan ve 2014 yılında basılmış her 100 makaleden 1 tanesinde R programına veya R paketlerine atıfta bulunulduğunu yazmıştır. 2014 yılında 2925 R paketi mevcuttur. 2016 yılı sonunda bu sayı 10000'e ulaşmıştır.

Data analizi ve istatistiksel modellemenin yanında, işlevsel grafikler çizme, dokümanlar oluşturma, sunum hazırlama ve simülasyon üretme gibi çeşitli amaçlar için kullanılabilen R, başta istatistikçiler olmak üzere, mühendisler, ekonometristler ve sosyal bilimlerde modern ve komplike modellerle çalışan araştırmacıların dikkatini çekmiştir.R programının yaygınlığı hakkındaki diğer göstergelere örnekler;

1. R veri analizikonusunda evrensel bir dil olmuştur ve yeni metotlar çoğu zaman R ile sunulur (Muenchen, 2011).
2. Elektrik ve elektronik mühendisleri enstitüsü (IEEE) üyeleri tarafından en çok kullanılan 5 programlama dilinden biridir. see
3. R programlama dilini lisans ve yüksek lisans düzeyinde ders olarak sunan üniversiteler ve uzaktan eğitim kurumları mevcuttur.
4. Özel şirketler tarafından kullanılan bir programlama dilidir.

Chapter 4

Windows için R kurulumu

R kurulumu oldukça kolaydır. R-project websitesinde yer alan basamaklar takip edilebilir veya sessiz olarak kaydedilmiş video izlenebilir (Video1 **??**).

R programını betik dosyası oluşturmadan doğrudan kullanmak mümkündğr fakat bir betik düzenleyici kullanmak kolaylık sağlar. En basit betik düzenleyicisi R programının içeresinde yer alır. R açık iken *File* ve sonrasında *new script* seçilerek betik düzenleyici açılır. Bu basamaklar Video2 **??**) ile gösterilmiştir.

Fakat R içerisinde yer alan betik düzenleyici çok basittir. Kullanışlılğı oldukça yüksek olan ve bu materyalin hazırlanmasında da kullanılmış olan betik dzenleyici R studio'dur. Kurulum basamakları Video3 **??** ile gösterilmiştir.

Chapter 5

Giriş

R veri analizi, grafik veya interaktif web uygulaması gibi basit olmayan çıktılar oluşturabilir. Bu bölümün amacı basit olmayan çıktılar oluşturmadan önce gereken temel prensipleri göstermektir.

5.1 Fonksiyonlar

R gibi programlanabilir gelişmiş hesap makineleri kullanıcıların fonksiyon yazmasına ve saklamasına izin verir. R kullanıcılarının fonksiyonların nasıl çalıştığını kavraması önemlidir.

5.1.1 R: Basit Hesap makinesi

R hesap yapabilir. Aşağıdaki işlemleri ve ilgili R kodlarını inceleyiniz.

$$1 + 1 = 2 \tag{5.1}$$

```
1+1
## [1] 2
```

$$1 - 1 = 0 \tag{5.2}$$

```
1-1
## [1] 0
```

$$1 + (2/3) - (2 * 6.5) = -11.33 \tag{5.3}$$

```
1 + (2 / 3) - (2 * 6.5)
## [1] -11.3
```

$$sin(30) + 4^3 + log(4) + e^3 + \sqrt{7} = 87.13 \tag{5.4}$$

```
sin(30) + 4^3 + log(4) + exp(3) + sqrt(7)
## [1] 87.1
```

Eşitlik (5.1) 'den (5.4) 'e kadar olan işlemler R tarafından tamamlanır fakat hafızada tutulmaz. Eğer yaptığınız bir işelmin sonucunu tekrar kullanmak istiyorsanız ona isim vermelisiniz. İsim verdiğiniz R çıktıları oturum süresince (session) tekrar erişime açıktır. Çıktılara oturum kapandıktan sonra da ulaşmak istiyorsanız kaydetmelisiniz. Kaydetme işlemleri ilerleyen bölümlerde ele alınmıştır. İsim verme işlemi fraklı şekillerde yapılabilir; "=", "<-" or "<<-". Bu materyal "=" operatörünü kullanır.

Eşitlik (5.1) 'den (5.4) 'e kadar olan işlemleri oturum süresince saklamak için;

```
a=1 - 1
b=1 + 1
c=1 + (2 / 3) - (2 * 6.5)
d=sin(30) + 4^3 + log(4) + exp(3) + sqrt(7)
```

İsim verdiğiniz çıktılar ile işlem yapabilirsiniz.

```
a+b+c+d
## [1] 77.8
```

İsim verdiğiniz bir çıktıyı değiştirebilirsiniz (overwrite)

```
e=3+2
e
## [1] 5
e=e+10
e
## [1] 15
```

Farklı bir isim vermek için (Not: R büyük harf küçük harf ayrımı yapar)

```
Equation1_output=a
Equation1_output + b + c + d  # a+b+c+d ile eşit
## [1] 77.8
```

5.1.2 R: Programlanabilir Hesap makinesi

En basit hali ile fonksiyon 3 parçadan oluşur, girdi, işlem, çıktı. Bu bölümde verilen fonksiyonların test puanlarının analiz basamağında kullanıldığını varsayalım.

5.1.2.1 Tek girdi - Tek Çıktı

Aşağıda verilen fonksiyonun adı *sabit5* olsun. *sabit5* fonksiyonu her öğrencinin puanına 5 puan ekleyecektir. Bir diğer deyişle, son derece basit bir fonksiyon olan *sabit5* verilen bir puana 5 ekleyerek çıktı oluşturur.

```
sabit5=function(girdi){
    cikti=girdi+5
    return(cikti)
    }

sabit5(girdi=50)
## [1] 55
sabit5(60)
## [1] 65
sabit5(80)
## [1] 85
```

sabit5 fonksiyonu girdiyi alır , 5 ekler *(input+5)*, ve bir çıktı oluşturur *(cikti=girdi+5)*, ve çıktıyı rapor eder *(return(cikti))*. Bütün bu işlemler *{ }* içinde verilmelidir.

Diğer basit bir fonksiyon *sistematik1* olarak isimlendirilmiştir ve verilen her bir puana %1 ekler.

```
systematic1=function(input){
    output=input+(input/100)
    return(output)
    }

systematic1(input=50)
## [1] 50.5
systematic1(100)
## [1] 101
systematic1(120)
## [1] 121
```

5.1.2.2 Çoklu Girdi-Tek Çıktı

Daha önce verilen fonksiyonlar tek girdi alıp tek çıktı oluşturmuştur. Bu örnekte iki farklı girdi ve tek bir çıktı vardır. Fonksiyona *eksipuan* adı verilmiştir. Ham puan ve yanlış sayısı verilen fonksiyon, her yanlış için 0.2 puan düşürür. Örneğin 90 puan ve 6 yanlış girildiğinde çıktı olarak (90-0.2*6) 88.8 verilir.

```
eksipuan=function(puan, yanlis){
    cikti=puan - (0.2 * yanlis)
    return(cikti)
    }
eksipuan(puan=90,yanlis=6)
## [1] 88.8
eksipuan(90,17)
## [1] 86.6
```

Bir R fonksiyonunda girdiler *argüman* (arguments) olarak isimlendirilir. *eksipuan* fonksiyonu 2 argümana sahiptir (puan ve yanlış) ve tek bir çıktı verir. Çoklu argüman ve çoklu çıktı içeren fonksiyonlar yazılabilir.

5.1.2.3 Çoklu Girdi ve Çoklu Çıktı

geridönüt fonksiyonu doğru yanıt sayısını ve her sorunun kaç puan olduğu argümanlarını alır, çıktı olarak toplam puanı ve 100 almak için eksik kalan soru sayısını hesaplar.

```
geridonut=function(dogruyanit, katsayi){
    total=dogruyanit*katsayi
    kalan=(100-total)/katsayi
    cikti=c(paste("Puan:", total," eksik:",kalan))
    return(cikti)
    }
geridonut(dogruyanit=20,katsayi=2)
## [1] "Puan: 40  eksik: 30"
geridonut(27,2)
## [1] "Puan: 54  eksik: 23"
```

5.1.2.4 Basit Hata

R fonksiyonlarının çalışması için argumanların doğru kullanılması gerekir. Eğer *geridönüt* fonksiyonuna *katsayı* parametresini girmezseniz bir hata ile karşılaşırsınız.

```
geridonut=function(dogruyanit, katsayi){
    total=dogruyanit*katsayi
    kalan=(100-total)/katsayi
    cikti=c(paste("Puan:", total," eksik:",kalan))
    return(cikti)
    }
geridonut(dogruyanit=20)
## Error in geridonut(dogruyanit = 20): argument "katsayi" is missing, with no default
```

5.1.2.5 Basit uyarı

R fonksiyonları uyarı içerebilir. Daha önce yazdığımız *eksipuan* fonksiyonunu düşünelim

```
eksipuan=function(puan, yanlis){
    cikti=puan - (0.2 * yanlis)
    return(cikti)
    }
eksipuan(puan=50,yanlis=10)
## [1] 48
```

Bu fonksiyona bir uyarı ekleyebiliriz. Örneğin hesaplanacak puan sıfırın altında ise bir uyarı verebbiliriz.

```
eksipuan2=function(puan, yanlis){
      cikti=puan - (0.2 * yanlis)
      if (cikti<0)
    warning("Yeni puan 0'dan düşük")
    return(cikti)
    }
eksipuan2(puan=10,yanlis=60)
## Warning in eksipuan2(puan = 10, yanlis = 60): Yeni puan 0'dan düsük
## [1] -2
```

5.1.2.6 Basit Sekte

Bir R fonksiyonu, yazarın belirlediği durumlarda sekteye uğrayabilir. Örneğin *eksipuan3* fonksiyonunu 20'den düşük puanlar için düzeltme yapmayacak şekilde yazabiliriz.

```
eksipuan3=function(puan, yanlis){

    if ((puan)<(20))
    stop("20den düşük puanlar için bu fonksiyon işlemez")

     cikti=puan - (0.2 * yanlis)
      return(cikti)
    }
eksipuan3(10,9)
## Error in eksipuan3(10, 9): 20den düsük puanlar için bu fonksiyon islemez
```

5.1.3 Yardım!

Her R kullanıcısı yeni fonksiyonlar yazmak zorunda değildir, fakat fonksiyonların nasıl çalıştığını bilmek önemlidir. Eğer R bir hata veriyorsa bu genellikle kullanıcı veya datadan kaynaklıdır. Her ne kadar çok

karşılaşılmasa da hatanın fonksiyonun kendisinden kaynaklandığı durumlar da olabilir.

R fonksiyonlar sayesinde çalışır. Dünyanın her yerinden araştırmacılar R fonksiyonları yazmakta, bu fonksiyonları bir R paketi olarak erişime açmaktadırlar. Hali hazırda 10 binden fazla R paketi vardır. R programını indirdiğinizde yaklaşık 30 R paketi bilgisayarınıza otomatik olarak indirilir. Bu 30 R paketinde binlerce fonksiyon bulunur.

R programınızı yüklediğinizde otomatik olarak yüklenen 30 paketten bir tanesi *base* dir. Bu paketin içinde 1200'den fazla fonksiyon bulunur. Örneğin *mean* fonksiyonu aritmetik ortalama hesaplar.Genellikle paketler detaylı açıklamalar ile birlikte sunulur. Kullanıcıların bu açıklamalara ulaşabilmesi için çeşitli yollar mevcuttur; *help*, *?*, *??* veya *example*

```
help("base")  #
help(mean)    # aritmetik ortalama fonksiyonu ve argümanları
?mean         # aritmetik ortalama fonksiyonu ve argümanları
??mean        # aritmetik ortalama fonksiyonu ve argümanları
example(mean) # aritmetik ortalama fonksiyonu ve argümanları
```

5.2 R Data Tipleri

Bu bölümde vektörler, matrisler, değişken çeşitleri, kayıp veriler ve data çerçeveleri (data frames) kısaca tanıtılmıştır.

5.2.1 Vektörler

R *c* fonksionu ile vektör oluşturabilir. 10 öğrenci için not girelim.

```
notlar=c(40,50,53,65,72,77,79,81,86,90)
notlar
##  [1] 40 50 53 65 72 77 79 81 86 90
```

R vektörler üzerinden işlem yapabilir.

```
notlar=c(40,50,53,65,72,77,79,81,86,90)
#her nota 10 ekle
notlar+10
##  [1]  50  60  63  75  82  87  89  91  96 100
#her nota yüzde 10 ekle
notlar+(notlar*0.10)
##  [1] 44.0 55.0 58.3 71.5 79.2 84.7 86.9 89.1 94.6 99.0
#kendi ile çarp
notlar*notlar
##  [1] 1600 2500 2809 4225 5184 5929 6241 6561 7396 8100
# yeni notlar
notlar2=c(30,40,46,58,64,66,69,72,74,81)
# notlar ve notlar2 nin ortalamasını al
(notlar+notlar2)/2
##  [1] 35.0 45.0 49.5 61.5 68.0 71.5 74.0 76.5 80.0 85.5

# ilk notların yüzde 40ı ile ikinci notların yüzde 60ını topla
notlar*0.4 + notlar2*0.6
##  [1] 34.0 44.0 48.8 60.8 67.2 70.4 73.0 75.6 78.8 84.6
```

Vektör oluşturmak için işeyarar birçok fonksiyon vardır. Örneğin *rep* fonksiyonu (bknz: example(rep)) aynı değerleri tekrarlamak için kullanışlıdır.

rnorm fonksiyonu normal dağılıma sahip veriler simüle etmek için işe yarar. Eğer *?rnorm* kullanılırsa bu fonksiyonun 3 argümanı olduğu görülür *rnorm(n, mean = 0, sd = 1)*. Bu fonksiyon vektör uzunluğu (değişken sayısı) *n* verilmediği sürece çalışmaz. Eğer sadece *n* verilirse, popülasyon ortalaması 0 ve standart sapması 1 olan dağılımdan rasgele seçilen değerler ile bir vektör oluşturulur. Bu parametreler değiştirilebilir. Örneğin *rnorm(12,mean=10,sd=2)* popülasyon parametreleri 10 ve 2 olan normal bir dağılımdan 12 adet gözlem çeker. Benzer bir fonksiyon *runif(n, min = 0, max = 1)* tekdüzey bir dağılımdan gözlem çeker.

```
a=1:12            # a 1 den 12ye tam sayılar
rep(0,12)         # 0 12 kez tekrarlanır
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0
rep(1:5,each=3)   # 1 den 5'e tam sayılar 3er kez tekrarlanır
##  [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5
rep(1:5,times=3)  # 3 kere 1'den 5'e tekrarla
##  [1] 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
seq(from=1,to=12) # 1'den 12'ye tam sayılar
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
seq(1,25,by=2)    # 1'den 25'e ikişer atla
##  [1]  1  3  5  7  9 11 13 15 17 19 21 23 25
seq(1,6,by=0.5)   # 1'den 6'ya 0.5 atla
##  [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
rnorm(12)        #  ~N(0,1) 12 gözlem
##  [1] -1.0091 -1.3822  1.3556 -0.4352 -2.3159 -0.1603 -0.0427 -1.2246
##  [9] -1.5322  1.0750 -0.5040 -0.5663
rnorm(12,mean=10,sd=2) #~ N(10,2) 12 gözlem
##  [1]  7.94  4.75  9.87 11.40 10.30  8.41  8.04 12.34  6.00 10.17  8.52
## [12]  9.43
runif(12, min = 10, max = 37)
##  [1] 29.1 32.0 23.5 25.3 14.2 11.4 12.2 19.5 15.6 18.6 35.6 17.1
```

5.2.2 Matrisler

R matrisler oluşturup işlem yapabilir.

```
A=matrix(1:16,ncol=4,nrow=4) #4x4 matris oluştur
A
##      [,1] [,2] [,3] [,4]
## [1,]    1    5    9   13
## [2,]    2    6   10   14
## [3,]    3    7   11   15
## [4,]    4    8   12   16
B=matrix(runif(16,min=20,max=40),ncol=4) #4x4 matris oluştur

# işlem örnekleri
A+B    # topla
##      [,1] [,2] [,3] [,4]
## [1,] 32.1 39.4 32.2 49.8
## [2,] 40.1 29.7 31.9 34.3
## [3,] 31.2 36.4 34.1 45.4
## [4,] 29.8 42.6 45.5 55.9
A*B    # çarp
##       [,1] [,2] [,3] [,4]
```

```
## [1,]  31.1  172  209  478
## [2,]  76.2  142  219  285
## [3,]  84.5  206  254  456
## [4,] 103.2  277  401  639
A%*%B  # matris çarp
##      [,1] [,2] [,3] [,4]
## [1,]  811  867  775  931
## [2,]  934  989  877 1059
## [3,] 1057 1111  978 1186
## [4,] 1180 1233 1080 1313
t(B)   # çevir
##      [,1] [,2] [,3] [,4]
## [1,] 31.1 38.1 28.2 25.8
## [2,] 34.4 23.7 29.4 34.6
## [3,] 23.2 21.9 23.1 33.5
## [4,] 36.8 20.3 30.4 39.9
```

5.2.3 Değişkenler

Çözümlenecek verisetinin özelliklerini bilmek çok önemlidir. R içerisinde çözümlenecek değişkenler genellikle sınıflama, sıralı, sürekli , kayıp veya tarih tipindedir.

5.2.3.1 Sınıflama

R'da bir sınıflama verisi alfanumerik şekilde girilebilir fakat yorumlanması sayısal değil sınıflama şeklindedir. Örneğin;

```
adres=c("AAX","BBZ","CBT","DBA","DDC","XZT")
cinsiyet=c("M","F","F","M","F","M")
id=sample(letters,6)
program=rep(c("var","yok"),each=3)
sehir=as.character(1:6)
```

5.2.3.2 Sıralı

Sıralı bir değişken sınıflama değişkenine göre daha çok bilgi içerir. Sıra ifade eder fakat değerler arasındaki farklılık anlmalı değildir. Örneğin koşucular birinci, ikinci ve üçüncü olarak sıralanabilir fakat bu sıralama verisi birinci ile ikinci arasında kaç dakika farklılık olduğunu belirtmez. Birinci koşucu ikinciden 5 saniye hızlı iken, ikinci koşucu üçüncü koşucudan yarım saat daha hızlı olabilir. R içerisinde *ordered* fonksiyonu ve *level* argümanı ile sıra belirtilebilir. Eğer *level* argüanı boş bırakılırsa R değerleri küçükten büyüğe sıralar.

```
soru1=ordered(c("zayif","orta","iyi","iyi","zayif","zayif"),
              levels=c("zayif","orta","iyi"))
ses=ordered(c(1,3,2,2,1,3),levels=c("1","2","3"))
```

5.2.3.3 Sürekli

Eşit aralıklı veya eşit oranlı değişkenler sıralı ve sınıflama değişkenlerine göre daha fazla bilgi içerir. Değerler arasındaki farklılık anlamlıdır.

```
notlar=c(52,75,39,62,24,86)
notlar=rnorm(n=6,mean=160,sd=5)
```

5.2.3.4 Tarih

as.Date fonksiyonu ile tarih verisi girilebilir.

```
dt=as.Date(c("1994-06-01","1988-10-20","1990-12-01",
                  "1978-03-23","1974-08-22","1994-11-04"))

dt
## [1] "1994-06-01" "1988-10-20" "1990-12-01" "1978-03-23" "1974-08-22"
## [6] "1994-11-04"

tatil=as.Date(c("01/01/2016","04/23/2016","05/19/2016","08/30/2016","09/29/2016"),
                format="%m/%d/%y")

tatil
## [1] "2020-01-01" "2020-04-23" "2020-05-19" "2020-08-30" "2020-09-29"

Sys.Date( )
## [1] "2018-02-27"
Sys.Date( )-dt
## Time differences in days
## [1]  8672 10722  9950 14586 15895  8516
```

5.2.3.5 Doğru-Yanlış (logical)

Bu değişken TRUE veya FALSE değerlerini alır. Eğer sayısal veri olmaya zorlanırsa 1 ve 0 değerlerini alır. Aşağıda verilen kod girilen notların ortalamadan düşük olup olmadığını gösterir.

```
notlar=c(52,75,39,62,24,86)    # notlar
notlar>mean(notlar)
## [1] FALSE  TRUE FALSE  TRUE FALSE  TRUE
as.numeric(notlar>mean(notlar)) # 1 ve 0.
## [1] 0 1 0 1 0 1
```

5.2.4 Faktörler

R içerisinde yer alan *factor* veri tipi sıralı ve sınıflama verileri için kullanılan bir çatıdır.

```
kurs=factor(c("muhasebe","garson","temizlik","garson","muhasebe","garson"))
ga1=factor(c(1,1,3,4,2,3),levels = 1:4,
           labels=c("tamamenkatilmiyorum","katilmiyorum","katiliyorum","tamamenkatiliyorum"))
ga2=factor(c(1,3,4,4,2,3),ordered = T)
ga3=gl(n=3,k=2,labels=c("A","B","C"),ordered=F)
```

Faktörler önemlidir. Faktörlerin alt sınıfları (levels) dikkatli bir şekilde incelenmelidir. Çözümleme aşamasında kullanılmayan alt sınıflar silinmelidir. Örneğin veri seti bölünmeden önce *Renk* faktörü girilmiş olsun, "mavi", "yeşil", "sarı" alt sınıflar olsun. Daha sonra veri seti bölündüğünde alt sınıf "sarı" kullanılmamış olsun. R *Renk* değişkenini hala 3 alt sınıflı olarak düşünecektir ve ona göre işlem yapacaktır. Bu hatalara sebep olur. Droplevel fonksiyonu kullanılarak faktör değişkeni düzeltilmelidir.

```
renk=factor(c(1,1,1,2,2,3),levels = 1:3,labels=c("mavi","yesil","sari"))
renk
## [1] mavi  mavi  mavi  yesil yesil sari
## Levels: mavi yesil sari
renk2=renk[1:5]  # renk2 değişkeni son degeri almadı
renk2  #fakat hala 3 level mevcut
## [1] mavi  mavi  mavi  yesil yesil
## Levels: mavi yesil sari
droplevels(renk2) #kullanılmayan level silindi
## [1] mavi  mavi  mavi  yesil yesil
## Levels: mavi yesil
```

5.2.5 Kayıp Veriler

Veri seti kayıp veriler içerebilir. R kayıp verileri *NA* (not available) ile belirtir.

```
gelir=c("maas","maas","destek",NA,NA,"maas")
hanekisi=c(3,2,3,NA,NA,4)
```

NOT: Kayıp veri belirleyiciler çetrefilli olabilir. NA, , " " (boşluk) veya önceden belirlenmiş bir sayı,örneğin -99 kayıp verileri temsil edebilir.

```
ornek = factor(c('maas','destek', NA, 'NA'," ",-99,"-99"))

# faktör içerisinde kayıp veriler <NA> olarak verilir.
#  < > içinde yer almayan NA faktör sınıfını gosterir.
# " " bu da faktör alt sınıfını gösterir
#-99 and "-99" aynı faktör sınıfını gösterir


# is.na fonksiyonu kayıp verileri gösterir.
# ornek için bakıldığında sadece 3. eleman kayıp veri olarak görünür
is.na(ornek)
## [1] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE


#çözüm 'NA', " ", -99 ve "-99" ları NA'ye çevirelim
ornek[ornek=='NA' | ornek==" "| ornek== -99 | ornek== "-99"]=NA

#kontrol
is.na(ornek)
## [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE

#droplevel kullanalım
ornek=droplevels(ornek)
```

5.2.6 Veri Çerçeveleri (Data Frames)

Bir veri çerçevesi değişkenlerden oluşur. Sosyal bilimcilerin genellikle değişkenler arası ilişkileri araştırdığını düşünürsek, veri çerçeveleri kullanıcılarının temel R ögesidir. Daha önce oluşturduğumuz değişkenleri bir veri çerçevesine alabiliriz;

```
# hatırlatma
# id=sample(letters,6)

# program=rep(c("var","yok"),each=3)

# cinsiyet=c("M","F","F","M","F","M")

# soru1=ordered(c("zayif","orta","iyi","iyi","zayif","zayif"),
#               levels=c("zayif","orta","iyi"))

# ses=ordered(c(1,3,2,2,1,3),levels=c("1","2","3"))

# notlar=c(52,75,39,62,24,86)

# gelir=c("maas","maas","destek",NA,NA,"maas")


# dt=as.Date(c("1994-06-01","1988-10-20","1990-12-01",
#                     "1978-03-23","1974-08-22","1994-11-04"))

# kurs=factor(c("muhasebe","garson","temizlik","garson","muhasebe","garson"))

basit_data=data.frame(id,program,cinsiyet,soru1,ses,
                      notlar,gelir,dt,kurs)
basit_data
##   id program cinsiyet soru1 ses notlar  gelir         dt     kurs
## 1  m     var        M zayif   1     52   maas 1994-06-01 muhasebe
## 2  l     var        F  orta   3     75   maas 1988-10-20   garson
## 3  c     var        F   iyi   2     39 destek 1990-12-01 temizlik
## 4  s     yok        M   iyi   2     62   <NA> 1978-03-23   garson
## 5  h     yok        F zayif   1     24   <NA> 1974-08-22 muhasebe
## 6  q     yok        M zayif   3     86   maas 1994-11-04   garson
```

Veri setleri el yordamı ile girildiğinde veya hazır olarak R a aktarıldığında (örneğin excel dosyasından) veri setinin yapısını incelemek önemlidir. *str* (structure) fonksiyonu kullanılabilir.

```
str(basit_data)
## 'data.frame':    6 obs. of  9 variables:
##  $ id      : Factor w/ 6 levels "c","h","l","m",..: 4 3 1 6 2 5
##  $ program : Factor w/ 2 levels "var","yok": 1 1 1 2 2 2
##  $ cinsiyet: Factor w/ 2 levels "F","M": 2 1 1 2 1 2
##  $ soru1   : Ord.factor w/ 3 levels "zayif"<"orta"<..: 1 2 3 3 1 1
##  $ ses     : Ord.factor w/ 3 levels "1"<"2"<"3": 1 3 2 2 1 3
##  $ notlar  : num  52 75 39 62 24 86
##  $ gelir   : Factor w/ 2 levels "destek","maas": 2 2 1 NA NA 2
##  $ dt      : Date, format: "1994-06-01" "1988-10-20" ...
##  $ kurs    : Factor w/ 3 levels "garson","muhasebe",..: 2 1 3 1 2 1
```

5.3 R Paketleri

R bilgisayarımıza kurulurken 30'dan fazla paket yükler. Bu paketler *sistem kütüphanesinde* saklanır. R paketleri otomatik olarak yüklenen bu 30 paketle sınırlı değildir, örneğin doğrusal karma etkiler modellerini

(linear mixed models) çözümlemek için *lme*(Bates et al., 2015) paketi kullanılabilir. Bu paket 60000'den fazla bilgisayara yüklenmiş ve 1500'den fazla akademik yayında kullanılmıştır. R paketleri genellikle CRAN (comprehensive R archive network) içerisinde bulunur. Paketler yazarlar tarafından güncellendiği sürece CRAN'da bulunur. R paketlerini CRAN'dan çekerek kendi bilgisayarınızda saklayabilirsiniz. Yüklediğiniz paketler *kullanıcı kütüphanesinde* tutulur. Paketleri bir R oturumunda kullanabilmek için onları aktif hale getirmeniz gerekir.

R ve R Studio'yu kurma aşamasında R-RStudio-R paketleri arasında bilgisayar tarafından sağlanan otomatik bir bağ olduğunu farketmiş olabilirsiniz. R studio R'dan sonra yüklendiğinde bilgisayarınızı tarayacak, R programının yerini bulacak ve ona bağlanacaktır. Hem R hem de R Studio R paketlerinizin yerini bulabilir (eğer siz yerlerini değiştirmediyseniz). Eğer R paketlerinizin nerede olduğunu öğrenmek isterseniz *.libPaths()* fonksiyonunu kullanabilirsiniz.

CRAN'da yer alan R kütüphaneleri bilgisayarınıza kolayca yüklenebilir. R studio'da yer alan *Packages* ve *install* sekmesinden veya *install.packages("paketismi")* fonksiyonu ile paketleri indirebilirsiniz. Paketlerin oturum esnasında aktif hale getirilmesi gerekir. Bu işlem R studio *Paketler* sekmesinde yer alan paket isimlerinin yanındaki kutucuğa tıklayarak veya *library("paketismi")* fonksiyonu ile tamamlanabilir. Bu basamaklar Video 4 ile gösterilmiştir.**??**.

5.4 Çalışma alanı (workspace)

Bir R oturumu açtığınızda ve R işlemleri yaptığınızda bu işlemler çalışma alanında yürütülür. Her adımınız R Studio sağ üst köşede yer alan *History* sekmesinde görülür. Çalışma alanınızı oturum sonunda kaydede-

bilirsiniz. Oturum esnasında oluşturduğunuz R çıktıları çalışma alanında tutulur. *ls()* fonksiyonu ile bu çıktıları görebilirsiniz.

R çıktıları çalışma alanına getirilebilir veya çalışma alanı dışına uzun süreliğine kaydedilebilir. Dosyaların yerlerini bulmak ile uğraşmak istemiyorsanız bütün işlemlerinizi aynı klasörde tamamlamayı tercih edebilirsiniz. *getwd()* fonksiyonu size hangi klasör içinde (working directory) olduğunuzu gösterir. Uzun süreliğine kaydetmek istediğiniz bir R çıktısı bu klasöre kolayca kaydedilebilir. Aktif olan klasörünüzü *setwd()* fonksiyonu ile değiştirebilirsiniz. Tabiki bilgisayarınızda her hangi bir klasörde, hatta internette sakladığınız bir nesneyi R çalışma alanınıza getirebilir veya çalışma alanınızda oluşturduğunuz bir dosyayı bilgisayarınızda her hangi bir klasöre kaydedebilirsiniz. Fakat bu durumlarda adresi (location) hatasız bir şekilde R'a bildirmeniz gerekir. Girdi ve Çıktı konuları bir sonraki bölümde ele alınmıştır.

Chapter 6

Veri Setleri

Verileri teker teker kaydetme 5.2.6 bölümünde verilmiştir. Fakat veri setleri genellikle çözümleme yapacak araştırmacıya hazır şekilde gelir. Bu bölüm (a) veri çekme , (b) basit veri işleme etme yöntemleri ve (c) veri kaydetme konularını içerir.

6.1 Veri Çekme

Bir veri seti farklı formatlarda bulunabilir. R kullanıcılarının çok karşılaştığı veri formatları arasında .csv, .sav, .Rdata, .txt sayılabilir. Çözümleme işleminden önce verilerin düzgün bir şekilde çalışma alanına getirilmesi önemlidir.Eğer çalışacağınız veri seti ve R betiği aynı klasör içerisinde ise, bir diğer deyişle veri setiniz çalışma klasörünün (working directory) içerisinde ise adres belirtmeden veriyi çalısma alanınıza çağırabilirsiniz.

6.1.1 CSV

CSV (comma separated values) virgülle ayrılmış değerler içeren dosyalardır. Microsoft Excel kullanıcıları excel formatında yer alan verileri kolayca csv olarak kaydedbilirler. Diğer excel formatları ile kıyaslandığında (xls,xlsx,xlsb, vd.) işlemesi daha kolay veri formatıdır. *read.csv* fonksiyonu ile veri çalısma alanına çağırılabilir. En basit hali ile;

```
data1=read.csv("dataismi.csv")   # eğer çalısma klasöründe dataisim.csv dosyası
                                 # mevut ise çalışır

#Windows için
data1=read.csv("C:\\Users\Desktop\\folderX\\dataismi.csv")  # adres (path)
data1=read.csv("C:/Users/Desktop/folderX/dataismi.csv")  # adres (path)
#NOTE:  \ karakteri hata verir / veya \\ kullanılmalıdır.
```

?read.csv komutu ile fonksiyonun argümanlarını görebilirsiniz. Önemli olan argümanlara örnek;

a) veya değişken isimleri mevcut ise *header=TRUE* aksi halde *header=FALSE* .

b) Kayıp veri belirticiler için *na.strings* . Örneğin *na.strings = "-99"* bütün -99 değerlerinin kayıp veriyi belirttiğini R'a iletir. Benzer şekilde *na.strings = c("-99" , "-9")* hem -99 hem de -9 değerlerinin kayıp veri olduğunu belirtir.

c) Eğer karakter verilerini faktör olarak kullanmak istiyorsanız *stringsAsFactors=TRUE* aksi halde *stringsAsFactors=FALSE*.

d) Veriyi çağırma esnasında değişkenlere yeni isim vermek istiyorsanız *col.names* argümanı kullanılabilir. Örneğin 3 değiskeniniz varsa *col.names=c("A1","B2","C3")* argümanı ile sütunlara isim verebilirsiniz.

Eğer csv dosyanızda ondalık sayılar nokta yerine virgül ile ayrılmış ise (Avrupa ve Türkiye) bu read.csv fonksiyonu için problem oluşturur. Çözüm olarak *read.csv2* fonksiyonunu kullanabilir, veya read.csv içerisinde *sep=";"* ve *dec=","* argümanlarını kullanabilirsiniz. CSV dosyasından veri çağırma basamakları Video 5 ile gösterilmiştir.

6.1.2 SPSS

SAV sosyal bilimciler tarafından kullanılan bir veri formatıdır. *foreign* (R Core Team, 2016a) paketinde yer alan *read.spss* fonksiyonu sav dosyalarını okumak için kullanılabilir.

```
require(foreign)
?read.spss
data=read.spss("dataismi.sav",to.data.frame=TRUE)
# eğer dataisim.sav çalışma klasöründe ise çalılşır
```

6.1.3 Rdata

Rdata formatı genelde daha az bilgisayar hafızası işgal eder. Rdata olarak kaydedilecek her R çıktısı ismi ile birlikte kaydedilir.

```
load("dataisim.Rdata")  #eğer dataisim.Rdata çalışma klasöründe ise çalılşır
```

6.1.4 Sanal Depolardan Veri Çekmek

R ile sanal dünyadan veri çekilebilir. Süreci basite indirgersek, (a) öncelikle verinin nerede yer aldığı doğru şekilde belirlenmelidir, (b) verinin formatı doğru şekilde belirlenmelidir, (c) veri indirilip R' çağrılır veya doğrudan R'a çağrılır. Aşağıda verilen komutlar 2.3 bölümünde tanıtılan dataWBT'nin çalışma alanınıza getirilmesini sağlar.

```
#sanal depodan CSV oku
urldosyasi='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataismi=read.csv(urldosyasi)
str(dataismi)


#sanal depodan Rdata oku
urldosyasi2='https://github.com/burakaydin/materyaller/blob/gh-pages/ARPASS/dataWBT.Rdata?raw=true'
load(url(urldosyasi2))
str(dataWBT)
```

Bu veri setleri Github Depo veya , excel dosyası olarak buradan indirilebilir.

6.1.5 R Stuido Aracılığı ile Veri Çağırma

Veri dosyanız bilgisayarınızda farklı bir klasörde (çalışma klasörü dışında) ise veya tıkla-bırak yöntemini (point-click) tercih ederseniz R Studio'nun sağ üst köşesinde *Environment* sekmesi altında yer alan *import dataset* ile veri çağırabilirsiniz. Bu basamaklar Video 6 ile gösterilmiştir.

6.2 Basit Veri İşlemleri

Genellikle çözümleme basamağına geçilmeden önce verinin işlenmesi gerekir. Bu bölüm (a) değişkenleri yeniden kodlama, (b) alt küme seçme, (c) yeni değişken oluşturma, (d) veri çerçevesini değiştirme, (e) değişken türünü değiştirme ve (f) veri silme işlemlerini kısaca özetler.

6.2.1 Değişkenleri yeniden kodlama

Bir satırı, bir sütünü veya bir satır/sütün keşiminde yer alan tek bir elemanı değiştirmek mümkündür. Değişkenlerin yeni isimler vermek mümkündür. *plyr* (Wickham, 2011) paketi değişkenleri yeniden kodlamada yardımcı olabilir.

```
# sanal depodan CSV oku
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya)

#URL adresini sil
rm(urldosya)


# 151. satır 16. sütunu değiştir ve 30 yap
```

```
veriseti1[151,16]=30

# aynı işlem satır ismi ve sütün ismi verilerek yapılabilir.
# id numarası 67034022 olan satırın yaş değerini 32 yap.
veriseti1[veriseti1$id==67034022,"age"]=32

# tekrar kodlama
# Veri setinde yer alan treatment değişkeni 0 ve 1 olarak girilmiştir.
# 1leri  "trt" ve 2leri  "cnt" yapmak için
veriseti1[veriseti1$treatment==1,"treatment"]="trt"
veriseti1[veriseti1$treatment==2,"treatment"]="cnt"

# ifelse fonksiyonu benzer şekilde çalışır
# "wage01" değişkeninde "wage01"  "Yes" ise 0.5, değil ise -0.5 olarak kodlayalım
veriseti1$wage01=ifelse(veriseti1$wage01=="Yes",0.5,-0.5)

# plyr paketi kullanarak
require(plyr)
# pension01yeni değişkeni pension01 değişkeni üzeriden tanımlanmıştır
# eski değerler olan Yes ve No yerine 1 ve 0 kodlayalım
veriseti1$pension01yeni <- mapvalues(veriseti1$pension01,
                                     from=c("Yes","No"),to=c("1","0"))


#bir değişkene yeni isim verelim
#4. ve 5. değişkenlere yeni isim verelim
colnames(veriseti1)[4]="kurs"
colnames(veriseti1)[5]="bolge"

#isim verme işlemini tek sıra kod ile yapalım
colnames(veriseti1)[c(17,21)]=c("Tgelir","maas1")

#plyr paketini kullanalım gen_att değişkenine toplumsalCinsiyet ismi verelim
veriseti1 <- rename(veriseti1,c('gen_att'='toplumsalCinsiyet'))

#kontrol etmek için head(veriseti) veya summary(veriseti) kullanılabilir

# veriseti1'i çalışma alanından sil
rm(veriseti1)
```

6.2.2 Alt Küme Seçme (Subsetting)

R ile alt küme oluşturmak oldukça kolaydır.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya)

# URL adresi sil
rm(urldosya)
```

```
# sadece İstanbul'u seç
istDAT=veriseti1[veriseti1$city=="ISTANBUL",]

# sadece İstanbul'dan ilk sekiz katılımcıyı seç
istDAT18=veriseti1[veriseti1$city=="ISTANBUL",1:8]

# sadece İstanbul'dan gen_att puanı 2den yüksek olanları seç
istDATGAT2=veriseti1[veriseti1$city=="ISTANBUL" | veriseti1$gen_att >2 ,]

# subset fonksiyonu
# sadece İstanbul'dan gen_att puanı 2den yüksek olan ilk sekiz katılımcıyı seç
istDATGAT2B=subset(veriseti1, city=="ISTANBUL" | veriseti1$gen_att >2, select=1:8)

#item 1 değeri 1,2 ve 3 olan katılımcıları seç
item1_123 <- veriseti1[veriseti1$item1 %in% c(1,2,3), ]

#çalışma alanını temizle
rm(list=ls())
```

6.2.3 Yeni Değişken Oluştur

Daha önce 5 bölümünde değişken oluşturma yöntemlerine değinilmiştir. Bu bölüm hatırlatma olarak görülebilir.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya)

#URL dosyası yükle
rm(urldosya)

# item2'den 6'ya kadar olan sütünları topla
veriseti1$itemTOPLAM=with(veriseti1,item2+item3+item4+item5+item6)

# item2'den 6'ya kadar olan sütünların ortalamasını al (na.rm =T önemli)
veriseti1$itemAVE=with(veriseti1,
                       rowMeans(cbind(item2,item3,item4,item5,item6),na.rm=T))

#veya rowMeans fonksiyonu
veriseti1$itemAVE=rowMeans(veriseti1[,10:14],na.rm = T)

# Şehirler için ortalama hesapla
veriseti1$CityAVEscore =with(veriseti1, ave(itemAVE,city,FUN=function(x) mean(x, na.rm=T)))

#veya
veriseti1=merge(veriseti1, aggregate(itemAVE ~ city, data = veriseti1, FUN=mean, na.rm=TRUE),
          by = "city", suffixes = c("", "citymean"),all=T)

#veya her bir soru için şehir ortalaması hesapla
veriseti1=merge(veriseti1, aggregate(cbind(item2,item3,item4,item5,item6) ~ city,
                                     data = veriseti1, FUN=mean, na.rm=TRUE),
                                     by = "city", suffixes = c("", "Citymean"),all=T)
```

```
# değişkenlerin kategorize edilmesi. Eğer item1AVE 2'den küçük ise 0 aksi halde 1
veriseti1$itemAVE01=ifelse(veriseti1$itemAVE<2,0,1)


# 0 ile 1.8 arasına 1 ver
# 1.8 ve 2.5 arasına 2 ver
# 2.5 ile 5 arasına 3 ver
veriseti1$itemAVE123=with(veriseti1,cut(itemAVE, breaks=c(0,1.8,2.5,5), labels = FALSE))
# cut fonksiyonu içerisinde yer alan right=T argümanına göz gezdirin
# örneğin  right=T  ise değeri tam olarak 1.8 olan değişkenler 1 olur
#          right=F  ise değeri tam olarak 1.8 olan değişkenler 2 olur
```

6.2.4 Veri Çerçevesini Değiştirme (Reshaping data)

Veri çerçevesini değitirmek, uzun formattan geniş formata geçmek veya tam tersi gerekli olabilir. *tidyr* (Wickham, 2016) yardımcı olabilir.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya)

#adresi sil
rm(urldosya)

# genişten uzuna item 1den 6 ya kadar olan sütünları item adı altında birleştir
library(tidyr)
data_long = gather(veriseti1, item, score, item1:item6, factor_key=TRUE)

#id'ye göre diz
data_long=data_long[order(data_long$id),]

# uzundan genişe.
data_wide = spread(data_long, item, score)

## belirlediğiniz nesneler dışında çalışma alanını temizle
rm(list=setdiff(ls(),c("veriseti1")))
```

6.2.5 Değişken Türünü Değiştirme

Sayısal girilen verileri faktöre çevirme gibi işlemler çözümleme basamağından önce gerekli olabilir.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya,stringsAsFactors = F)

#URL sil
rm(urldosya)

#treatment değişkenini incele
str(veriseti1$treatment)
```

```
#sayısal veriyi faktöre çevir
veriseti1$treatmentFactor=factor(veriseti1$treatment,labels=c("treatment","control"))

#karakter olarak girildiğinde faktörleri sayıya çevirme

veriseti1$iv1=factor(rep(c("1","2","3"),length=nrow(veriseti1)))
veriseti1$iv1numeric=as.numeric(levels(veriseti1$iv1))[veriseti1$iv1]
#veya
veriseti1$iv1numeric=as.numeric(as.character(veriseti1$iv1))

#NAleri -99'a çevir
veriseti1[is.na(veriseti1)]= (-99)

#çalışma alanını temizle
rm(list=ls())
```

6.2.6 Veri Silme

Bir tek hücreyi, bir satırı veya bir sütunu silmek gerekebilir.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya,stringsAsFactors = F)

#URL sil
rm(urldosya)

#3. satır 5. sütunda yer alan hücreyi sil
veriseti1[3,5]=NA

#3. satırı sil
veriseti1[3,]= NA

#veya
veriseti1=veriseti1[-3,]

#course taken isimli sütunu sil
veriseti1$course_taken=NULL

#gösterim amaçlı veri oluştur
temp=veriseti1[,1:10]

# kayıp verili satırı silme (listwise)
temp=na.omit(temp)

#çalışma alanını temizle
rm(list=ls())
```

6.3 Veri Kaydetme

Adres belirtmediğiniz sürece kaydetme işlemi mevcut çalışma klasörünüz (working directory) içeriside tamamlanır.

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
veriseti1=read.csv(urldosya,stringsAsFactors = F)

#URL sil
rm(urldosya)

#nesne oluştur.
subset1=veriseti1[1:20,1:5]
object2=mean(veriseti1$item1,na.rm = T)


#çalışma klasörünüzü kontrol edin
getwd()

# Rdata olarak sakla
save(subset1,file="subset1Rfile.Rdata")
# adres vererek sakla
save(object2,file="C:/Users/Desktop/object2Rfile.Rdata")

# csv olarak sakla
write.csv(subset1,file="subset1CSVfile.csv",row.names = F)

#sps dosyası olarak sakla
library(foreign)
write.foreign(subset1, "subset1SPSfile.txt","subset1SPSfile.sps", package="SPSS")

#çalışma alanını temizle
rm(list=ls())
```

Chapter 7

Betimleyici İstatistikler ve Hipotez Testi

Betimleyici istatistikler örneklemi tanımlamayı amaçlar. Bu bölüm içeriside daha önce tanıtılan dataWBT (2.3) kullanılarak (a) betimleyici istatistikler hesaplanmış (b) basit grafikler çizilmiş ve (c) hipotez testi açıklanmıştır.

Bu bölümde yer alan R kodlarını kullanmak isteyen araştırmacıların bir önceki bölümü inceledikleri varsayılmıştır. Bu bölümde yer alan basamakların atlanmadan takip edildiği varsayımı da yapılmıltır. dataWBT çalışma alanınıza çağırmak için;

```
# CSV yükle
urldosya='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urldosya)

#remove URL
rm(urldosya)
```

7.1 Betimleyici İstatistikler

Bu alt bölümde ortalama, ortanca, varyans, standart sapma, çarpıklık ve basıklık hesaplanmıştır. Örneklerde toplumsal cinsiyet algısı (gen_att) değişkeni kullanılmıştır.

7.1.1 Ortalama

Eşitlik (7.1) 'de verildiği gibi, ortalama, bir değişkeni oluşturan değerlerin toplamının toplam değer sayısına bölünmesi ile hesaplanır.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{7.1}$$

```
# gen_att değişkeninin ortalamasını hesapla
mean(dataWBT$gen_att,na.rm = T)
## [1] 1.94
```

```
# birden fazla değişkenin ortalamasını hesapla
# ?colMeans
colMeans(dataWBT[,c("gen_att","item1")],na.rm = T)
## gen_att   item1
##    1.94    3.45
```

7.1.2 Ortanca

Büyükten küçüğe veya küçükten büyüğe dizilmiş bir değişkenin orta noktasına ortanca denir. Eğer değişkenin eleman sayısı (n) tek sayı ise $((n+1)/2)$. sırada yer alan, eğer çift sayı ise $(n/2)$. ve $((n+1)/2)$. değerin ortalaması ortancayı verir.

```
# Ortanca hesapla
median(dataWBT$gen_att,na.rm = T)
## [1] 2
```

7.1.3 Varyans

Varyans değişkenin ne kadar yayıldığını anlamada çok kullanılan bir ölçüdür. Eşitlik (7.2) ile hesaplanır.

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \tag{7.2}$$

```
#varyans hesapla
var(dataWBT$gen_att,na.rm = T)
## [1] 0.364
```

7.1.4 Standart Sapma

Varyansın kareköküdür ve Eşitlik (7.3) ile hesaplanır.

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \tag{7.3}$$

```
#SS hesapla
sd(dataWBT$gen_att,na.rm = T)
## [1] 0.603
```

7.1.5 Çarpıklık (Skewness)

Çarpıklık değeri dağılımın şekli hakkında bilgi verir. Tamamen simetrik olan bir dağılımın çarpıklık değeri 0'dır.

Dağılımın sol kuyruğu sağ kuyruğuna nazaran uzun olduğunda çarpıklık değerinin sıfırdan küçük çıkması tipiktir. Bu tür dağılımlar sola çarpık veya negatif çarpık olarak isimlendirilir. Bu tür dağılımlarda medyan ortalamadan yüksektir

Dağılımın sağ kuyruğu sol kuyruğuna nazaran uzun olduğunda çarpıklık değerinin sıfırdan büyük hesaplanması tipiktir. Bu tür dağılımlar sağa çarpık veya pozitif çarpık olarak isimlendirilir. Bu tür dağılımlarda medyan ortalamadan küçüktür.

Örneklem için çarpıklık formülü [1]

$$\sqrt{n}\frac{\sum_i^n \left(X_i - \bar{X}\right)^3}{\left(\sum_i^n \left(X_i - \bar{X}\right)^2\right)^{3/2}} \tag{7.4}$$

Örneklem için çarpıklık değeri *moments* (Komsta and Novomestky, 2015) paketinde yer alan *skewness* fonksiyonu ile hesaplanabilir.

```
#çarpıklık hesapla
library(moments)
skewness(dataWBT$gen_att,na.rm = T)
## [1] 0.377
```

```
NOT: Çarpıklık ve basıklık değerleri için standart hata ve sonrasında z-puanı hesaplanabilir.
Hesaplanan bu z-puanı seçilen bir kritik değer ile (ör. 1.96) kıyaslanarak
çarpıklık veya basıklığın istatistiksel olarak anlamlı olup olmadığı sınanabilir.
Benzer şekilde normallik testleri de (ör. Shapiro-Wilk) yapılabilir. Fakat bu testler
örneklem büyüklüğüne hassasstır. Bir diğer deyişle örneklem büyüdükçe çok küçük farklılıklar
istatistiksel olarak anlamlı bulunabilir.
Çarpıklık,basıklık veya normallik testlerinin varsayım ihlallerini tespit etmek
üzere kullanılışı nispeten eskimiş yöntemlerdir. Bu testleri kullanmak yerine normallik
grafik üzerinden incelenip, dirençli tahminleyicilerin (robust estimators) veya
Monte Carlo simulasyon tekniklerinin çıktıları incelenebilir.
```

7.1.5.1 Çarpıklık örnekleri

Normal bir dağılım ve çarpıklık istatistiği;

Sola çarpık sürekli değişken;

Sağa çarpık sürekli değişken;

7.1.6 Basıklık (Kurtosis)

Basıklık değeri dağılımın şekli hakkında bilgi verir. Normal bir dağılımın Pearson basıklık değeri 3'tür. Eşitlik (7.5) basıklık değerinin hesaplanışını gösterir.

$$n\frac{\sum_i^n (X_i - \bar{X})^4}{(\sum_i^n (X_i - \bar{X})^2)^2} \tag{7.5}$$

Eşitlik (7.5) sıfırdan küçük değerler vermez. 0 ile 3 arasında yer alan değerler genellikle düz dağılımlarda hesaplanır, örneğin tekdüzey dağılımlar. Uzun kuyruklu dağılımlarda 3'ten büyük değerler görülebilir. Alanyazında yorumu kolaylaştırmak için Eşitlik (7.5) 'ten 3 çıkarıldığı durumlar mevcuttur.

Örneklem için Pearson basıklık değeri *moments* (Komsta and Novomestky, 2015) paketinde yer alan *kurtosis* fonksiyonu ile hesaplanabilir.

```
#basıklık hesapla
library(moments)
kurtosis(dataWBT$gen_att,na.rm = T)
## [1] 2.9
```

[1] Bu formül popülasyon için yanlı (biased) bir tahminleyicidir. R yanlı olmayan çarpıklık ve basıklık istatistikleri hesaplayabilir, *describe* fonksiyonu *type* argümanı incelenebilir.

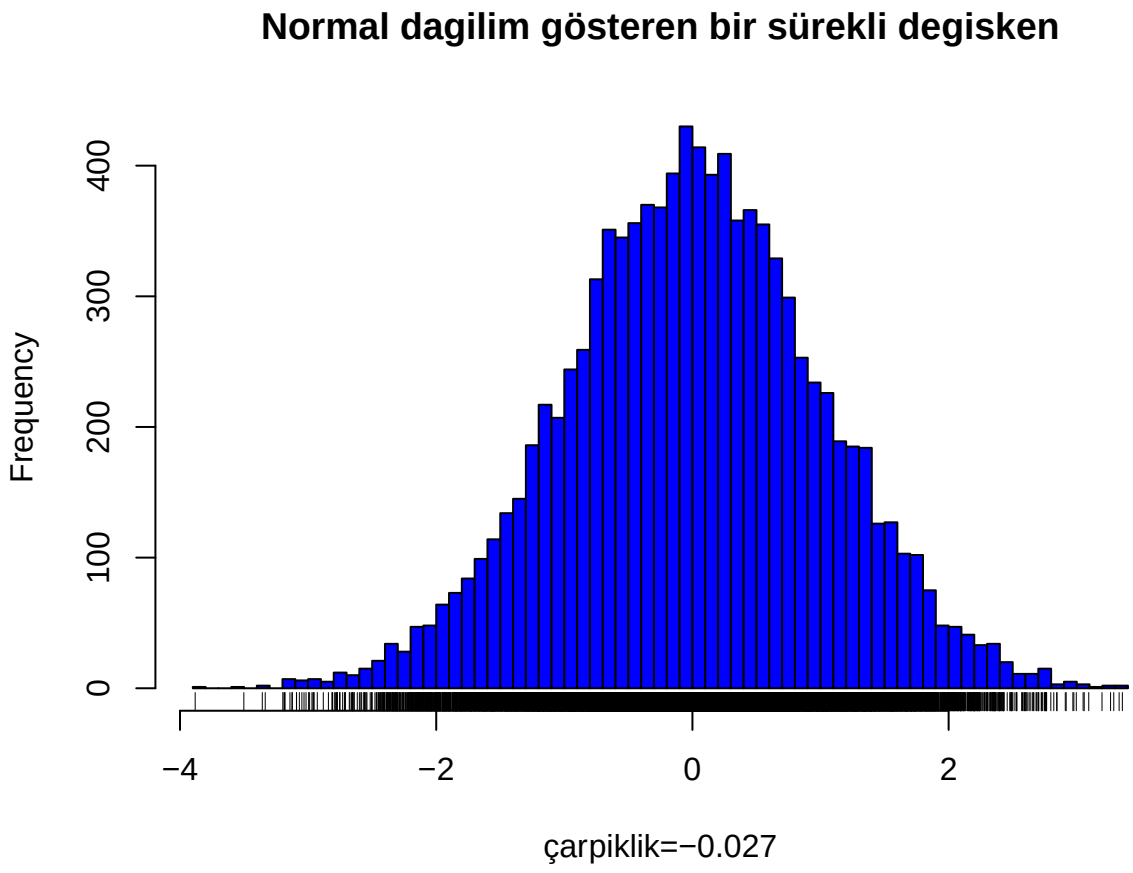


Figure 7.1: Normal dagilim gösteren bir sürekli degisken

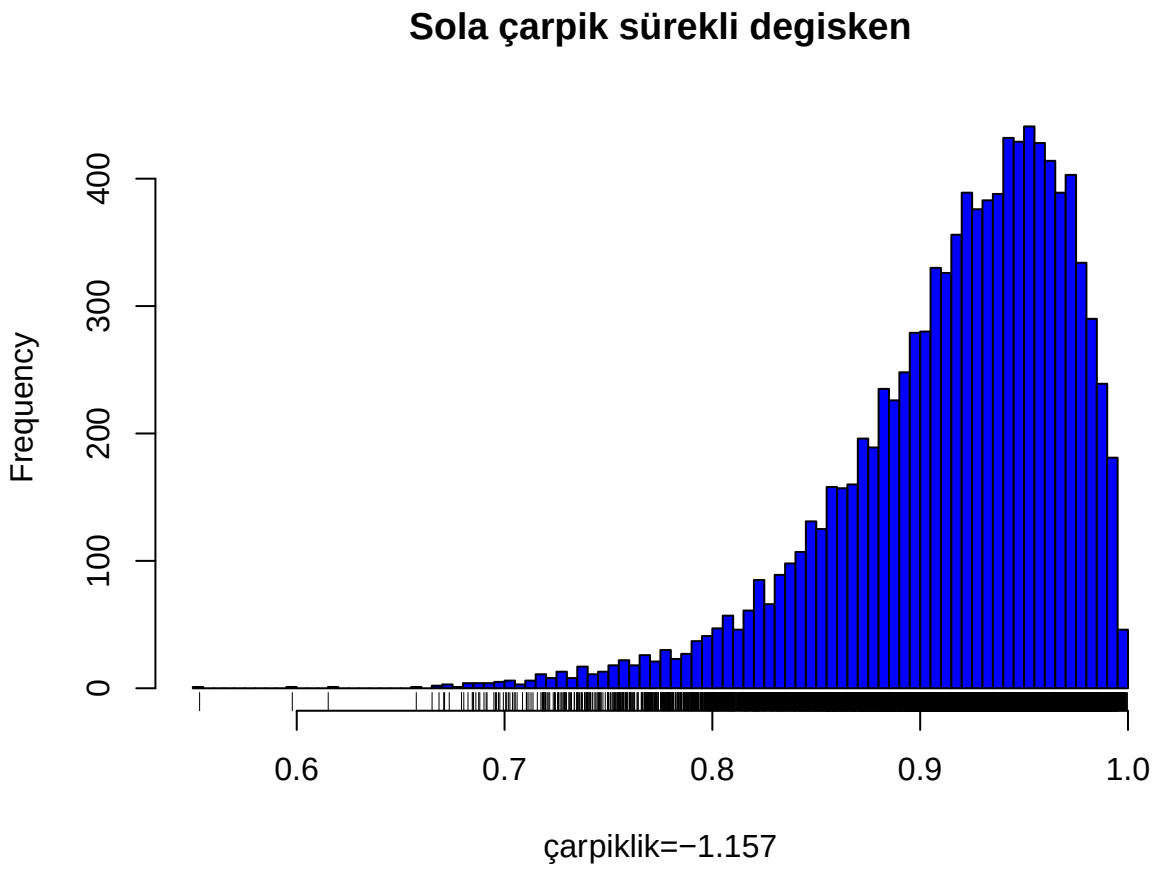


Figure 7.2: Sola çarpik sürekli degisken

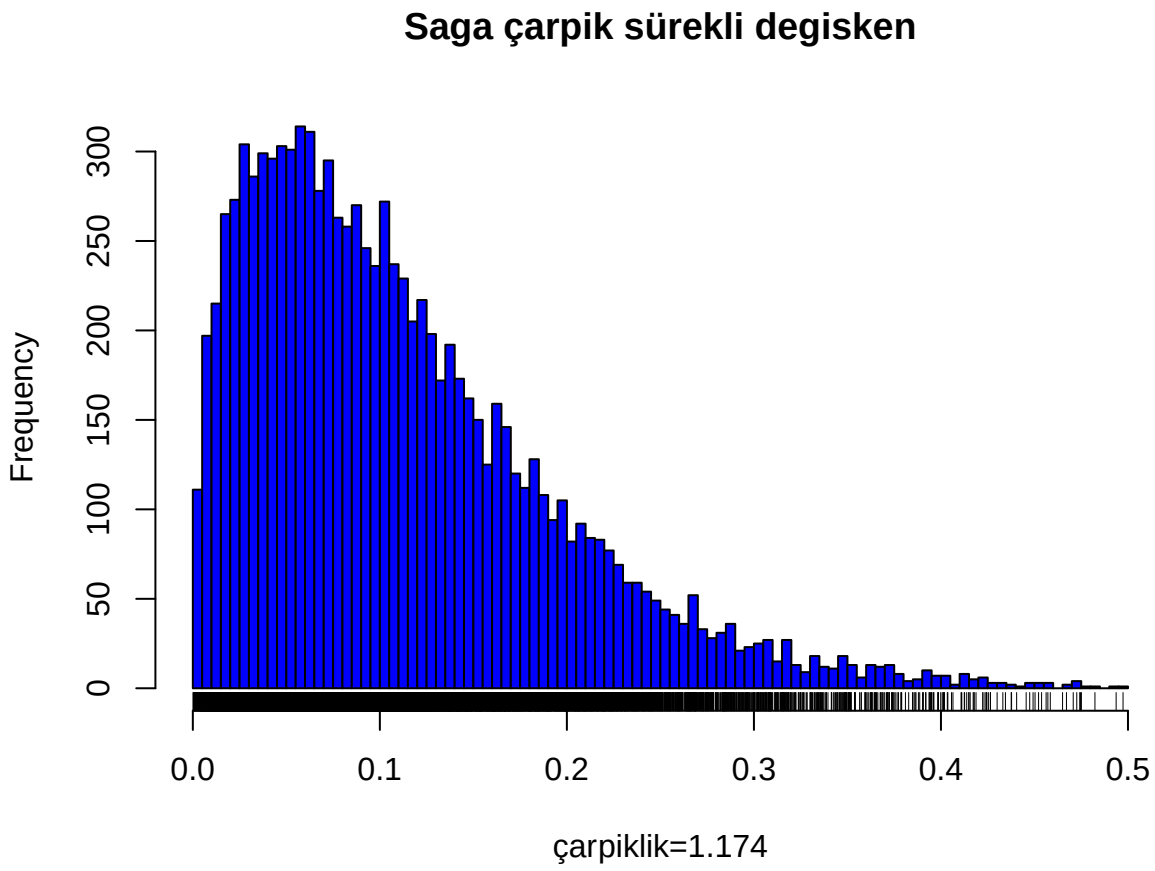


Figure 7.3: Saga çarpik sürekli degisken

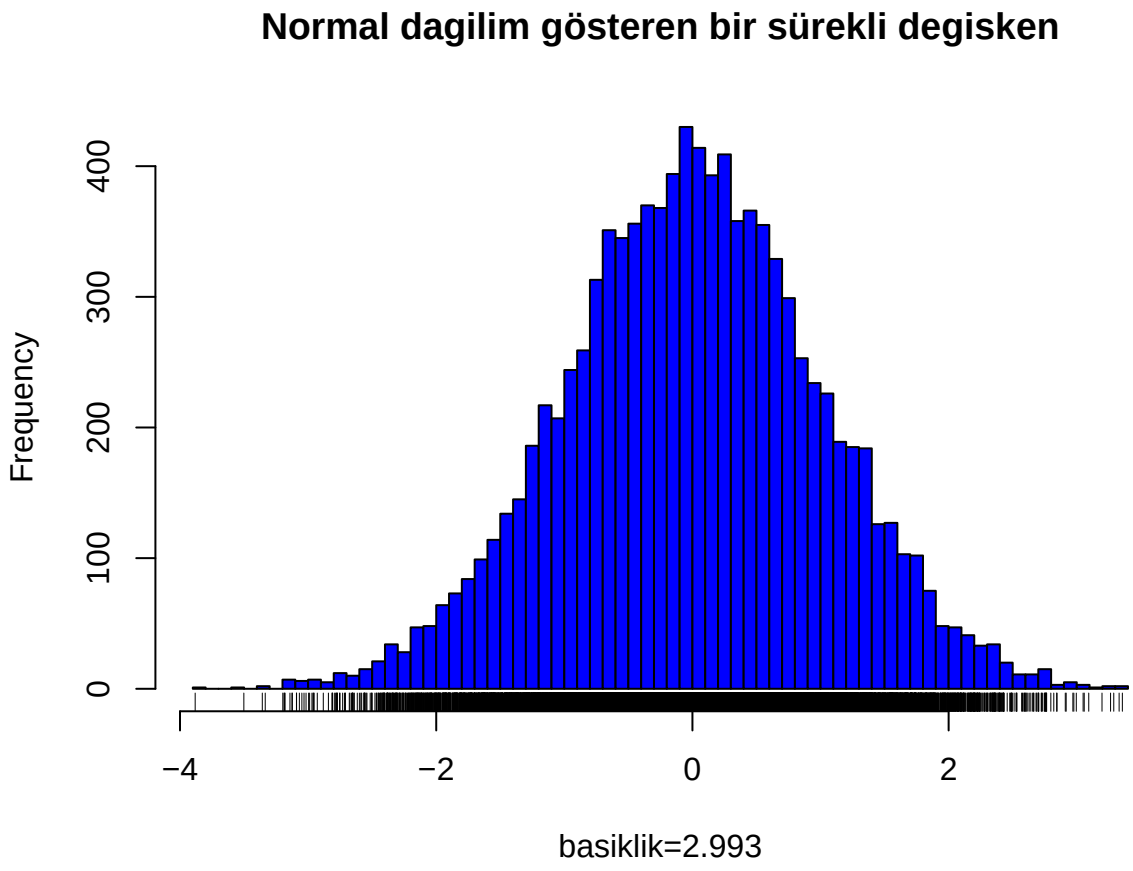


Figure 7.4: Normal dagilim gösteren bir sürekli degisken

7.1.6.1 Basıklık Örnekleri

Normal bir dağılım ve basıklık ölçüsü

Tekdüzey bir dağılım ve basıklık değeri

Beta dağılımı gösteren bir sürekli değişken

7.1.7 Betimleyici İstatisklerin Raporlanması

psych (Revelle, 2016), *doBy* (Højsgaard and Halekoh, 2016) ve *apaStyle* (de Vreeze, 2016) paketleri betimsel analizleri rapor etmede yardımcı olabilir.

```
# psych paketi describe fonksiyonu sırasıyla;
# n: gözlem sayısı (kayıp veriler hariç)
# ortalama, ss, ortanca, budanmış ortalama (trim=0.05 5% budanmış)
# ortanca mutlak dağılımı (median absolute deviation),
# minimum, maksimum, ranj
# çarpıklık ve basıklık-3 (type=2 popülasyon basıklık ve çarpıklık)
# standart hata
library(psych)
```

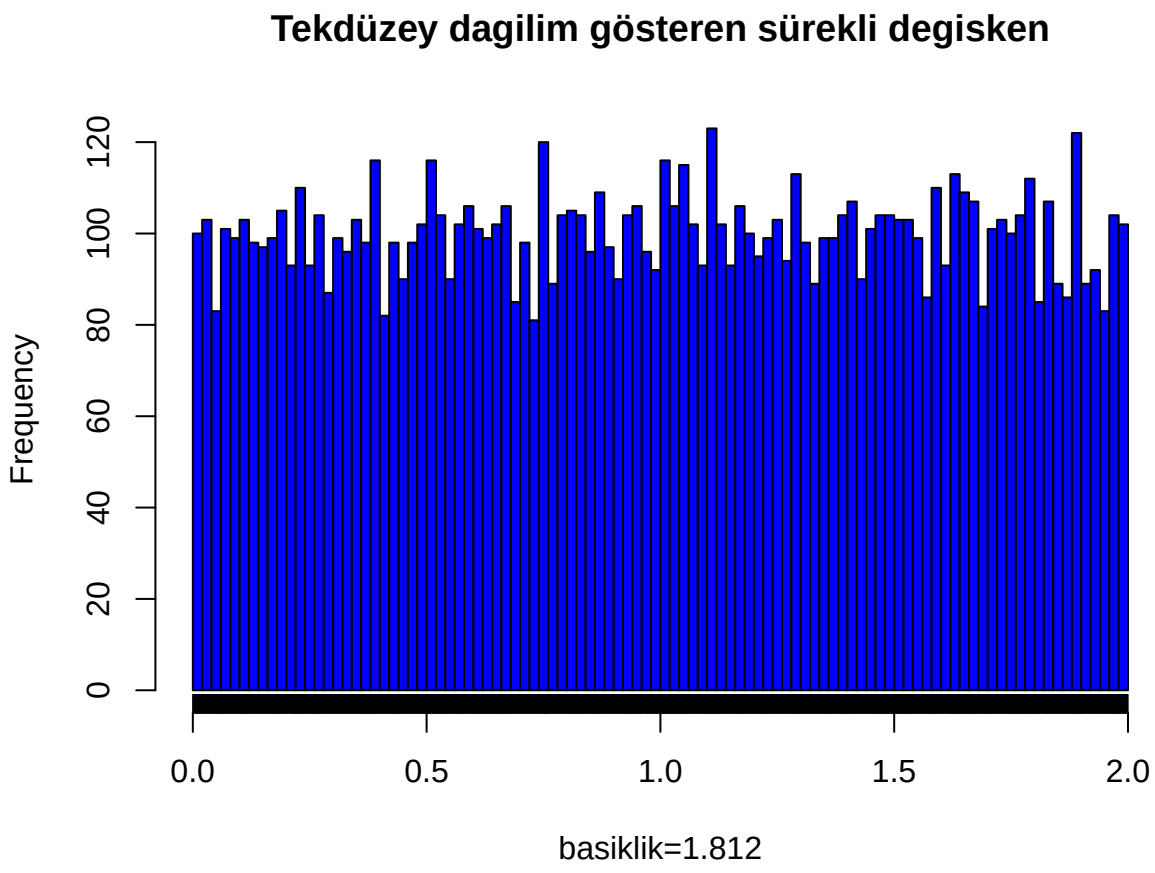


Figure 7.5: Tekdüzey dagilim gösteren sürekli degisken

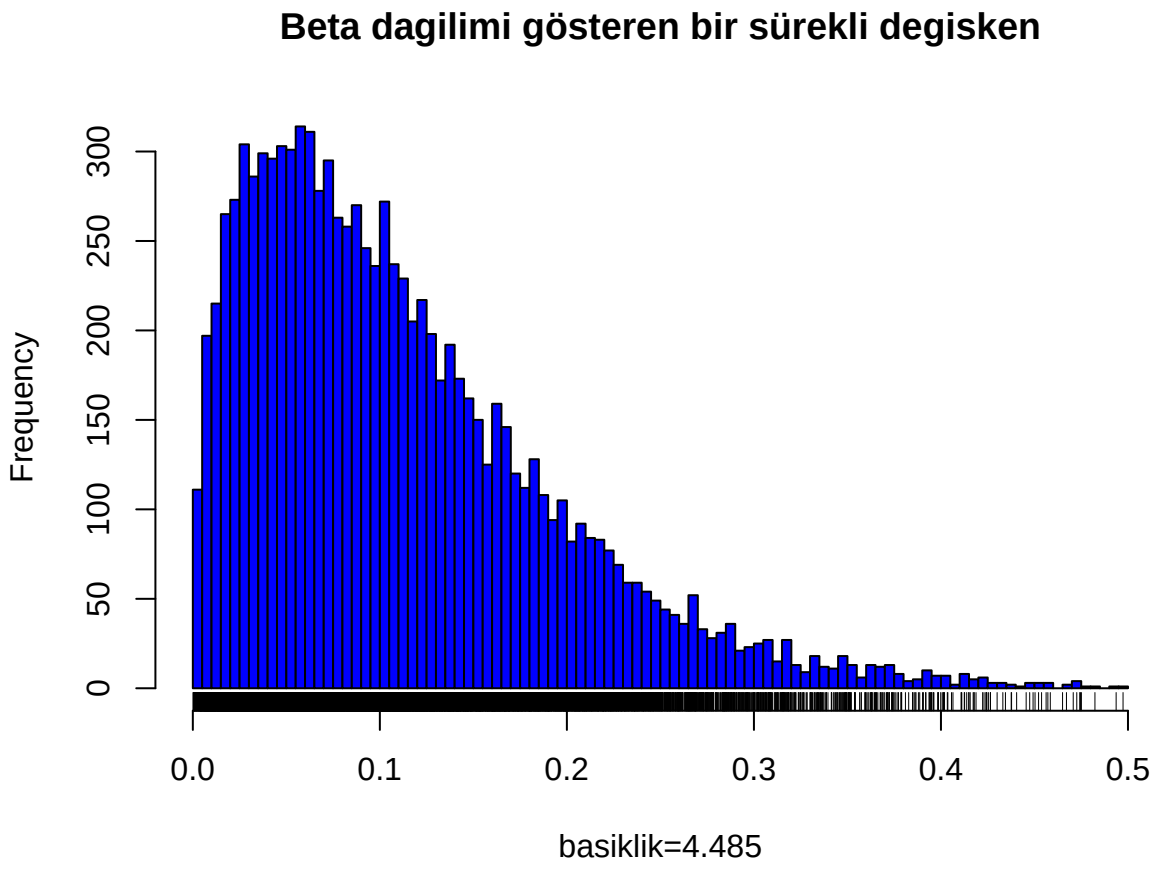


Figure 7.6: Beta dagilimi gösteren bir sürekli degisken

```
desc1=describe(dataWBT[,c("gen_att","age")],trim = 0.05,type=3)
desc1
##         vars    n  mean   sd median trimmed  mad min max range skew
## gen_att    1 5302  1.94 0.60      2    1.92 0.59   1   4     3 0.38
## age        2 5308 27.08 7.21     25   26.62 5.93  15  60    45 0.96
##         kurtosis   se
## gen_att    -0.10 0.01
## age         0.63 0.10

# kaydet
write.csv(desc1,file="pscyhbetimsel.csv")

#doBy
# program değişkenine göre betimleyici istatistikler
library(doBy)
library(moments)
desc2=as.matrix(summaryBy(gen_att+age~treatment, data = dataWBT,
          FUN = function(x) { c(n = sum(!is.na(x)), nmis=sum(is.na(x)),
                                m = mean(x,na.rm=T), s = sd(x,na.rm=T),
                               skw=moments::skewness(x,na.rm=T),
                               krt=moments::kurtosis(x,na.rm=T)) } ))

#yuvarlama
round(desc2,2)
##   treatment gen_att.n gen_att.nmis gen_att.m gen_att.s gen_att.skw
## 1         1      2736          265      1.93       0.6        0.38
## 2         2      2566          335      1.95       0.6        0.38
##   gen_att.krt age.n age.nmis age.m age.s age.skw age.krt
## 1        2.90  2739      262  26.9  7.17    0.99    3.69
## 2        2.91  2569      332  27.3  7.24    0.93    3.57
write.csv(round(desc2,2),file="doBydesc.csv")

#apaStyle
# APA formatında tablo
library(apaStyle)
apa.descriptives(data = dataWBT[,c("gen_att","age")],
                 variables = c("Gender Attitude","Age"), report = c("M", "SD"),
                 title = "APAtableGenderAge", filename = "APAtableGenderAge.docx",
                 note = NULL, position = "lower", merge = FALSE,
                 landscape = FALSE, save = TRUE)
##
## Word document succesfully generated in: C:/Users/Burak/Desktop/github/SARP

#apaStyle paketi hata veriyorsa;
#https://www.r-statistics.com/2012/08/how-to-load-the-rjava-package-after-the-error-java_home-cannot-be

#Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre1.8.0_111')
```

7.1.7.1 Betimsel İstatistik Rapor Örneği

Toplumsal cinsiyet algısı puanları 5302 katılımcı için 1 ve 4 arasında değişmiştir, ortanca 2, ortalama 1.94 ve standart sapma 0.6 olarak hesaplanmıştır. Puanların dağılımı örneklem bazında 0.38 çarpıklık ve -0.1

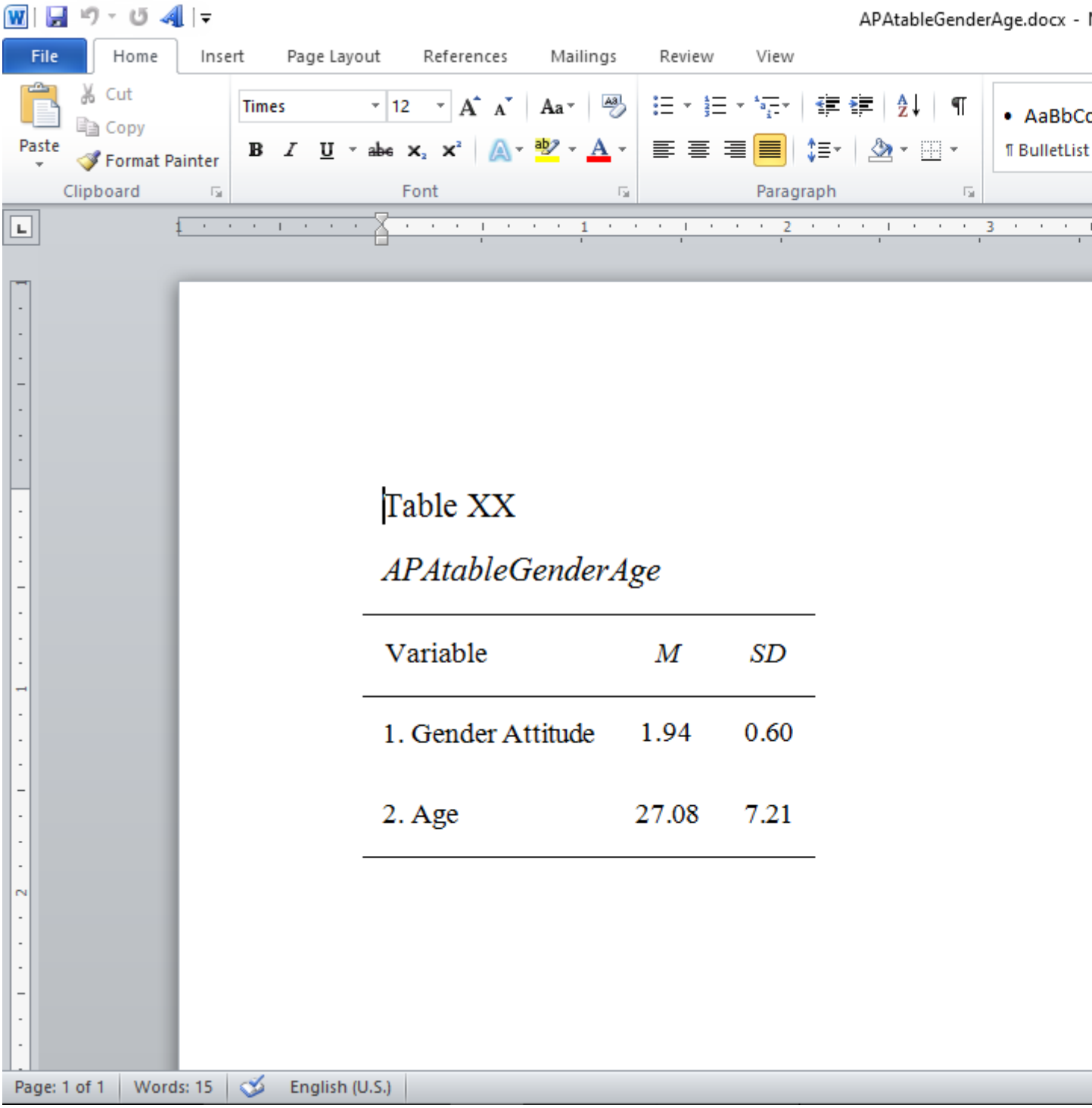


Figure 7.7: APA Tablo.docx

basıklık değerine sahiptir.

7.2 Basit Grafikler

R programı ile basit olmayan grafikler çizilebilir. Popüler olan grafik oluşturma yöntemlerinden dört tanesi, base(R Core Team, 2016b), lattice(Sarkar, 2016), ggplot2(Wickham and Chang, 2016) ve plotrixtir(Lemon et al., 2016). Bu materyal ggplot2 kullanmıştır. Bir ggplot fonksiyonunda argüman sayısı oldukça fazladır, bu sayede kullanıcılar grafiğin her noktasında değişiklik yapabilirler.

7.2.1 Histogram

Dikdörtgenlerden oluşan histogram grafikleri değişken içerisinde yer alan değerlerin frekanslarına göre oluşturulur.

7.2.1.1 Tek değişken için histogram

Dağılım hakkında bilgi sahibi olmak için kullanışlıdır.

```
library(ggplot2)
ggplot(dataWBT, aes(x = gen_att)) +
  geom_histogram(binwidth = 0.2)+ theme_bw()+labs(x = "Toplumsal Cinsiyet Algısı ")+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

7.2.1.2 Tek değişken tek faktör histogram

Gruplara dayalı farklılıkları görmek için kullanışlı

```
dataWBT$HEF=droplevels(factor(dataWBT$higher_ed,
                    levels = c(0,1),
                    labels = c("Lise ve altı", "Üniversite")))

ggplot(dataWBT, aes(x = gen_att, fill=HEF,drop=T)) +
  geom_histogram(breaks=seq(1, 4, by =0.2),alpha=.5,col="black")+
  theme_bw()+labs(x = "Toplumsal Cinsiyet Algısı",fill='Yüksek Öğretim Durumu')+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))

dataWBT2=na.omit(dataWBT[,c("gen_att","HEF")])
ggplot(dataWBT2, aes(x = gen_att)) +
  geom_histogram(breaks=seq(1, 4, by =0.2),alpha=.5,col="black")+
  theme_bw()+labs(x = "Toplumsal Cinsiyet Algısı")+ facet_wrap(~ HEF)+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

```
library(ggplot2)
ggplot(dataWBT, aes(x = gen_att)) +
  geom_histogram(binwidth = 0.2)+ theme_bw()+
   facet_wrap(~city, ncol = 8)
```

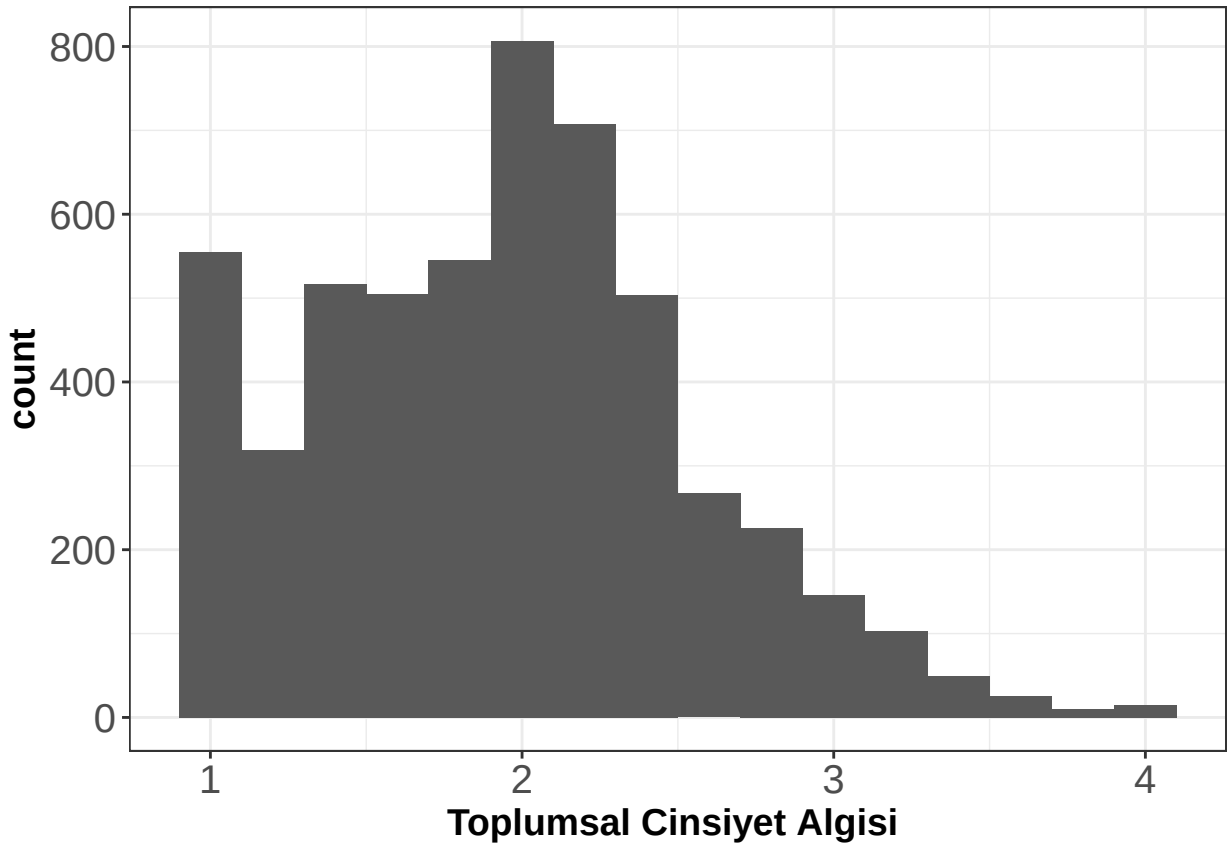


Figure 7.8: Toplumsal Cinsiyet Algisi Puan Dagilimi

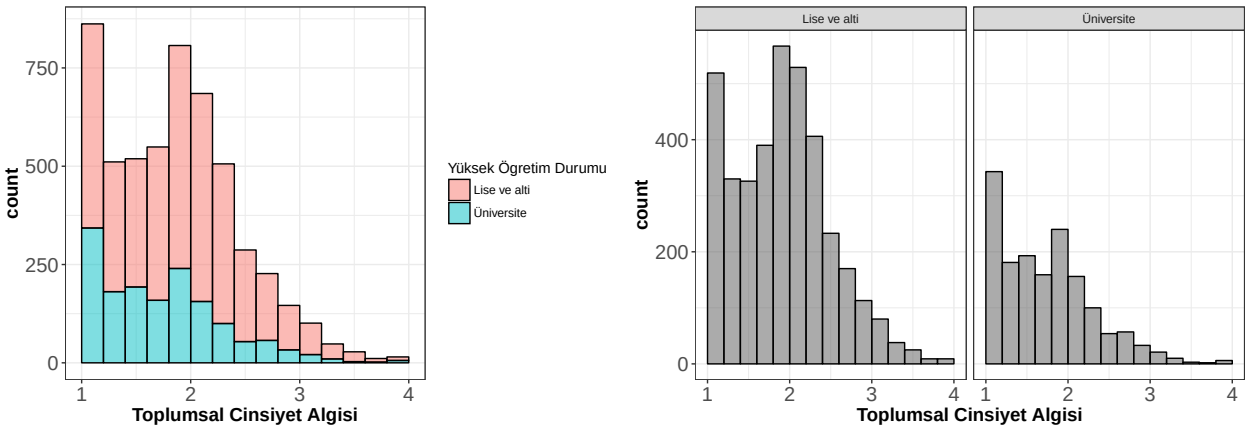
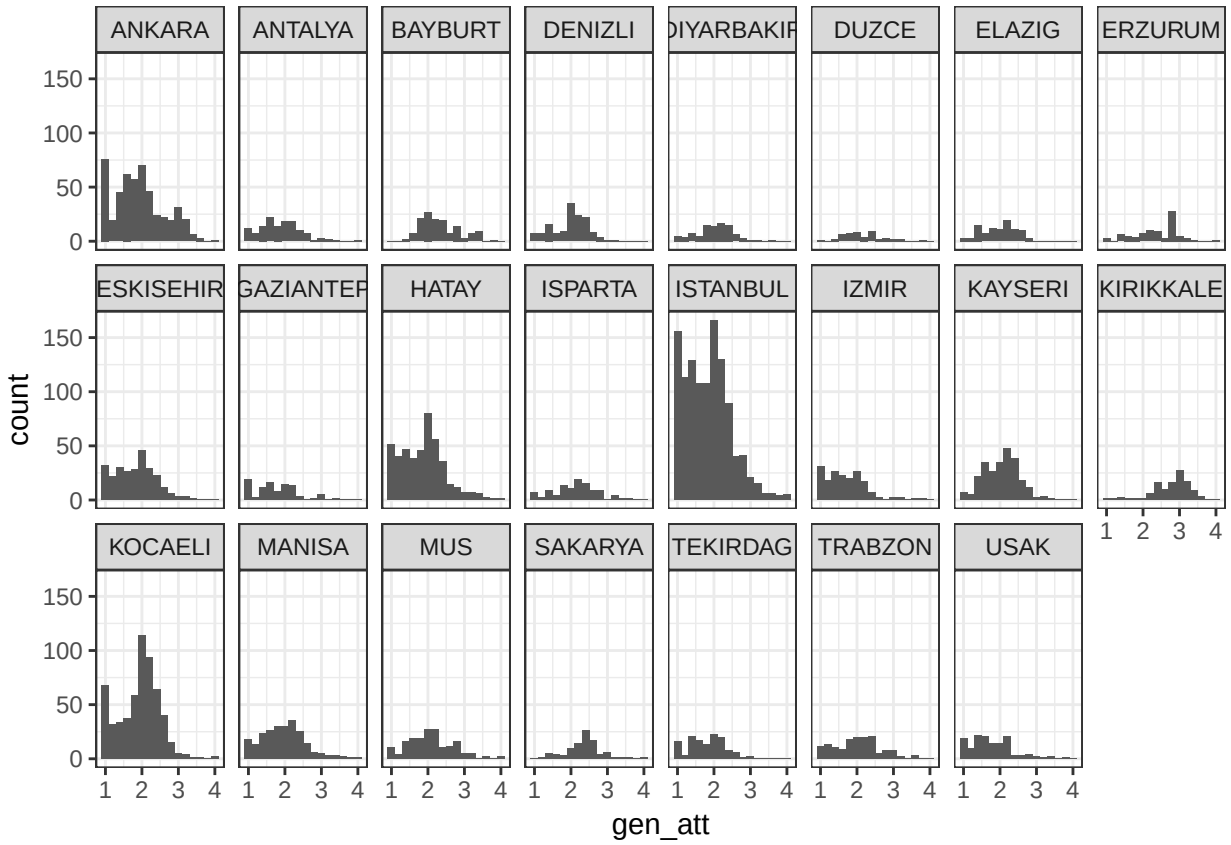


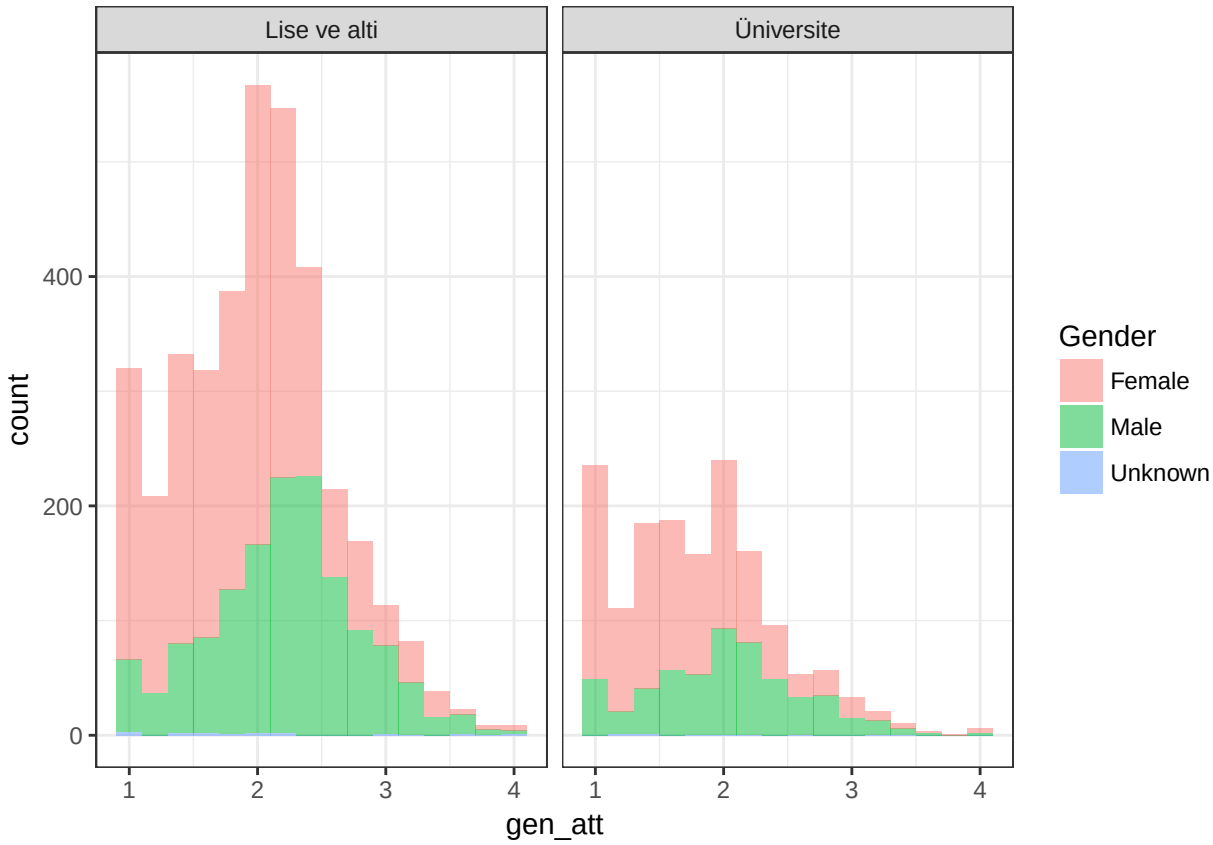
Figure 7.9: Egitime Göre Toplumsal Cinsiyet Algisi



7.2.1.3 Tek değişken iki faktör histogram

Useful for two way interactions

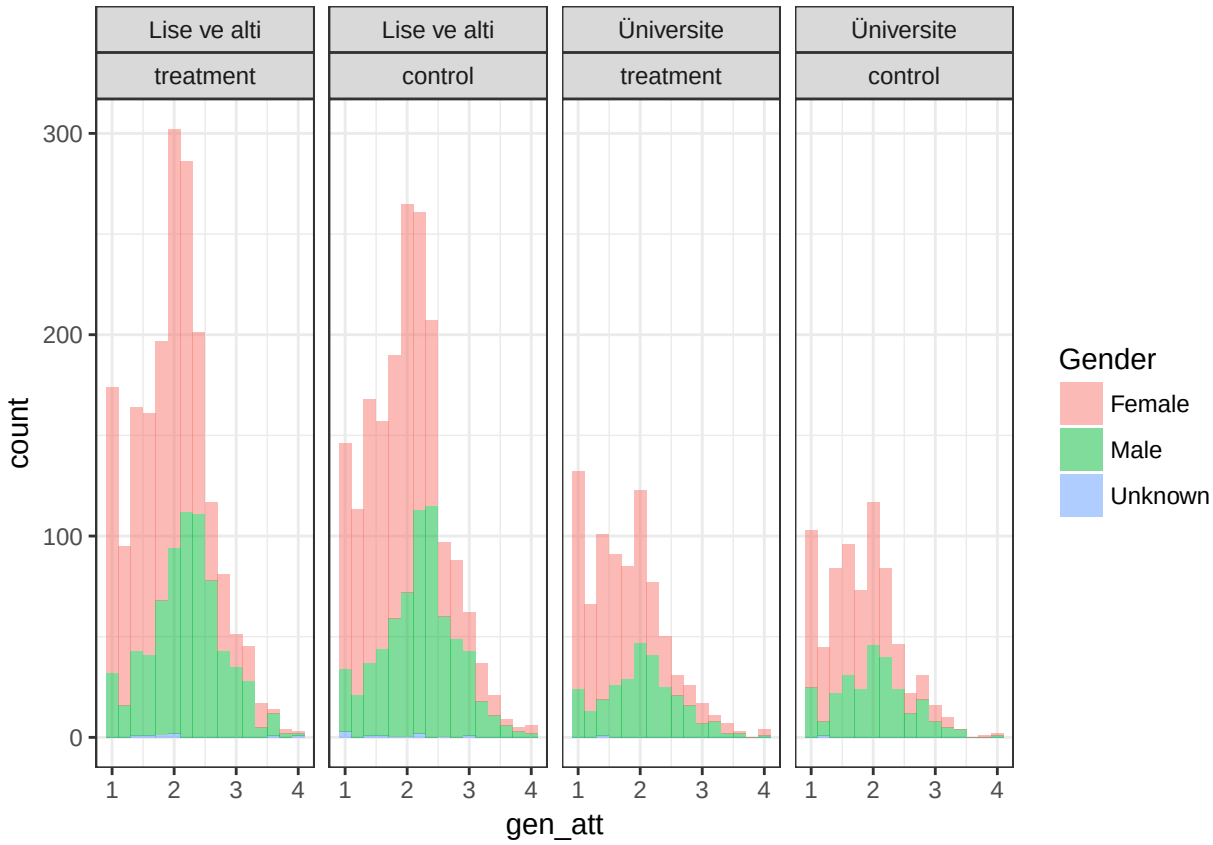
```
dataWBT2=na.omit(dataWBT[,c("gen_att","HEF","gender")])

ggplot(dataWBT2, aes(x = gen_att,fill=gender)) +labs(fill='Gender')+
  geom_histogram(binwidth = 0.2,alpha=.5)+ theme_bw()+
    facet_grid(~HEF)
```



7.2.1.4 Tek değişken üç faktör histogram

Etkileşim (interaction) açıklamada kullanışlı

```
dataWBT$Condition=droplevels(factor(dataWBT$treatment,
                    levels = c(1,2),
                    labels = c("treatment", "control")))

dataWBT2=na.omit(dataWBT[,c("gen_att","HEF","gender","Condition")])

ggplot(dataWBT2, aes(x = gen_att,fill=gender)) +labs(fill='Gender')+
  geom_histogram(binwidth = 0.2,alpha=.5)+ theme_bw()+
    facet_grid(~HEF+Condition)
```



7.3 Hipotez Testi Tanıtım

Cambridge sözlüğü evren (popülasyon) tanımı olarak "aynı ülke, aynı alan veya aynı yerde yaşayan insan veya canlı grubu" cümlesini kullanır. Sosyal bilimlerde evren genellikle "belli bir gruba ait bütün insanlar" olarak belirlenir. Örneğin *sekiz yaşındaki tüm öğrenciler*, *belli bir ülkede bulunan sekiz yaşındaki tüm öğrenciler*, *8 yaşında disleksi teşhisi konulan öğrenciler*.

Sosyal bilimciler araştırma soruları doğrultusunda ilgili evreni tanımlar. Evrende yer alan bireylerin (unit) gözlenlenebilen karakteristik özellikleri değişkenleri oluşturur. Diğer bir ifade ile değişkene ait popülasyon tanımlanabilir. Bölüm 5.2.3 'de değişken türleri açıklanmıştır. Evren değişkene ait bütün değerleri kapsar, bu değerlere ait bir ranj ve görülme olasılığı (probability of occurence) vardır. Yoğunluk (probability, sürekli değişken için) ve çoğunluk (mass, sürekli olmayan değişken için) fonksiyonları görülme olasılıklarını formüle etmek için kullanışlıdır. Dağılım hakkında yapılacak geçerli bir varsayım ile örneklemden evrene genellemeler yapılabilir.

Seçkisiz seçilen bir örneklem ile ulaşılan değişken evrende sahip olduğu bütün değerleri içermeyebilir. Fakat, özellikle gözlem sayısı küçük değil ise, seçkisiz seçme işleminde sistematik bir yanlılık görülmeyeceği düşünüldüğünde, örneklemin evrene benzer özellikler göstermesi beklenir.Bu bilgi kullanılarak evrene ait bir parametre örneklemden yola çıkarak tahmin edilebilir. Bu işlem genellikle bir model sayesinde olur. Modelin veriye gösterdiği uyum araştırmacı tarafından değerlendirilir. Hipotez testleri kullanılan bir model sonrasında araştırma soruları ile ilgili karara varma sürecidir.

7.3.1 Örnekleme Dağılım (Sampling Distribution)

Seçkisiz bir örneklem için hesaplanmış bir istatistik aslında bir değişkendir ve belli bir dağılıma sahiptir. Örnekleme dağılım konusununu açıklamak için en çok kullanılan istatistik aritmetik ortalamadır. Merkezi limit teoremine göre basit seçkisiz örneklem kullanıldığında [2] değişkenin evrende gösterdiği dağılımdan bağımsız olarak, o değişkene ait örneklem ortalamalarının dağılımı yaklaşık olarak normaldir. Örneklem büyüdükçe ;

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}). \tag{7.6}$$

Eğer evren bazında dağılım normal ise (7.6) küçük örneklemler için de doğrudur. Ortalamanın örnekleme dağılımına ait standart sapma *ortalamanın standart hatası* olarak isimlendirilir ve istatistiksel çıkarımlarda kullanılır. Eşitlik (7.6) içinde yer alan μ ve σ^2 bilinmezdir. Fakat bu eşitlik örnekleme ait ortalamanın evrene ait ortalamayı ne derecede kestirebileceğini anlamada önemlidir. Örneğin bir araştırmacının basit seçkisiz yöntemle 10 kişilik bir örneklem seçtiğini düşünelim. Araştırmacının bilemeceği parametrelerin ise $\mu = 100$ ve $\sigma = 15$ olduğunu düşünelim. Bu durumda örnekleme dağılım için standart sapma $(15\sqrt{10}) = 4.74$ olarak bulunur. %95 olasılıkla araştımacının 10 kişilik örneklem ile ulaşacağı aritmetik ortalama 90.7 ve 109.3 arasında olacaktır. Bu oldukça geniş bir aralıktır. Fakat araştırmacı 10 kişi yerine 100 kişiyi aynı örneklem ile seçseydi ulaşacağı örneklem ortalaması %95 olasılıkla 97.1 ve 102.9 olacaktır.

Bu noktada önem taşıyan konu, örneklemden gelen bilgilerle evrene ait parametrelerin hangi tahminleme yöntemleri (estimator) ile yansız, tutarlı ve keskin (unbiased, consistent and efficient) olarak kestirebileceğidir. Örneğin μ, Eşitlik (7.1) ile, σ^2 ise Eşitlik (7.2) ile yansız olarak kestirilebilir.

7.3.1.1 Yansız tahminleme ve örnneklem seçimi

Eklenecek

7.3.2 Güven Aralıkları (The Confidence Intervals (CI))

Dağılım hakkında yapılacak bir varsayım, örneklemden gelen bilgi ve uygun bir tahminleyici (estimator to produce a point estimate) kullanılarak güven aralıkları oluşturulabilir. Bir güven aralığı evren parametresinin muhtemelen hangi aralıkta olduğunu gösterir. Fakat bu evren parametresinin bu aralıkta kesinlikle yer aldığı anlamına gelmez. Örnekleme ait aritmetik ortalamadan yola çıkarak evren parametresi için güven aralığı hesaplamak oldukça basittir.Değişkene ait dağılımın normal olduğu varsayıldığında, ortalamanın örnekleme dağılımı da normaldir ve örnekleme ait ortalama yansız bir tahmindir. Normal dağılım bilindik özellikleri vardır, yoğunluk fonksiyonu değerlerin %95'inin ortalamadan 1.96 standart sapma aşağıda ve yukarıda olduğunu gösterir. Normal dağılımın bu özelliği grafik 7.10 ile gösterilmiştir. Mavi ile işaretlenen bölgeden bir gözlem yapma olasılığı %5tir. Benzer şekilde, mavi veya sarı ile işaretlenen bölgeden gözlem yapma olasılığı da %10'dur. Gri bölge (± 1) yoğunluğun yaklaşık olarak %68'ini kapsar.Bu bilgi kullanışlıdır. Örneklem için hesaplanan ortalama ve varyans μ için güven aralığı hesaplamada kullanılabilir.

7.3.2.1 Güven Aralığı Örneği

Toplumsal Cinsiyet Algısına ait ortalama ve güven aralığı hesaplamaları

```
# gözlem sayısı, n
GA_n=sum(!is.na(dataWBT$gen_att))

#ortalama
```

[2] evrende yer alan her üyenin eşit seçilme olasılığı olduğunda ve seçilen bir üyenin diğer bir üyenin seçilme olasılığını etkilemediği durumlarda

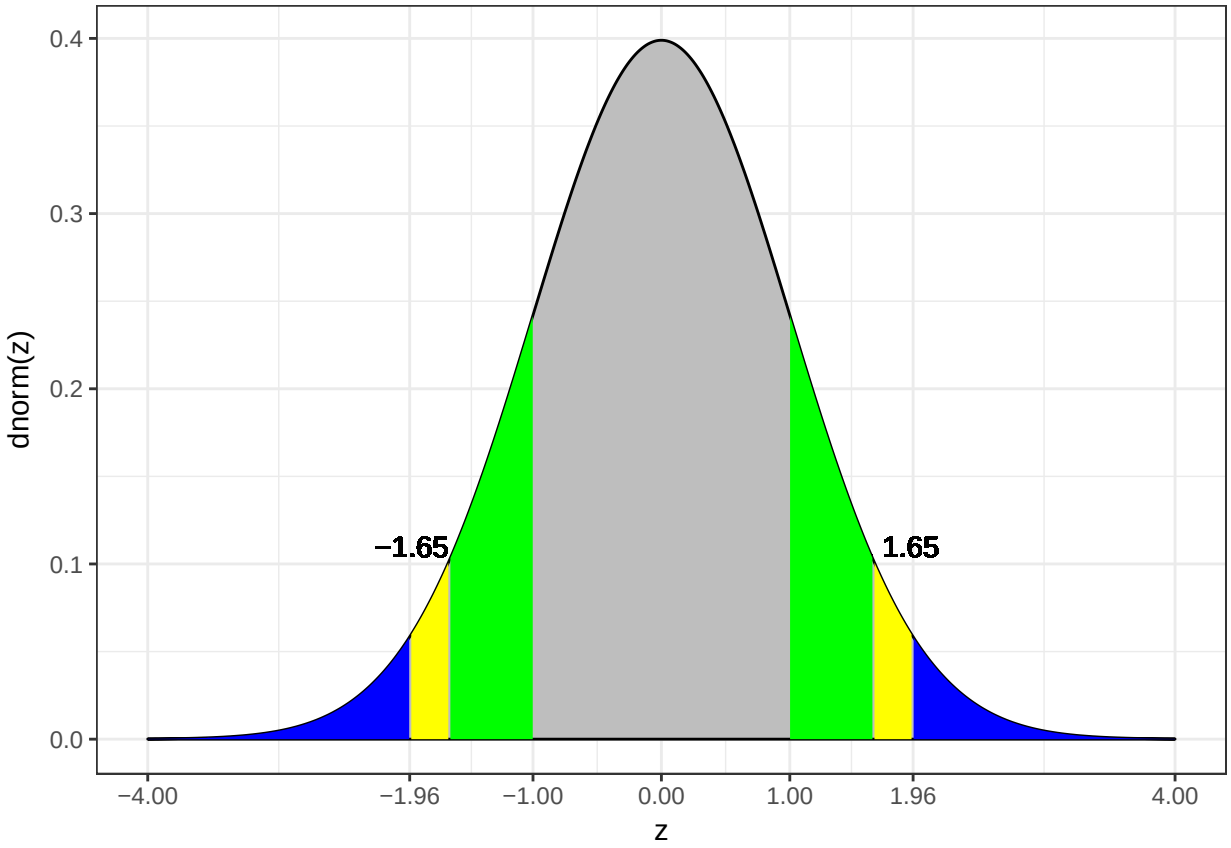


Figure 7.10: The z distribution

```
GA_m=mean(dataWBT$gen_att,na.rm = T)

#ss
GA_s=sd(dataWBT$gen_att,na.rm = T)

#95% güven aralığı
alt=GA_m - 1.96 * (GA_s/sqrt(GA_n))
alt
## [1] 1.92
ust=GA_m + 1.96 * (GA_s/sqrt(GA_n))
ust
## [1] 1.96

#veya
GA_m +c(-1,1)*1.96 * (GA_s/sqrt(GA_n))
## [1] 1.92 1.96


#1.96 için qnorm(0.975)
```

7.3.2.2 Ortalama için güven aralığı raporlama

5302 katılımcı için Toplumsal Cinsiyet Algısı puanlarına ait ortalama 1.94 (%95 GA [1.92-1.96]), standart sapma 0.60 bulunmuştur.

7.3.3 Boş Hipotez

Hipotez testinin amacı evren hakkında öne sürülen iki hipotezden hangisinin örneklem tarafından desteklendiğine karar vermektir. Bir hipotez testi beş basamaktan oluşur;

1) Boş hipotezin[3] belirlenmesi (örneğin $\mu = 0$)
2) Alternatif hipotezin belirlenmesi. (örneğin $\mu \neq 0$)
3) Test istatistiğinin seçilmesi
4) Test istatistiği ve belirlenen kritik değeri karşılaştırarak karar verilmesi. Eğer hesaplanan test istatistiği kritik değerden daha yüksekse boş hipotezin ret edilmesi (kritik değer alternatif hipoteze göre değişir).
5) Sonucun açıklanması. Araştırma sorusuna yanıt vermek üzere kararın ifade edilmesi.

Boş hipotez (H_0) ve alternatif hipotez (H_1) araştırma sorusunu cevaplamak üzere belirlenir. İstatistiksel kanıtlar boş hipotezini kabul veya terketmek için kullanılır. Boş hipotezi kabul etmek veya terketmek bir karardır, teorik istatistik açısından bu karara yönelik oluşabilecek durumlar şunlardır;

Gerçekte Durum	Karar	Sonuç
H_0	Kabul H_0	Doğru Karar
H_0	Red H_0	Yanlış Karar *(Tip I hata,α)*
H_1	Red H_0	Doğru Karar
H_1	Kabul H_0	Yanlış Karar *(Tip II hata, β)*

Tip-I hata: Gerçekte doğru olduğu halde boş hipotezin terkedilmesi durumudur. Alfa , α, boş hipotezin

[3] yokluk veya sıfır hipotezi olarak da bilinir

gerçekte doğru olduğu halde red edilme olasılığıdır. Hipotez testi sürecinde önemli olan noktalardan biri α'nın yeterince küçük olması gerektiğidir. Sosyal bilimlerde sıklıkla $\alpha = .05$ kullanılır.

Tip-II hata: Gerçekte yanlış olan bir boş hipotezin terkedilememisidir. Beta, β, gerçekte yanlış olan bir boş hipotezi kabul etme olasılığıdır.

7.3.4 z-puanı ve z-testi

z-puanı için genel formül;

$$z_X = \frac{X - \bar{X}}{s_X}$$

Hesaplanan bu z değişkeni 0 ortalamaya ve 1 standart sapmaya sahiptir. Eğer X normal dağılım gösteriyorsa z de normal dağılım gösterir.

```
# z puanı hesapla

GA_m=mean(dataWBT$gen_att,na.rm = T)
GA_s=sd(dataWBT$gen_att,na.rm = T)
z_GA=(dataWBT$gen_att-GA_m)/GA_s

#veya
z_GA=scale(dataWBT$gen_att, center=T, scale=T)

# scale fonksiyonu ile birden fazla değisken için z hesaplanabilir.
# center=T her X puanından ortalamayı çıkarır
# scale=T farkı standart sapmaya böler
# scale(dataWBT$gen_att, center=3, scale=2)her değerden 3 çıkarıp 2'ye böler.
```

Ortalama için z-istatistiği hesaplamak kolaydır;

$$z = \frac{\bar{X} - \mu_{hipotez}}{StandartHata} = \frac{\bar{X} - \mu_{hipotez}}{\sigma_X / \sqrt{n}}$$

Hesaplanan bu z istatistiği bir z dağılımı kullanılarak yorumlanabilir (Figür 7.10);

a) Eğer alternatif hipotez, gözlemlenen ortalamanın, hipotez değerinden küçük olacağını belirtiyorsa, hesaplanan z istatistiği z_{alpha} veya $-z_{(1-alpha)}$ ile kıyaslanır. Eğer hesaplanan z , z_{alpha}'ya eşit veya küçük ise boş hipotez terkedilir.

b) Eğer alternatif hipotez, gözlemlenen ortalamanın, hipotez değerinden farklı olacağını belirtiyorsa, hesaplanan z istatistiğinin mutlak değeri $z_{1-(alpha/2)}$ ile kıyaslanır. Eğer mutlak z , $z_{1-(alpha/2)}$'ya eşit veya büyük ise boş hipotez terkedilir.

c) Eğer alternatif hipotez, gözlemlenen ortalamanın, hipotez değerinden büyük olacağını belirtiyorsa, hesaplanan z istatistiği $z_{1-alpha}$ ile kıyaslanır. Eğer hesaplanan z , $z_{1-(alpha)}$'ya eşit veya büyük ise boş hipotez terkedilir.

Burada dikkat edilmesi gereken nokta a ve c senaryolarında (yönlü alternatif) kullanılan kritik değerin b senaryosunda (yönsüz alternatif) kullanılan kritik değerden farklı oluşudur. Araştırmacılar alternatif hipotezlerinin yönlü veya yönsüz oluşunu savunabilmelidir.

7.3.4.1 z testi örnek-1 (yönsüz)

Boş hipotez $H_0 : \mu_{CinsiyetAlgisi} = 2$ ve alternatif hipotez $H_1 : \mu_{CinsiyetAlgisi} \neq 2$ ve $\alpha = 0.05$;

```
#  n
GA_n=sum(!is.na(dataWBT$gen_att))

#ortalama
GA_m=mean(dataWBT$gen_att,na.rm = T)

#ss
GA_s=sd(dataWBT$gen_att,na.rm = T)

# boş hipotez
mu_hyp=2

# z istatistiği
(GA_m-mu_hyp)/(GA_s/sqrt(GA_n))
## [1] -7.17

#alpha=0.05 ve yönsüz alternatif için kritik değer
qnorm(1-(0.05/2))
## [1] 1.96
```

5302 katılımcıya ait Toplumsal Cinsiyet Algısı puanları için ortalama 1.94 ve standart sampma 0.6 olarak hesaplanmıştır. Tek örneklem için hesaplanan z testi, gözlemlenen ortalamanın, hipotez ile öne sürülen 2'den 7.17 standart hata daha düşük olduğunu göstermiştir. Kritik değer olarak 1.96 ($z_{1-(0.05/2)}$) seçildiğinde ve gözlemlenen ortalama ve hipotez ile öne sürülen ortalama arasındaki farkın istatistiksel olarak anlamlı olduğu kararı verilmiştir.

7.3.4.2 z testi örnek-2 (yönlü)

Bu örnekte Toplumsal Cinsiyet Algısı değişkeninin evren bazında ortalaması 1.9 ve standart sapması 0.75 olarak varsayılmıştır. Eğer evren bazında standart sapma biliniyorsa istatistik hesaplarken kullanılmalıdır. $H_0 : \mu_{CinsiyetAlgisi} = 1.9$ ve $H_1 : \mu_{CinsiyetAlgisi} > 1.9$ ve $\alpha = 0.01$;

```
# boş hipotez
mu_hyp=1.9

# z istatistik
(GA_m-mu_hyp)/(0.75/sqrt(GA_n))
## [1] 3.94

#yönlü alternatif ve alfa=.01
qnorm(1-(0.01))
## [1] 2.33
```

Kritik değer olarak 2.33 kullanıldığında ($z_{0.99}$) gözlemlenen ortalamanın hipotez ile öne sürülen ortalamadan büyük olduğu ve bu farkın istatistiksel olarak anlamlı olduğu kararına varılmıştır $z = 3.94$.

7.3.5 Tek örneklem t-testi

Evrene ait dağılım normal olsa dahi küçük örneklemler için z dağılımı yerine t dağılımı kullanmak daha geçerlidir. bu t dağılımının serbestlik derecesi n-1'dir.

7.3.5.1 t test örnek-1 (yönsüz)

Örnek için Düzce ilinde yaşadığını belirten katılımcılara ait Toplumsal Cinsiyet Algısı puanları kullanılmıştır.

$H_0 : \mu_{CinsiyetAlgisi} = 1.94$ ve alternatif $H_1 : \mu_{CinsiyetAlgisi} \neq 1.94$ ve $\alpha = 0.05$;

```
dataWBT_DUZCE=dataWBT[dataWBT$city=="DUZCE",]
#betimleyici
describe(dataWBT_DUZCE[,"gen_att"],type=3)
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 47 2.18 0.55      2    2.14 0.59   1 3.8   2.8 0.56     0.28 0.08

#t test
t.test(dataWBT_DUZCE$gen_att,
       alternative="two.sided",
       mu=1.94,
       conf.level = 0.95)
##
##  One Sample t-test
##
## data:  dataWBT_DUZCE$gen_att
## t = 3, df = 50, p-value = 0.005
## alternative hypothesis: true mean is not equal to 1.94
## 95 percent confidence interval:
##  2.01 2.34
## sample estimates:
## mean of x
##      2.18

#kritik değer
qt(.975,df=46)
## [1] 2.01
```

Düzce ilinde yaşayan katılımcıların Cinsiyet Algısı puanları 1 ve 3.8 arasında değişmiştir, ortanca 2, ortalama 2.18, standart sapma 0.55, örnekleme ait dağılımın çarpıklığı 0.56 ve basıklığı 0.28 olarak hesaplanmıştır. Düzce şehrinde yaşayan katılımcılara ait Toplumsal Cinsiyet Algısı ortalaması, hipotez ile öne sürülen 1.94 değerinden farklıdır ve bu farklılık istatistiksel olarak anlamlıdır, t(46)=2.94 ve $t_{.975,46} = 2.01$

7.3.5.2 t test örnek-2

$H_0 : \mu_{CinsiyetAlgisi} = 1.94$ ve alternatif $H_1 : \mu_{CinsiyetAlgisi} \leq 1.94$ ve $\alpha = 0.05$;

```
#t test
t.test(dataWBT_DUZCE$gen_att,
       alternative="less",
       mu=1.94,
       conf.level = 0.95)
##
##  One Sample t-test
##
## data:  dataWBT_DUZCE$gen_att
## t = 3, df = 50, p-value = 1
## alternative hypothesis: true mean is less than 1.94
## 95 percent confidence interval:
##  -Inf 2.31
```

```
## sample estimates:
## mean of x
##      2.18

#kritik  değer
qt(.05,df=46)
## [1] -1.68
```

Test istatistiği t(46)=2.94 ve kritik değer $t_{.05,46} = -1.68$ kullanılarak örneklemin, evrene ait ortalamanın 1.94'ten küçük olduğu hipotezini destekleyecek kanıtı içermediği kararı verilmiştir.

7.3.6 p-değeri

Tek örneklem ile t testi için kullanılan t.test fonksiyonu bir p değeri rapor etmiştir. Bu p değeri hesabı, boş hipotezin ve dağılım için yapılan varsayımın doğru olduğu kabulü üzerine yapılır. Bu p-değerinin amacı araştıracıyı hesaplanan istatistiğin sıradan olup olmadığı yönünde bilgilendirmektir. Uzun yıllardır araştırmacılar hesapladıkları bu p-değerini daha önceden belirledikleri bir alfa kriteri ile kıyaslayıp, buldukları sonuçların istatistiksel olarak anlamlı olup olmadığına karar vermişlerdir.

7.3.7 p değeri örnek-1

Bir z-dağılımın geçerli olduğu ve z değerinin 1.80 hesaplandığı durum için grafik;

Mavi alan yoğunluğun %3.6'sını gösterir, p=0.0359

```
1-pnorm(1.8)
```

```
## [1] 0.0359
```

Bu p değeri yönlü bir alternatif hipotez için hesaplanmıştır. Yönsüz alternatif için geçerli değildir. Yönsüz alternatif için ;

Mavi alan yoğunluğun %7.2'sini gösterir, p=0.0719;

```
2*(1-pnorm(1.8))
```

```
## [1] 0.0719
```

7.3.8 İstatiksel Güç

İstatistiksel güç, gerçekte yanlış olan bir boş hipotezi terketme olasılığıdır ve $(1-\beta)'ya$ eşittir. Bu olasılık istatistiksel sınamalar yapıldıktan önce (a-priori) veya sonra (post-hoc) hesaplanabilir, fakat sınama sonrasında yapılan güç analizi genellikle işlevsizdir. Sınama yapılmadan önce, daha doğrusu veriler toplanmadan önce yapılacak bir güç hesabı ile çalışma tasarısı gözden geçirebilir, yeniden düzenlenebilir ve örneklem sayısı belirlenebilir. Amerika'da bir çok proje başvurusu istatistiksel güç analizlerini mecbur tutar.

İstatisksel gücü açıklamak için R ile çizilen grafik ve kod[4];

```
x <- seq(-4, 8, 0.02)
zdat <- data.frame(x = x, y1 = dnorm(x, 0, 1), y2 = dnorm(x, 2.5, 1))
ggplot(zdat, aes(x = x)) +
  geom_line(aes(y = y1), size=2) +
  geom_line(aes(y = y2), color='red',size=2) +
  geom_vline(xintercept = c(0,2.5), color="black", linetype = "longdash")+
```

[4]partially based on http://multithreaded.stitchfix.com/blog/2015/05/26/significant-sample/

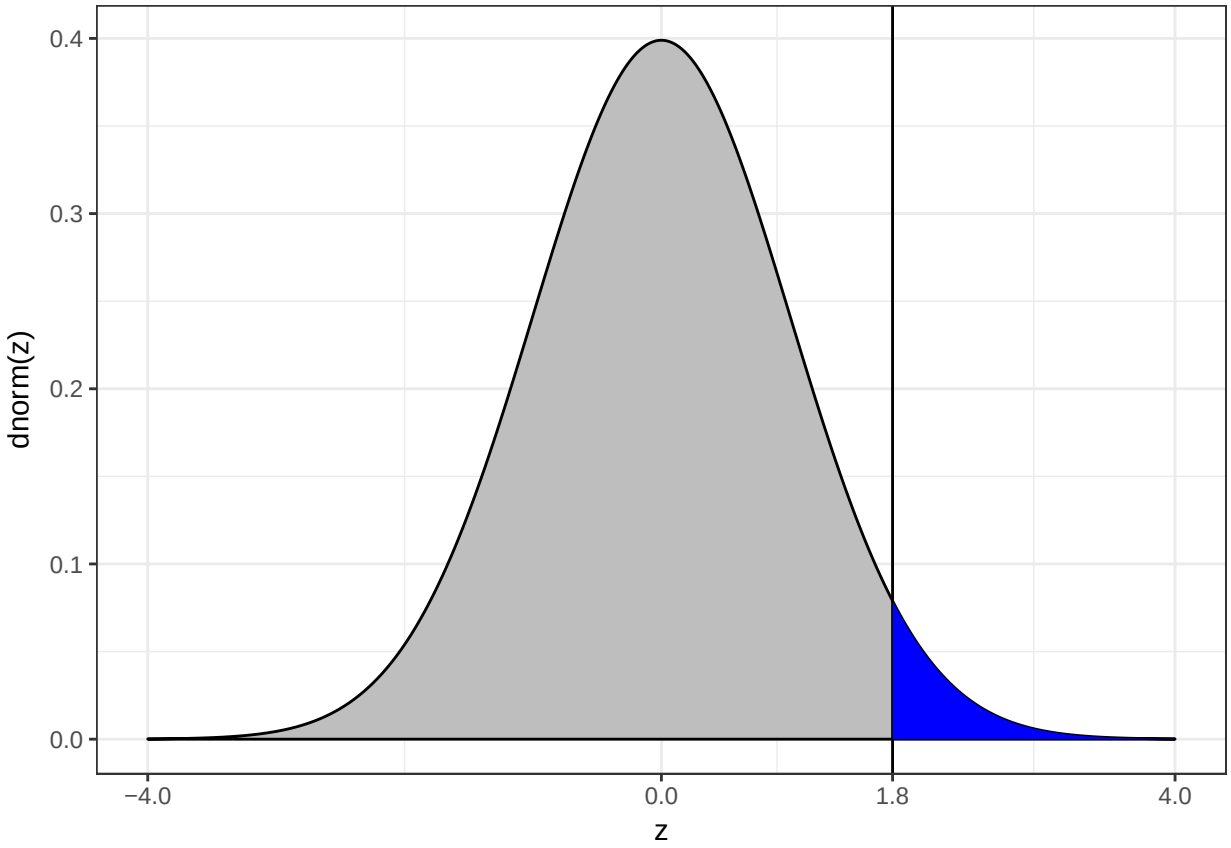


Figure 7.11: z dagilimi ve z=1.8

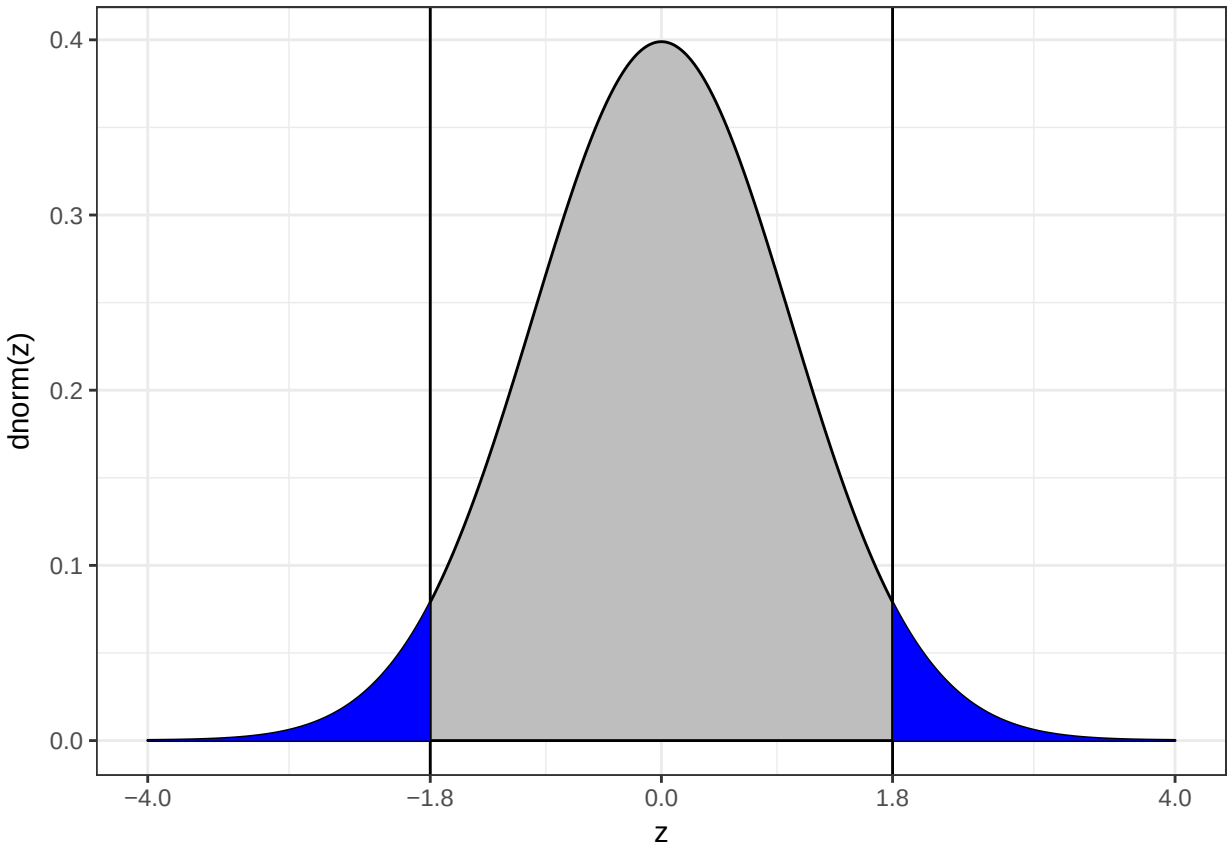


Figure 7.12: z dagilimi ve |z|)=1.8

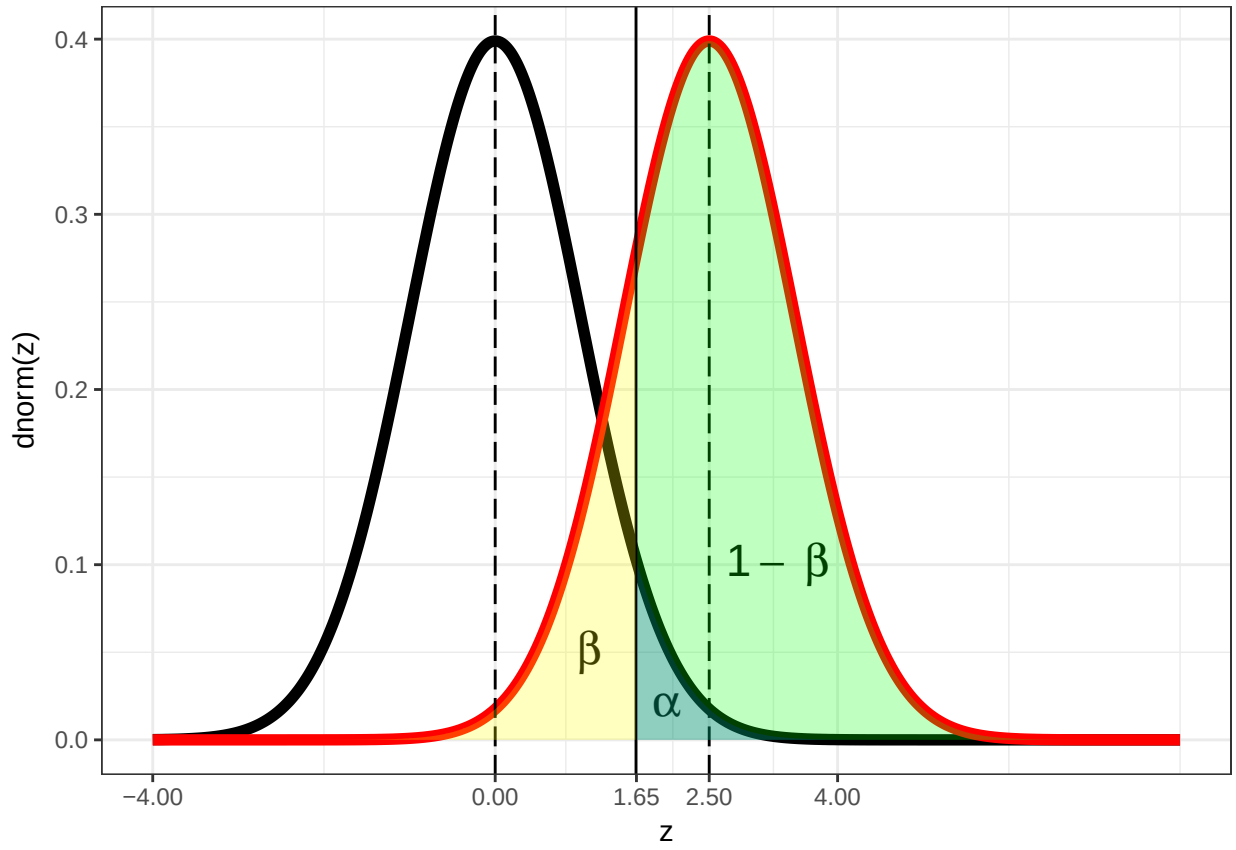


Figure 7.13: z dagilimi ile istatistiksel güç

```
  geom_vline(xintercept = qnorm(1 - 0.05))+
  scale_x_continuous(breaks = c(-4,0,1.65,2.5,4))+
  annotate("text", label="beta" , x=1.1, y=0.05, parse=T, fontface =2, size=6)+
  annotate("text", label="alpha", x=2  , y=0.02, parse=T, fontface =2, size=6)+
  annotate("text", label="1-~beta", x=3.3, y=0.1, parse=T, fontface =2,size=6)+
  geom_area(aes(y=y1, x = ifelse(x > qnorm(.95), x, NA)), fill = 'blue'  , alpha=0.25) +
  geom_area(aes(y=y2, x = ifelse(x > qnorm(.95), x, NA)), fill = 'green' , alpha=0.25) +
  geom_area(aes(y=y2, x = ifelse(x < qnorm(.95), x, NA)), fill = 'yellow', alpha=0.25) +
  xlab("z") + ylab("dnorm(z)") + theme_bw()
```

$H_0 : \mu = 0$ boş hipotezini doğru kabul eden bir z dağılımı siyah çizgiler ile gösterilmiştir, bu dağılımın ortalaması sıfırdır ve kesik çizgiler ile gösterilmiştir. Kırmızı çizgiler ile gösterilen dağılım (a) boş hipotezin yanlış olduğu ve (b) evrene ait ortalama ve standard sapma ile hesaplanan z-istatistiğinin $((\mu - \mu_{hypothesis})(\sigma\sqrt{n}) = 2.5)$ olduğu varsayımı ile çizilmiştir. Bu görsel yönlü bir alternatif hipotez ve alfa=0.05, dolayısıyla kritik değer $z_{0.95} = 1.65$ için geçerlidir. Mavi alan α'yı, sarı alan β'yı ve yeşil alan istatistiksel gücü yansıtır. Bu görselde istatistiksel güç .804'tür.

```
1-pnorm(qnorm(0.95),mean=2.5)
```

```
## [1] 0.804
```

Figüre[5] 7.13 , istatistiksel güç hesaplamak için iki farklı dağılımın, bir alfa değerinin ve bir test istatistiğinin olması gerektiğini gösterir. Bu bilinenler ile istatistiksel güç hesaplanabilir. Burda önemli olan detay bir

[5]accurate only for post hoc power

test istatistiğinin kendi bileşenleri olduğudur, genellikle bu bileşenler bir bölünen ve bir bölendir. z testi için bölünen , hipotez ile öne sürülen değer ile gözlemlenen değer arasındaki fark, bölen ise ortalamanın standart hatasıdır ($\sigma/\sqrt{(n)}$). Eğer istatistiksel güç sabit tutulursa (örneğin .8) eşitlik seçilen bir bilinmeyene göre çözülebilir. Genellikle de seçilen bilinmeyen örneklem sayısıdır, n.

İstatistiksel güç ilerleyen bölümlerde tekrar değinilmiştir. Test istatistikleri, parametre kestirimleri, standart hatalar, dağılımsal varsayımlar tasarlanan araştırmaya göre değişecektir. Tek örneklem için t testi düşünüldüğünde *power.t.test* fonksiyonu işeyarardır.

```
#power.t.test
power.t.test(delta=.1, sd=.6,sig.level=0.05, power=0.9,
             type="one.sample", alternative="one.sided")
##
##      One-sample t test power calculation
##
##               n = 310
##           delta = 0.1
##              sd = 0.6
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
```

Bu örnek, belirlenmiş bir ortalama fark 0.1, standart sapma 0.6, alfa 0.05, yönlü alternatif ve istenilen güç 0.9 için örneklemin 310 olması gerektiğini gösterir. Bir diğer ifade ile, araştırmacı 310 kişiden gelen bir veride, ortalama fark 0.1 , standart sapme 0.6 tespıt eder ve alfa 0.05 ile yönlü bir test kullanırsa boş hipotezi ($H_0 : \mu = 0$) terketme olasılığı 0.90'dır.

7.3.9 z ve t dağılımları geçerli değil ise

Bilinenlerden (örneklem) bilinmeyenlere (evren) genelleme varsayımların yapılmasını gerektirir. Bir test istatistiğine ait örnekleme dağılımın, belli bir örneklem sayısında, belirli bir varsayımın ihlal edilmesi ile büyük ölçüde değişmemesi durumu *direnç* (robustness) olarak isimlendirilir (Verzani (2014)). Burda dikkat edilmesi gereken, bir test istatitiğinin bir varsayım ihlaline dirençli iken başka bir varsayım ihlaline dirençsiz olabileceğidir. Ayrıca, bir varsayım ihlaline dirençli olan bir test istatistiği, ikinci bir varsayım ihlalinin de yaşanması durumunda direncini yitirebilir. Bir test istatistiği dirençli olduğu için kullanılması şart değildir çünkü aynı şartlar altında daha iyi çalışan bir başka test istatistiği olabilir.

z istatistiği, örneklem 30'dan büyük ise normallik varsayımının ihlallerine karşı dirençlidir(Field et al. (2012), page 198). Örnekleme dağılımın z dağılımına yakınlığını etkileyen bir faktör diğer faktörde örneklemin nasıl bir dağılım gösterdiğidir. Ayrıca, evrene ait dağılımın normal olduğu varsayımı ile küçük örneklemler için t dağılımı geçerlidir.

Tek örneklem ile aritmetik ortalama için yapılacak dirençli istatistikler detaylı olarak Wilcox (2012) tarafından verilmiştir. Verilen R kodu bootstrap-t metodu içim %95 güven aralığı hesaplar (Wilcox (2012), page 117).

```
#ikinci tür bootstrap t metodu
# Düzce katılımcılarını seç
dataWBT_DUZCE=na.omit(dataWBT[dataWBT$city=="DUZCE",c("id","gen_att")])

# normallik varsayımı ve t test kullanarak
# evren ortalamasının 1.94 olup olmadığını sına
t.test(dataWBT_DUZCE$gen_att,mu=1.94,conf.level = 0.95)
##
##  One Sample t-test
##
```

```
## data:  dataWBT_DUZCE$gen_att
## t = 3, df = 50, p-value = 0.005
## alternative hypothesis: true mean is not equal to 1.94
## 95 percent confidence interval:
##  2.01 2.34
## sample estimates:
## mean of x
##      2.18
```

```
# bootstrap ile 95% GA  (normallik varsayımı yok)
set.seed(04012017)
B=5000       # bootstrap sayısı
alpha=0.05   # alfa

#x değişken
# xBAR gözlemlenen ortalama
tstar=function(x,xBAR) sqrt(length(x))*abs(mean(x)-xBAR)/sd(x)


output=c()
for (i in 1:B){
  output[i]=tstar(sample(dataWBT_DUZCE$gen_att,
                         replace=T,
                         size=length(dataWBT_DUZCE$gen_att)),
                  xBAR=mean(dataWBT_DUZCE$gen_att))
  }
output=sort(output)
Tc=output[as.integer(B*(1-alpha))]

#bootstrap GA
mean(dataWBT_DUZCE$gen_att)+c(-1,1)*(Tc*sd(dataWBT_DUZCE$gen_att)/sqrt(length(dataWBT_DUZCE$gen_att)))
## [1] 2.01 2.34
```

7.3.9.1 Raporlama

Düzce ilinden 47 katılımcının verdiği yanıtlar ile hesaplanan Toplumsal Cinsiyet Algısı puanları 1 ve 3.8 arasında değişmiş, ortancası 2, ortalaması 2.18, standart sapması 0.55 bulunmuştur. Puanların dağılımına ait çarpıklık değeri 0.56, basıklık değeri 0.28 olarak hesaplanmıştır. Kritik değer olarak 2.01 ($t_{.975,46}$) kullanıldığında, tek örneklem t testi anlamlı bir farklılığa işaret etmiştir,t(46)=2.94, bu şehirdeki katılımcıların puanları hipotez ile öne sürülen 1.94 değerinden farklıdır. Normallik varsayımı yapıldığında %95 güven aralığı [2.01,2.34] olarak bulunmuştur. Normallik varsayımı yapılmadığıda 5000 tekrarlı bootstrap metodu ile hesaplanan güven aralığı [2.01,2.34] olarak bulunmuştur.

Chapter 8

İki Ortalamanın Karşılaştırılması, t-testi

Bölüm 7.3.1 örnekleme dağılım (sampling distribution) konusunun ana hatlarını tek bir aritmetik ortalama üzerinden ele almıştır. Eğer iki farklı aritmetik ortalama kıyaslanmak isteniyorsa t testi kullanılabilir, bu prosedür aritmetik ortalamaların örnekleme dağılımı üzerine inşaa edilmiştir.

$\bar{Y_1} - \bar{Y_2}$ ($\mu_{\bar{Y_1}-\bar{Y_2}}$) sonucunun örnekleme dağılım ortalaması daima $\mu_1 - \mu_2$ 'dir. Fakat örnekleme dağılım standart sapması ($\sigma_{\bar{Y_1}-\bar{Y_2}}$) araştırmanın tasarısına göre değişir.

Örnek Bir cerrahın yaraların iyileşmesi konusunda araştırma yaptığını düşünelim. Cerrahın kapanan bir yaradan sonra gerilme direncini araştırdığını, yara bandı ve dikiş atma tedavisinin arasında bir fark olup olmadığını araştırdığını varsayalım. Bu çalışmada tek bir faktör vardır, yara kapatma yöntemi ve bu faktöre ait iki alt sınıf vardır, yara bandı ve dikiş. Bu araştırmayı iki farklı şekilde tasarlamak mümkündür.

Bağlı gözlemler (within-subjects) 10 tavşanın her birinin sırtına (omurganın sağına ve soluna) 2 kesik oluşturulur. 2 kesikten bir tanesi yeni geliştirilen bir yara bandı ile diğeri dikiş ile kapatılır, hangi kapatma yönteminin hangi yarayı kapatacağı rassal (random) seçilmelidir. Bu tasarı *bağlı-gözlemler* olarak isimlendirilmiştir çünkü faktöre ait iki alt sınıf aynı tavşan üzerinde gözlemlenmiştir.

Bağlı olmayan gözlemler (between-subjects) 20 tavşan rassal olarak 2 gruba ayrılır, birinci grupta yer alan tavşanların yaraları bant ile, ikinci grupta yer alan tavşanların yaraları ise dikiş ile kapatılır. Yaralar omurganın sağ veya sol tarafında rassal olarak açılmalıdır. Bu tasarı *bağlı olmayan gözlemler* olarak isimlendirilmiştir çünkü faktöre ait alt sınıflar farklı tavşanlar üzerinde gözlemlenmiştir ve bu tavşanların herhangi bir şekilde eşlenmiş değillerdir. Örneğin aynı anneden gelen iki tavşan rassal olarak gruplara atansa idi gözlemler bağlı olurdu.

Her iki yara kapatma yönteminden sonra yapılacak gerilme direnci ölçümlerinin evren bazında bir ortalaması ve standart sapması olduğu aşikardır.

Konuyu açıklama amaçlı, yara bandı yönteminden sonra yapılan gerilme direnci ölçümlerine ait ortalamanın ve standart sapmanın bağlı gözlem veya bağlı olmayan gözlem tasarılarında aynı olduğunu düşünelim.

Parametres	Bant	Dikiş
Ortalama	μ_B	μ_D
Standart Sapma	σ_B	σ_D
Örneklem	n_B	n_D

Buradan itibaren μ_B yerine μ_1, μ_D yerine μ_2, σ_B yerine σ_1,σ_D yerine σ_2 kullanılmıştır.

Örnekleme dağılım parametresi	Bağlı olmayan gözlem	Bağlı gözlem
Ortalama ($\mu_{\bar{Y}_1-\bar{Y}_2}$)	$\mu_1 - \mu_2$	$\mu_1 - \mu_2$
Standart sapma ($\sigma_{\bar{Y}_1-\bar{Y}_2}$)	$\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}$	$\sqrt{\frac{\sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2\rho_{12}}{n}}$

1. ρ_{12} bağlı gözlemlerde, bant ve dikiş sonrası yapılan ölçümlerin arasındakı korelasyondur.
2. Standart sapmadaki değişiklik ρ_{12}'den kaynaklanır. Eğer bu korelasyon 0 ise iki tasarının da örneklem dağılımına ait standart sapma aynıdır.

Bir araştırma tasarısında hedeflerden biri standart hatayı mümkün olduğunca küçük tutmaktır. Standart hatanın küçük olması elde edilen test istatistiğinin tahmin edilen parametreye yakın olduğu anlamına gelir.

Veri çözümleme sürecinde hata varyansı (error variance) hesaplamak için bir formül seçilir. Yanlış formülün kullanılması büyük bir hatadır.

Uygulamada, standart hatanın hesaplanışı tasarının bağlı mı bağsız mı olduğuna göre değişir. Tasarının yanlış sınıflandırılması, çözümleme sürecinde büyük bir hatadır.

8.1 Bağımsız gruplar t-test (The Independent Groups t-test)

Uşak ilinde yaşayan katılımcıların Toplumsal Cinsiyet Algısı (TCA) puanları yüksek öğretim durumuna göre karşılaştırılmıltır (2.3). Her grup için yoğunluk grafikleri;

```
# csv yükle
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#URL sil
rm(urlfile)
dataWBT_USAK=dataWBT[dataWBT$city=="USAK",]

# factor ve droplevels fonksiyonları bölüm 5.2.4 ile verilmiştir.
# yeni oluşturulan HEF (Higher Education Factor)
# katılımcı lise veya altı diplomaya sahipse 0, non-college
# katılımcı lise üstü diplomaya sahip ise 1,  college
dataWBT_USAK$HEF=droplevels(factor(dataWBT_USAK$higher_ed,
                    levels = c(0,1),
                    labels = c("non-college", "college")))

require(ggplot2)
plotdata=na.omit(dataWBT_USAK[,c("gen_att","HEF")])
ggplot(plotdata, aes(x = gen_att)) +
  geom_histogram(aes(y = ..density..),col="black",binwidth = 0.2,alpha=0.7) +
  geom_density(size=2) +
  theme_bw()+labs(x = "Uşak ilinde Yüksek Öğretim Durumuna göre TCA puanları")+ facet_wrap(~ HEF)+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

8.1.1 R betiği: Bağımsız gruplar t testi

Takip edilen basamaklar;

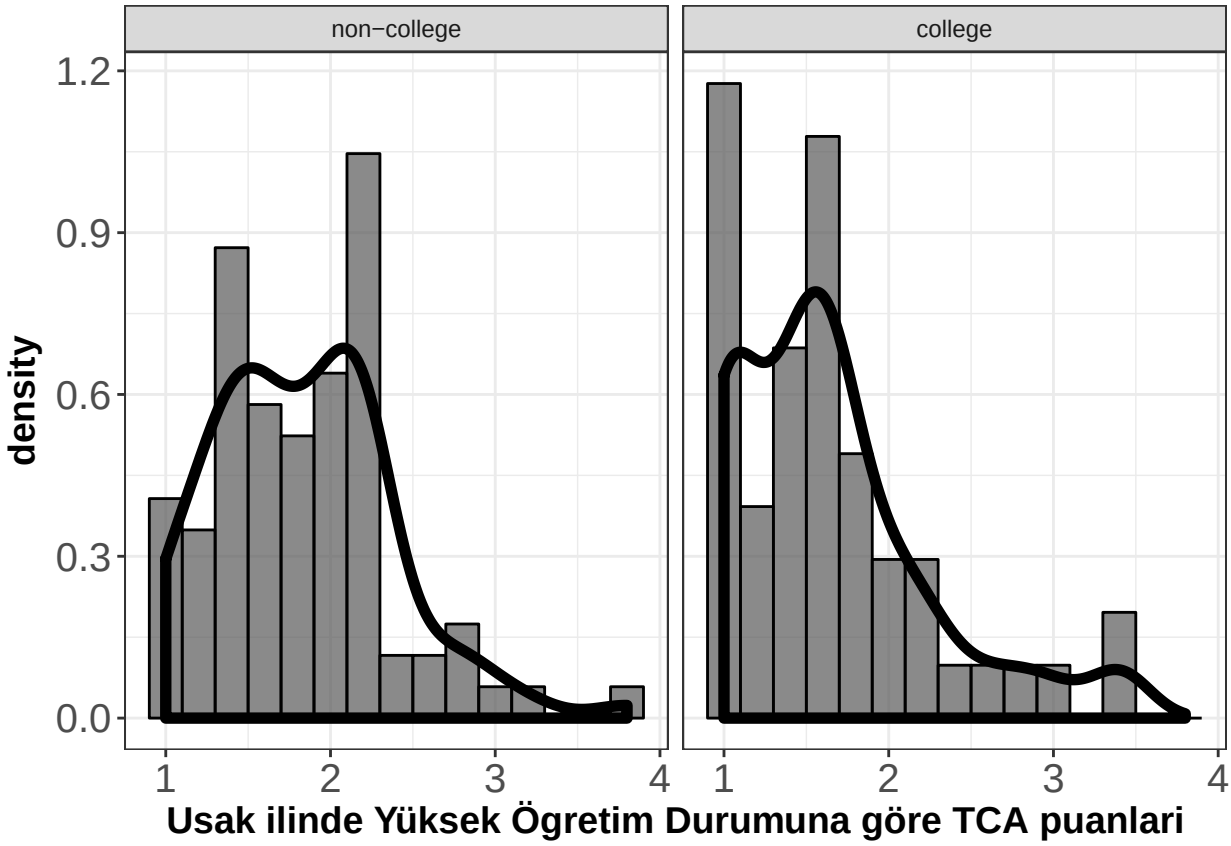


Figure 8.1: Yüksek Ögretim Durumuna göre TCA puanlari

1. Betimsel istatistikler
2. Test istatistiğinin hesabı

$$t = \frac{\bar{Y_1} - \bar{Y_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

3. Kritik değerin hesabı $\pm t_{\alpha/2, n_1+n_2-2}$

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

```
library(psych)
descIDT=with(dataWBT_USAK,describeBy(gen_att, HEF,mat=T,digits = 2))
descIDT
##     item      group1 vars  n mean   sd median trimmed  mad min max range
## X11    1 non-college    1 86 1.83 0.54    1.8    1.80 0.59   1 3.8   2.8
## X12    2     college    1 51 1.64 0.61    1.6    1.54 0.59   1 3.4   2.4
##     skew kurtosis   se
## X11 0.72     0.90 0.06
## X12 1.19     1.09 0.09
# rapor etmek için
#write.csv(descIDT,file="independent_t_test_desc.csv")
#türkçe excel için
# #write.csv2(descIDT,file="independent_t_test_desc.csv")


# ss
sp=sqrt((85*.543^2 + 50*.608^2)/(86+51-2))

# t-istatistik
tstatistic=(1.832-1.635)/(sp*sqrt(1/86+1/51))

# alfa=0.05 kritik değer
qt(.975,df=135)
## [1] 1.98
```

1.963 kritik değer $t_{.975,135} = 1.978$'den küçük olduğu için, H_0 kabul edilir.

$H_1 : \mu_1 - \mu_2 > 0$ alternatif hipotezi kurulsa idi, kritik değer $t_{.95,135} = 1.66$, 1.963'ten küçük olduğu için H_0 terkedilirdi.

$H_1 : \mu_1 - \mu_2 < 0$ alternatif hipotezi kurulsa idi, 1.963 kritik değer olan $t_{.05,135} = -1.66$' ten küçük olmadığı için H_0 kabul edilirdi (Burada alternatif hipotez ile örneklem arası farklılık zıt yönde).

Daha kullanışlı bir R betiği;

```
# dataWBT HEF faktörünü içermez, yukarıda HEF faktörü oluşturulmuştur.

t.test(gen_att~HEF,data=dataWBT_USAK,var.equal=T,
                                     alternative="two.sided",
                                     conf.level=0.95)
##
```

```
##  Two Sample t-test
##
## data:  gen_att by HEF
## t = 2, df = 100, p-value = 0.05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0019  0.3949
## sample estimates:
## mean in group non-college     mean in group college
##                      1.83                      1.64

# büyüktür
t.test(gen_att~HEF,data=dataWBT_USAK,var.equal=T,
                                    alternative="greater",
                                    conf.level=0.95)
##
##  Two Sample t-test
##
## data:  gen_att by HEF
## t = 2, df = 100, p-value = 0.03
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0303    Inf
## sample estimates:
## mean in group non-college     mean in group college
##                      1.83                      1.64


# küçüktür
t.test(gen_att~HEF,data=dataWBT_USAK,var.equal=T,
                                    alternative="less",
                                    conf.level=0.95)
##
##  Two Sample t-test
##
## data:  gen_att by HEF
## t = 2, df = 100, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##   -Inf 0.363
## sample estimates:
## mean in group non-college     mean in group college
##                      1.83                      1.64
```

8.1.1.1 Yönsüz alternatif için rapor örneği

Uşak ilinde yaşayan katılımcılardan yüksek öğretim diplomasına sahip olan 51 kişi için TCA puanları ortalaması 1.64, standart sapması 0.61, puanlara ait dağılım çarpıklık değeri 1.19 ve basıklık değeri 1.09 olarak bulunmuştur. Aynı ilde yüksek öğretim diploması olmayan 86 katılımcının ise TCA puanları ortalaması 1.83, standart sapması 0.54, puanlara ait dağılımın çarpıklığı 0.72 ve basıklığı 0.90 olarak hesaplanmıştır.Bağımsız gruplar t-testi sonuçları, Uşak ilinde yaşayan katılımcıların Toplumsal Cinsiyet Algısı puanlarının yüksek öğretim durumuna göre değişeceği tezini desteklememiştir, t(135)=1.96, p=0.052. Puanlar arasındaki fark

için %95 güven aralığı [-0.002,0.395] olarak hesaplanmıştır.[1]

8.1.1.2 Yönlü alternatif için rapor örneği

Uşak ilinde yaşayan katılımcılardan yüksek öğretim diplomasına sahip olan 51 kişi için TCA puanları ortalaması 1.64, standart sapması 0.61, puanlara ait dağılım çarpıklık değeri 1.19 ve basıklık değeri 1.09 olarak bulunmuştur. Aynı ilde yüksek öğretim diploması olmayan 86 katılımcının ise TCA puanları ortalaması 1.83, standart sapması 0.54, puanlara ait dağılımın çarpıklığı 0.72 ve basıklığı 0.90 olarak hesaplanmıştır. Bağımsız gruplar t testi sonuçları, yüksek öğretim diploması olmayanların TCA puanları yüksek öğretimlere nazaran daha yüksektir tezini desteklemiştir, t(135)=1.96, p=0.026. Puanlar arasındaki fark için %95 güven aralığı [0.030,∞] olarak hesaplanmıştır.

8.1.2 Varsayımlar: Bağımsız gruplar t testi

Geleneksel t-testi sonuçlarının geçerliği 3 varsayımın ihlal edilmemesi ile mümkündür.

1. Yanıtların bağımsızlığı (independence). Her gruba ait puanların dağılımı birbirinden bağımsız olmalıdır. Yanıtların bağımsızlığını tehdit eden durumlardan biri aynı grup içeresinde yer alan bireylerin birbirlerinin yanıtlarını etkilemesidir (Yantların bağımsızlığı 9.2.1.4 bölümünde daha detaylı ele alınmıştır).

2. Normallik. Her gruba ait puanlar normal bir dağılımdan çekilmiştir. Myers et al. (2013) grupların örneklem sayısı (n) eşit olduğunda ve toplam örneklemin 40 veya daha fazla olduğu durumlarda t test istatistiğinin normallik varsayımı ihlallerine dirençli olduğunu belirtmiştir. Fakat bu direnç normal dağılımdan büyük çaplı sapmalar (extreme) için geçerli değildir. Bu kitabın yazarları normallik testlerinin bu varsayımı kontrol etmek için kullanılmasına sıcak bakmamaktadır. Görsel bir değerlendirmeden sonra özellikle küçük örneklemlerde, normallik varsayımının ihlal edildiği kaygısı varsa araştırmacılar dirençli tahminleme yöntemlerini kullanmalıdırlar.

3. Eş varyanslılık. Varyans homojenliği olarak da bilinen bu varsayım, iki grupta yer alan puanların evren bazında eşit varyanslı dağılımlardan çekildiğini kabul eder. Myers et al. (2013) grup örneklem sayılarının **eşit** ve en az 5 olduğu durumlarda varyans eşitliği sağlanamasa dahi 1. tip hata oranlarının önemli ölçüde değişmeyeceğini belirtmiştir. Fakat bu direnç büyük çaplı heterojenlikler için geçerli değildir, örnğin $s_1^2/s_2^2 > 100$. Field et al. (2012) Levene testi gibi eşvaryanslılık testlerinin örneklem sayılarının eşit olmadığı ve küçük örneklemlerde sağlıklı sonuçlar vermediğini belirtmiştir, halbuki bu testlere en çok ihtiyaç durumlar küçük örneklem ve eşit olmayan örneklem sayısı durumlarıdır. Eş varyanslılık testleri ile ilgili bir diğer kaygı, bu test sonucunda varyansların eşit sayılabileceği kararı verilmiş olsa da varyanslar arası farklılığın t test istatistiğini etkileyebileceğidir. *t.test* fonksiyonu , aksi istenmediği sürece, varyans eşitliği varsayımı yapmaz ve Welch t testini hesaplar.

Varsayımlar hakkındaki bu kısa tanıtımın ele almadığı durumlar vardır. Örneğin hem normalliğin hem eş varyanslılığın ihlal edildiği durumlar tartışılmamıştır. Ayrıca direnç konusu tartışmaya açıktır. Önreğin n1=n2=10 örneklem sayısı ve eş olmayan varyans durumu için 100000 tekrarla yaptığımız bir simülasyon, alfa=.01 ve yönsüz bir t testi için 1. tip hata oranını .018 bulmuştur. Bu oranın kabul edilebilir olup olmadığı tartışmaya açıktır.

Sonuç olarak, eğer yanıtların bağımsızlığı kabul ediliyorsa, örneklem sayısı eşitse ve her grupta en az 20 ise bağımsız gruplar t istatistiğinin büyük ölçüde dirençli olduğu kabulu makul bir kabuldur. Diğer durumlarda varsayım ihlalleri tespit edildi ise araştırmacılar alternatif çözümleme yöntemlerini kullanabilirler.

[1] Çözümlemeler R programlama dili kullanılarak tamamlanmıştır, betimsel istatistikler *psych* paketi(Revelle, 2016), t test istatistiği ise stats paketi (R Core Team, 2016b) ile hesaplanmıştır.

8.1.3 Welch t test

Normalliğin ciddi ölçüde zedelenmediği ve örneklem sayılarının her grup için en az 20 olduğu durumlarda Welch testi geçerli sonuçlar üretir. Bu test varyans eşdeğerli varsayımı yapmaz.

```
t.test(gen_att~HEF,data=dataWBT_USAK,var.equal=F,
                                     alternative="two.sided",
                                     conf.level=0.95)
##
##  Welch Two Sample t-test
##
## data:  gen_att by HEF
## t = 2, df = 100, p-value = 0.06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.00848  0.40146
## sample estimates:
## mean in group non-college     mean in group college
##                      1.83                      1.64
```

8.1.3.1 Raporlama örneği: Welch t testi

Uşak ilinde yaşayan katılımcılardan yüksek öğretim diplomasına sahip olan 51 kişi için TCA puanları ortalaması 1.64, standart sapması 0.61, puanlara ait dağılım çarpıklık değeri 1.19 ve basıklık değeri 1.09 olarak bulunmuştur. Aynı ilde yüksek öğretim diploması olmayan 86 katılımcının ise TCA puanları ortalaması 1.83, standart sapması 0.54, puanlara ait dağılımın çarpıklığı 0.72 ve basıklığı 0.90 olarak hesaplanmıştır.Bağımsız gruplar Welch t-testi sonuçları, Uşak ilinde yaşayan katılımcıların Toplumsal Cinsiyet Algısı puanlarının yüksek öğretim durumuna göre değişeceği tezini desteklememiştir, t(95.89)=1.9, p=0.06. Puanlar arasındaki fark için %95 güven aralığı [-0.082,0.402] olarak hesaplanmıştır.

Eğer normallik varsayımı ciddi ölçüde süpheli ise ve özellikle iki grup farklı şekillerde dağılım gösteriyor ise yüzdeli bootstrap prosedürü (percentile bootstrap) kullanılabilir (Wilcox (2012),page 171).

```
#bootstrap ile %95 güven aralığı (normallik varsayımı yok)
set.seed(04012017)
B=5000       # bootstrap tekrar sayısı
alpha=0.05   # alfa

# grupları tanımla
GroupCollege=na.omit(dataWBT_USAK[dataWBT_USAK$HEF=="college","gen_att"])
GroupNONcollege=na.omit(dataWBT_USAK[dataWBT_USAK$HEF=="non-college","gen_att"])

output=c()
for (i in 1:B){

  x1=mean(sample(GroupCollege,replace=T,size=length(GroupCollege)))
  x2=mean(sample(GroupNONcollege,replace=T,size=length(GroupNONcollege)))
  output[i]=x2-x1
  }
output=sort(output)

## yönsüz
# D yıldız alt
output[as.integer(B*alpha/2)+1]
```

```
## [1] -0.0134

# D yıldız üst
output[B-as.integer(B*alpha/2)]
## [1] 0.39

##Yönlü x2>x1
# D yıldız alt
output[as.integer(B*alpha)+1]
## [1] 0.022

#hatalı yön x2<x1
# D yıldız üst
output[as.integer(B*(1-alpha))]
## [1] 0.358
```

8.1.3.2 Yüzdeli bootstrap yöntemi için raporlama örneği

Uşak ilinde yaşayan katılımcılardan yüksek öğretim diplomasına sahip olan 51 kişi için TCA puanları ortalaması 1.64, standart sapması 0.61, puanlara ait dağılım çarpıklık değeri 1.19 ve basıklık değeri 1.09 olarak bulunmuştur. Aynı ilde yüksek öğretim diploması olmayan 86 katılımcının ise TCA puanları ortalaması 1.83, standart sapması 0.54, puanlara ait dağılımın çarpıklığı 0.72 ve basıklığı 0.90 olarak hesaplanmıştır.Puanlar arasındaki fark için %95 bootstrap güven aralığı [-0.013,0.390] olarak hesaplanmıştır. Güven aralığı 0 değerini içerdiği için puanlar arasındaki farkın istatistiksel olarak anlamlı olduğu savunulamaz.

Yönlü test Alternatif hipotez, yüksek öğretim mezunu olmayan katılımcıların puanlarının yüksek olacağını belirtmiş ise; Puanlar arasındaki fark için %95 bootstrap güven aralığı [0.022,∞] olarak hesaplanmıştır. Eldeki veri yüksek öğretim mezunu olmayanların puanlarının daha yüksek olacığı tezini desteyecek kanıt sunmuştur, $H_0 : \mu_{non-college} = \mu_{college}$ in favor of $H_1 : \mu_{non-college} - \mu_{college} > 0$.

Yönlü test Alternatif hipotez, yüksek öğretim mezunu olmayan katılımcıların puanlarının düşük olacağını belirtmiş ise; Puanlar arasındaki fark için %95 bootstrap güven aralığı [$-\infty$,0.358] olarak hesaplanmıştır. Eldeki veri yüksek öğretim mezunu olmayanların puanlarının daha düşük olacağı tezini desteklememektedir. $H_0 : \mu_{non-college} = \mu_{college}$ ve $H_1 : \mu_{non-college} - \mu_{college} < 0$.

Yönlü ve yönsüz alternatif testler farklı kriterler kullanır. Yönlü testlerde yönün nasıl belirlendiği savunulmalıdır. Kullanılan örnekte yönsüz alternatif kullanıldığında Welch t testi p değeri 0.06 bulunmuştur. Bu sonuç marjinal anlamlılık olarak yorumlanabilir. Ülkemizde TCA puanları ve yüksek öğretim ilişkisi hakkında alanyazın sınırlı olduğu için yönlü bir alternatif savunulması zordur.

8.1.4 Etki büyüklüğü: Bağımsız gruplar t testi için

Bir t testi istatistiği ortalamaların birbirinden farklı olup olmadığı hakkında karar vermeye yardımcı olsa da, bu farklılığın büyüklüğünü yorumlamak için etki büyüklüğü hesaplamaları geliştirilmiştir. Bağımsız gruplar t testi için Cohen etki büyüklüğü , ortalamarın farkını bileşik standart sapmaya (pooled) bölünmesi ile hesaplanır.

$$EB = \frac{t}{\sqrt{\frac{n_1 n_2}{n_1+n_2}}}$$

Cohen (1962) etki büyüklüğü sınıflaması

Etki büyüklüğü	Tanım
.2	Küçük

Etki büyüklüğü	Tanım
.5	Orta
.8	Büyük

```
##  normallik ve eş varyanslılık varsayımı yapıldığında
## (dirençli yöntem benzer sonuç verdiğiiçin varsayımların kabulü makuldür.)
n1=51
n2=86
tval=1.96

EB=tval/sqrt((n1*n2)/(n1+n2))
EB
## [1] 0.346

#veya effsize paketi ile
t.test(gen_att~HEF,data=dataWBT_USAK,var.equal=F,
                                    alternative="two.sided",
                                    conf.level=0.95)
##
##  Welch Two Sample t-test
##
## data:  gen_att by HEF
## t = 2, df = 100, p-value = 0.06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.00848  0.40146
## sample estimates:
## mean in group non-college     mean in group college
##                      1.83                      1.64
library(effsize)
cohen.d(gen_att~HEF,data=dataWBT_USAK, paired=F, conf.level=0.95,noncentral=F)
##
## Cohen's d
##
## d estimate: 0.346 (small)
## 95 percent confidence interval:
##      inf      sup
## -0.00579  0.69815
# noncentral=T argümanını araştırabilirsiniz.
```

effsize paketi (Torchiano, 2016) etki büyüklüğünü 0.35 ve ilgili %95 güven aralığını [-0.008, 0.701] olarak hesaplamıştır.

8.1.5 Extra: Pratikte anlamlılık ve istatistiksel anlamlılık

Bir çalışma sonucunda elde edilen sonuçların istatistiksel olarak anlamlı olmadığı fakat prtaikte anlamlı olduğu öne sürülebilir. Bu sakıncalı bşr durumdur ve sadece küçük örneklemlerde görülür. Küçük bir örneklem ile yapılan çalışmadan sonra pratikte anlamlılıktan bahsetmek tezat oluşturur.

Bir çalısma sonucunda elde edilen sonuçların istatistiksel olarak anlamlı olduğu fakat pratikte anlamlı olmadığı öne sürülebilir. Bu doğru olabilir. 400 kişi ile tamamlanan bir çalışmada istatistiksel anlamlı farklılık .05 etki büyüklüğüne sahip olabilir. Eğer .05 etki büyüklüğü çalışmanın yapıldığı alanda küçük adlediliyorsa istatististiksel anlamlılık pratikte anlamlılığı desteklemez.

8.1.6 Kayıp veriler ile bağımsız gruplar t testi

Eklenecek

8.1.7 Destekleyici grafikler

Eklenecek

8.1.8 İstatisksel güç

İstatistiksel güç konusunun ana hatlarına bölüm 7.3.8'de değinilmiştir.

```
#power.t.test
power.t.test(delta=.35, sd=.6,sig.level=0.05, power=0.95,
             type="two.sample", alternative="two.sided")
##
##      Two-sample t test power calculation
##
##               n = 77.4
##           delta = 0.35
##              sd = 0.6
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Bu örnekte belirlenmiş değerler ortalamalar farkı 0.35, standart sapma 0.6, alfa 0.05, yönsüz test ve hedeflenen güç 0.95 seçildiğinde gereken örneklem sayısı her grup için 78'dir. Diğer bir ifade ile, 156 katılımcı ile belirlenen değerlere (ortalamalar farkı 0.35, standart sapma 0.6, alfa 0.05, yönsüz test) ulaşılması durumunda $H_0 : \mu_1 - \mu_2 = 0$ boş hipotezinin terkedilme olasılığı %95'tir.

8.2 Bağlı gruplar t-testi (Within-subjects t-test)

20 tavşan ile gerçekleştirilen deneyde, yara bandı ve dikiş yöntemlerinin yara kapandıktan 10 gün sonra ölçülen germe mukavemeti değerleri üzerinde etkisi araştırılmıştır.

```
gerMUK=data.frame(tavid=1:20,
                  bant=c(6.59,9.84 ,3.97,5.74,4.47,4.79,6.76,7.61,6.47,5.77,
                         7.36,10.45,4.98,5.85,5.65,5.88,7.77,8.84,7.68,6.89),
                  dikis=c(4.52,5.87,4.60,7.87,3.51,2.77,2.34,5.16,5.77,5.13,
                          5.55,6.99,5.78,7.41,4.51,3.96,3.56,6.22,6.72,5.17))

# Grafik verisi
library(tidyr)
plotdata=gather(gerMUK, metot, mukavemet, bant:dikis, factor_key=TRUE)

require(ggplot2)
ggplot(plotdata, aes(x = mukavemet)) +
  geom_histogram(aes(y = ..density..),col="black",alpha=0.7) +
  geom_density(size=2) +
  theme_bw()+labs(x = "mukavemet")+ facet_wrap(~ metot)+
```

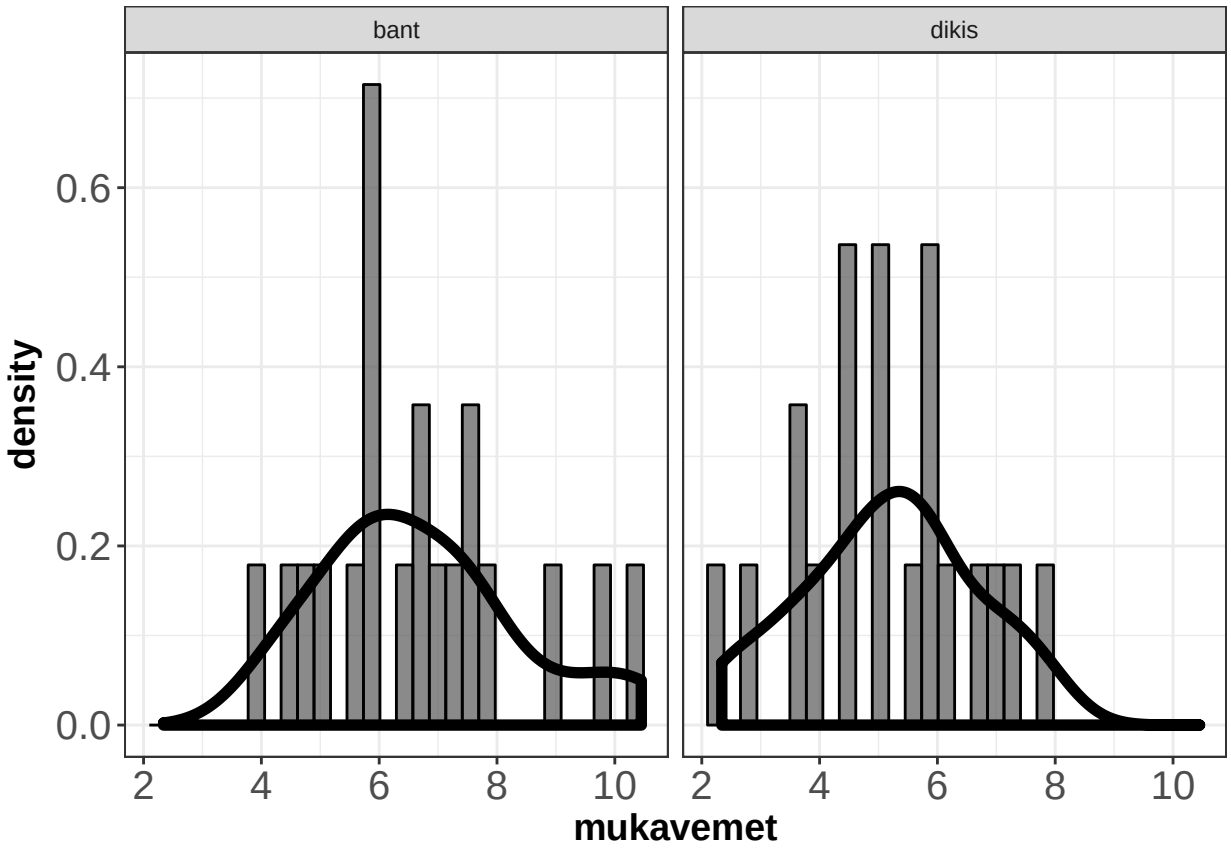


Figure 8.2: Bagli gruplar örnegi

```
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

8.2.1 Bağlı gruplar için t testi

Takip edilen basamaklar;

1. Betimsel istatistikler
2. Test istatistiğinin hesabı

$$t = \frac{\bar{Y_1} - \bar{Y_2}}{\sqrt{\frac{S_1^2 + S_2^2 - 2S_1S_2r_{12}}{n}}}$$

3.Kritik değerin hesabı $\pm t_{\alpha/2,n-1}$,

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

```
library(psych)
descDT=with(gerMUK,describe(cbind(bant,dikis)))
descDT
```

```
##       vars  n mean   sd median trimmed  mad  min   max range  skew
## bant     1 20 6.67 1.71   6.53    6.54 1.45 3.97 10.45  6.48  0.55
## dikis    2 20 5.17 1.49   5.17    5.19 1.30 2.34  7.87  5.53 -0.08
##       kurtosis   se
## bant     -0.45 0.38
## dikis    -0.87 0.33

corDT=with(gerMUK,cor(bant,dikis,use="complete.obs"))
corDT
## [1] 0.354


# tahmin edilen standart hata (tsh)
tsh=sqrt(((1.71^2+1.49^2)-(2*1.71*1.49*corDT))/(20))

# t-istatistik
tstatistic=(6.67-5.17)/tsh

#  alfa=0.05 kritik değer
qt(.975,df=19)
## [1] 2.09
```

Hesaplanan 3.67, kritik değerden ($t_{.975,19} = 2.09$) büyük olduğu için boş hipotez terkedilebilir.

Daha kolay bir R satırı;

```
library(psych)
with(gerMUK, t.test(bant,dikis,paired=T,
                                    alternative="two.sided",
                                    conf.level=0.95))
##
##  Paired t-test
##
## data:  bant and dikis
## t = 4, df = 20, p-value = 0.002
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.643 2.352
## sample estimates:
## mean of the differences
##                     1.5
```

8.2.1.1 Raporlama örneği:

Bağlı gruplar t testi hesaplarına dayanarak, gerilim mukavemetinin bant tedavisi (ort=6.67,SS=1.71,çarpıklık=0.55,basıklık=-0.45) ve dikiş tedavisi (ort=5.17,SS=1.49,çarpıklık=-0.08,basıklık=-0.87) arasında farklılık gösterdiği sonucuna varılmıştır, t(19)=3.67, p=0.002 ,r=0.35. Aritmetik ortalamaların farkı için 95% güven aralığı [0.64,2.35] olarak hesaplanmıştır.

8.2.2 Varsayımlar: bağlı gruplar için t testi

Puanların farkı ($Y_{1i} - Y_{2i}$) normal bir dağılımdan çekilmiş olmalıdır. Puanların farkları her birey için bağımsız olmalıdır. Bağlı gruplar t testi , örneklem küçük değil ise, normallik varsayımı ihlallerine karşı genellikle

dirençlidir.

8.2.3 Dirençli tahminleme yöntemi: bağlı gruplar için t testi

Eğer dağılım normallikten büyük ölçüde ayrılıyor ise, yüzdeli bootstrap prosedürü kullanılabilir (Wilcox (2012),page 201).

```
#bootstrap ile %95 güven aralığı (normallik varsayımı yok)
set.seed(04012017)
B=5000       # bootstrap tekrar sayısı
alpha=0.05   # alfa

gerMUK=data.frame(ratid=1:20,
                  bant=c(6.59,9.84 ,3.97,5.74,4.47,4.79,6.76,7.61,6.47,5.77,
                         7.36,10.45,4.98,5.85,5.65,5.88,7.77,8.84,7.68,6.89),
                  dikis=c(4.52,5.87,4.60,7.87,3.51,2.77,2.34,5.16,5.77,5.13,
                          5.55,6.99,5.78,7.41,4.51,3.96,3.56,6.22,6.72,5.17))

output=c()
for (i in 1:B){
  #satırları örnekle
  bs_rows=sample(gerMUK$ratid,replace=T,size=nrow(gerMUK))
  bs_sample=gerMUK[bs_rows,]
  mean1=mean(bs_sample$bant)
  mean2=mean(bs_sample$dikis)
  output[i]=mean1-mean2
  }
output=sort(output)

## yönsüz
# d yıldız alt
output[as.integer(B*alpha/2)+1]
## [1] 0.686

# d yıldız üst
output[B-as.integer(B*alpha/2)]
## [1] 2.24

##Yönlü x2>x1
# d yıldız alt
output[as.integer(B*alpha)+1]
## [1] 0.837

#yanlış yön x2<x1
# d yıldız üst
output[as.integer(B*(1-alpha))]
## [1] 2.14
```

8.2.3.1 Yönsüz yüzdeli bootstrap için örnek rapor:

Yaraların tedavisinden 10 gün sonra gerilim mukavemeti ölçümleri yapılmıştır. Yara bandı ile (ort=6.67,SS=1.71,çarpıklık=0.55,basıklık=-0.45) dikiş tedavisi (ort=5.17,SS=1.49,çarpıklık=-0.08,basıklık=-0.87) ölçümleri arasındaki fark için 5000 tekrarlı bootstrap prosedürü %95 güven aralığı [0.667,2.2555]

olarak hesaplanmıştır. Yeni geliştirilen yara bandı sonrası gerilim mukavemetinin daha yüksek olduğu ve bu farklılığın istatistiksel olarak anlamlı olduğu sonucuna varılmışştır.

8.2.4 Etki büyüklüğü: bağlı gruplar t testi

En basit etki büyüklüğü hesaplama yöntemlerinden biri; [2]

$$ES = \frac{t}{\sqrt{n}}$$

```
## dirençli prosedürler farklı sonuç vermediği için
## normallik ve varyans eşitliği varsayımları yapılmıştır.
n=20
tval=3.6678

EB=tval/sqrt(n)
EB
## [1] 0.82

library(effsize)
cohen.d(gerMUK$bant,gerMUK$dikis,
        paired=T, conf.level=0.95,noncentral=F)
##
## Cohen's d
##
## d estimate: 0.82 (large)
## 95 percent confidence interval:
##   inf   sup
## 0.154 1.487
```

effsize paketi (Torchiano, 2016) etki büyüklüğünü 0.820 ve ilgili %95 güven aralığını [0.135, 1.505] olarak hesaplamıştır.

8.2.5 Kayıp veriler ile bağlı gruplar t testi

Eklenecek

8.2.6 Destekleyici grafikler

Eklenecek

8.2.7 İstatistiksel güç: bağlı gruplar t testi

```
#power.t.test
power.t.test(delta=.35, sd=.6,sig.level=0.05, power=0.95,
             type="paired", alternative="two.sided")
##
##      Paired t test power calculation
```

[2] Lakens (2013) Eşitlik 7, fakat bu değer korelasyon 1'e yaklaştıkça sonsuza yaklaşır. Lakens (2013) Eşitlik 10 daha uygun bir tercih olabilir. $\frac{meandifference}{(SD_1+SD_2)/2}$

```
##
##              n = 40.2
##          delta = 0.35
##             sd = 0.6
##      sig.level = 0.05
##          power = 0.95
##    alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Bu örnekte belirlenmiş değerler ortalamalar farkı 0.35, standart sapma 0.6, alfa 0.05, yönsüz test ve hedeflenen güç 0.95 seçildiğinde gereken örneklem sayısı 41'dir. Diğer bir ifade ile, 41 katılımcı ile belirlenen değerlere (ortalamalar farkı 0.35, standart sapma 0.6, alfa 0.05, yönsüz test) ulaşılması durumunda $H_0 : \mu_1 - \mu_2 = 0$ boş hipotezinin terkedilme olasılığı %95'tir.

8.3 Yaygın Tasarılar

Sosyal bilimlerde iki ortalamayı kıyaslamak üzere kurulan tasarılardan yaygın olanları özetlenmiştir.

8.3.1 Grupların ilişkili olduğu durumlar

Ortalamaların hesaplandığı puanların farklı gruplar altında ilişkili olup olmadığı önemlidir.

8.3.1.1 Tekrarlanan ölçümler

Aynı katılımcı içiin aynı değişkenin birden fazla ölçülmesi durumu.

1. **Bireylerin kendi kontrol grubunu oluşturması:**

Semantik hafızanın aktivasyon hızını ölçmek üzere 100 üniversite öğrencisi ile araştırma yapılmıştır. Her öğrenci çiftler halinde verilen kelimeleri okumuştur. Okudukları ilk kelime ya bir silah (örneğin mermi, hançer) ya da silah olmayan bir kelimedir. Okudukları ikinci kelime mutlaka agresif bir kelimedir (örneğin yarala, imha et). Her bir öğrenci 192 çift kelimeyi bilgisayarda görüp sesli olarak okumuştur. Bilgisayar ilk kelimeyi 1.25 saniye ekranda tutmuş, .5 saniye siyah ekran göstermiş ve ikinci kelimeyi ekranda göstermiş, her ikinci kelimeden önce reaksiyon zamanını ölçmüştür. Her bireyin reaksiyon zamanı ortalamaları alınmış ve analiz edilmiştir.

	İlk kelime	
Öğrenci	Silah	Silah Değil
1		
2		
...		
100		

Her öğrencinin kendi okuma hızı olduğundan, Silah ve Silah olmayan kelimeleri okumak için gereken reaksiyon zamanları birbiri ile ilişkilidir.

2. **Boylamsal tasarılar:** 6. sınıf öğrencilerinin matematik başarısı dönem başında ve dönem sonunda ölçülmüştür. Amaç puanların değişip değişmediğini görmektir.

	Zaman	
Öğrenci	Dönem başı	Dönem sonu
1		
2		
...		
48		

Ölçümler aynı öğrenciler ile tekrarlandığı için puanlar ilişkilidir.

8.3.1.2 Blok tasarıları

Katılımcıların blok olarak ikili eşleştirilmesidir. Her çiftin (blok içindeki) benzer davranması beklenir.

1. **Rassal blok tasarı:** Okuma hızını artırmak için bir program geliştirilmiştir, etkililiğini araştırmak üzere, 30 ikinci sınıf öğrencisi bir okuma testini cevaplamış, ve puanlarına göre çiftler oluşturmuştur.

Çift	Okuma testi puan sırası
1	1,2
2	3,4
...	...
15	29,30

Görüldüğü gibi en hızlı okuyan 2 öğrenci ilk çifti, en yavas okuyan iki öğrenci son çifti oluşturmuştur.Çiftlerden her biri rassal olarak yeni program grubuna veya kontrol grubuna atanmıştır. As shown, the students with the two highest scores were in the first pair, the students with the

Yeni program son bulduktan sonra öğrencilerin okuma hızları ölçülmüştür

Çift	Yeni Program	Kontrol
1		
2		
...		
...		
15		

2. **Rassal olmayan blok tasarı:** Şiddete maruz kalmış grup çocuğun, şiddet görmemiş daha kalabalık bir gruptan çocuklar ile genel kaygı düzeyi üzerinden eşleştiğini düşünelim. Sonrasında stres anında gösterdikleri endişe durumlarını ölçelim

Çift	Şiddete maruz kalan	Kontrol
1		
2		
...		
20		

3. **Kalıtsal tasarılar (Familial):** 25 anne ve erişkin kızlarının politik görüşleri alınmıştır.

	Çift	
	Anne	Kız
1		

Çift
2
...
25

4. **Dyadik tasaıları:** Afrikalı-Amerikalı ve Avrupalı-Amerikalı çocuklardan oluşturulan çiftler işbirliği gerektiren küçük oyunlar oyamışlardır. Her bir çocuk eşinin işbirlikçiliğini puanlamıştır.

Etnisite		
Çift	Afrikalı-Amerikalı	Avrupalı-Amerikalı
1		
2		
...		
25		

8.3.2 Bağlı olmayan grup tasarılarına örnekler

1. **Tamamen rassal tasarı**
İki farklı manyetik ağrı kesici makinesinin performansı karşılaştırılacaktır. 50 hasta rassal olarak, 25-25, iki makineye alınmış, tedaviden sonra ağrı düzeylerini rapor etmişlerdir.

Makine	
1	2
.	
.	
.	

2. **Rassal olmayan tasarı:** Sekizinci sınıf öğrencilerinden 50 kız ve 50 erkek seçilmiş 2 basamaklı toplama işlemi yapma hızları ölçülmüştür.

Chapter 9

Varyans Analizi (ANOVA)

Varyans analizi (ANOVA) bölümünde kısaca tanıtılan konular; (a) terminoloji, (b) gruplar-arası ANOVA, ve (b) gruplar-içi ANOVA olarak üç başlıktır. Kitabın bir sonraki versiyonunda karma ANOVA (mixed ANOVA) eklenecektir.

9.1 Terminoloji

Varyans analizi kapsamında kullanılan terimler kısaca açıklanmıştır. Fakat bu bölüm bir önceki bölümde tanıtılan *yaygın tasarılar* (8.3) paragraflarının devamıdır.

Faktör Kategorik bağımsız değişkenler ANOVA çerçevesinde faktör olarak isimlendirilir. Önreğin bir önceki bölümde (8.3.2) rassal blok tasarısı içerisinde tanımlanan *okuma hızını artırmaya yönelik yeni program* ve *kontrol* grubu üyeliğini belirteniki kategorili (iki alt sınıflı) değişken bir faktör oluşturur. Toplumsal Cinsiyet Algısının (TCA) yüksek öğretim durumuna göre (yüksek öğretim mezunları ve yüksek öğretim mezunu olmayanlar)değişip değişmediğini test eden bir modelde TCA bağımlı değişken, yüksek öğretim durumu ise bir faktördür.

Alt sınıf Faktörü oluşturan kategorilere alt sınıf denir. TCA örneğinde faktör yüksek öğretim durumudur. Bu faktöre ait alt sınıflar *yüksek öğretim mezunu* ve *yüksek öğretim mezunu değil* olmak üzere iki tanedir.

Kesişen faktörler (Crossed factors) Bir faktöre ait alt sınıfların diğer bir faktörün bütün alt sınıfları ile kesişmesi durumudur. Tekarlanan ölçümlerin tanıtımı amaçlı bir önceki bölümde verilen örneği düşünelim,

	İlk kelime	
Öğrenci	Silah	Silah Değil
1		
2		
...		
100		

Öğrenciler de bir faktör olarak düşünüldüğünde, her öğrencinin reaksiyon süreleri ilk kelime faktörünün her iki alt sınıfında da ölçülmesi sebebi ile öğrenci ve ilk kelime faktörü kesişmiştir.

Kesişmeyen faktörler (Nesting) Bir faktöre ait bir alt sınıfın, ikinci faktörün sadece bir alt sınıfında gözlemlenmesi durumudur. Tamamen rassal tasarının tanıtımı amaçlı bir önceki bölümde verilen örneği düşünelim,

Makine	
1	2
H_1	H_{n+1}
H_2	H_{n+2}
H_3	H_{n+3}
...	
H_n	H_{2n}

Her bir hasta 1. veya 2. makine ile tedavi edildiğinden, katılımcı faktörü makine faktörünün içinde düşünülür ve faktörler kesişmemiştir.

Bağlı gözlemler faktörü (Within-subjects factor) katılımcı faktörü (öğrenci,çalışan,hasta vb.) ile kesişen faktörlerdir. Tekrarlanan ölçümler örneğinde olduğu gibi her bir katılımcı diğer bir faktörün bütün alt sınıflarında gözlemlenir. Boylamsal çalışmalarda zaman (örneğin öntest, sontest) bağlı gözlemler faktörüne örnektir.

Bağlı bloklar faktörü bloklar ile kesişen faktördür. Rassal blok tasarısımı tanıtım amaçlı bir önceki bölümde verilen örneği düşünelim,

Çift	Yeni Program	Kontrol
1		
2		
...		
...		
15		

Rassal olmayan blok, kalıtsal ve diyadik tasarılar da bağlı bloklar faktörüne örnektir.

Bağlı gözlemler faktörü ve bağlı bloklar faktörü aynı şekilde analiz edildiğinden çoğu zaman aynı isimle kullanılır, bu bölümde de her iki faktör *bağlı gözlemler* olarak kullanılmıştır.

Gözlemler arası faktör veya Bağlı olmayan gözlemler faktörü (Between-subjects factor) katılımcı faktörü ile kesişmeyen faktörlerdir. Toplumsal Cinsiyet Algısının (TCA) yüksek öğretim durumuna göre (yüksek öğretim mezunları ve yüksek öğretim mezunu olmayanlar)değişip değişmediğini test eden bir modelde TCA bağımlı değişken, yüksek öğretim durumu ise gözlemler arası faktördür.

9.2 Bağlı olmayan gözlemler varyans analizi (Between Subjects ANOVA)

Gözlemlerin (katılımcıların) tek bir alt sınıf kombinasyonuna ait olduğu durumlarda yapılan çözümlemelerdir.

9.2.1 Tek-yönlü bağlı olmayan gözlemler varyans analizi

Tek faktörlü bağlı olmayan gözlemler varyans çözümlemesi $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ eşitliğinde yer alan parametrelerin tahmini ile tamamlanabilir. Bu eşitlikte Y_{ij} j alt sınıfında yer alan i katılımcısına ait puanı, μ bütün puanların ortalaması, α_j j alt sınıfının etkisi ve ϵ_{ij} hata terimidir. $\mu_j = \mu + \alpha_j$'dir, μ_j j alt sınıfında yer alan katılımcıların aritmetik ortalamasıdır.

Genellikle ilgi α_j üstünedir, çünkü $\mu_j - \mu$ 'yi temsil eder. Bu ilgi $H_0 : \mu_1 = \mu_2 = \cdots = \mu_j$ boş hipotezinin test edilmesini gerektirir.Alternatif hipotez en az bir alt sınıfa ait ortalamanın farklı olduğunu belirtir. Bu boş hipotez varyansın ayrıştırılması ile test edilebilir. Myers et al. (2013) tarafından verilen notasyon kullanırsak;

VK	sd	KT	KTO	F
Toplam	$N-1$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{..})^2$		
A	$J-1$	$\sum_{j=1}^{J} n_j(\bar{Y}_{.j}-\bar{Y}_{..})^2$	SS_A/df_A	$MS_A/MS_{S/A}$
S/A	$N-J$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2$	$SS_{S/A}/df_{S/A}$	

VK	BKTO
Toplam	
A	$\sigma^2_{S/A}+\frac{1}{J-1}\sum_j n_j(\mu_j-\mu)^2$
S/A	$\sigma^2_{S/A}$

VK = vayans kaynağı, sd = serbestlik derecesi, KT = kareler toplamı, KTO = kareler toplamı ortalaması, BKTO = beklenilen kareler toplamı ortalaması.

A , J adet alt sınıfa sahip gruplar arası faktörü, S/A A fatörü içerisindeki katılımcıları, N toplam örneklem sayısını, j=1,...,J faktör alt sınıflarını, i=1,...,n_j katılımcıları, Y_{ij} katılımcı puanlarını, $\bar{Y}_{..}$ genel ortalamayı, $\bar{Y}_{.j}$ j alt sınıfı katılımcı ortalamasını temsil eder.

$MS_A/MS_{S/A}$ oranı, boş hipotez doğru olduğunda ve varsayımlar ihlal edilmediğinde, J-1 ve N-J serbestlik derecesine sahip bir F dağılımını takip eder. Dolayısı ile $MS_A/MS_{S/A}$, $F_{\alpha,J-1,N-J}$ kritik değerinden büyük ise boş hipotez terkedilir.

9.2.1.1 Etki büyüklüğü tek-yönlü bağlı olmayan gözlemler varyans analizi

Tanıtımı kolaylaştırmak amaçlı her alt sınıfın eşit sayıda gözlem içerdiğini düşünelim, $n_1 = n_2 = \cdots = n_j = n$. Bu durumda A faktörü BKTO $\sigma^2 + n\theta^2_A$ 'dır ve ;

$$\theta^2_A = \sum_{j=1}^{J} \frac{(\mu - \mu_j)^2}{J-1}$$

.

θ^2_A parametresinin tahmini olan $\hat{\theta}^2_A$, $\frac{MS_A - MS_{S/A}}{n}$ ile ve $\sigma^2_{S/A}$ parametresinin tahmini olan $\hat{\sigma}_{S/A}$, $MS_{S/A}$ ile hesaplanır.

8.1.4 bölümünde değinildiği gibi, aritmetik ortalamalar arasındaki farkın büyüklüğünü yorumlamak adına etki büyüklüğü hesaplaması yapılabilir. Günümüzde varyans çözümlemesi kullanmış çoğu bilimsel makalede etki büyüklüğü de raporlanmıştır. Bu etki büyüklükleri arasında omega-kare ($\hat{\omega}^2$), eta-kare ($\hat{\eta}^2$) ve f en çok raporlananlar arasındadır.

9.2.1.1.1 Omega-kare

Omega-kare faktör tarafından açıklanan varyansın toplam varyansa oranını tahmin etmek için türetilmiştir. $\hat{\omega}^2 = \frac{(J-1)\hat{\theta}^2/J}{((J-1)\hat{\theta}^2/J)+\hat{\sigma}^2_{S/A})}$

Omega-kare 0.01 ise küçük, 0.06 ise orta ve 0.14 ise büyük etki olarak yorumlanır Myers et al. (2013).

9.2.1.1.2 Eta-kare

Eta-kare de , $\hat{\eta}^2 = \frac{SS_A}{SS_{Total}}$, faktör tarafından açıklanan varyansın toplam varyansa oranını tahmin etmek için türetilmiştir.

Aynı çözümleme için, $\hat{\eta}^2$, $\hat{\omega}^2$ 'den büyüktür , çünkü $\hat{\eta}^2$, özellikle n küçük ise, pozitif yönde yanlı bir tahmindir. Buna rağmen $\hat{\eta}^2$ bildiğimiz kadarı ile en çok raporlanan etki büyüklüğüdür. $\hat{\eta}^2$, regresyon çözümlemesi çerçevesinde R^2 olarak da raporlanır.

9.2.1.1.3 Etki büyüklüğü f

Cohen'in f katsayısı, $f = \frac{\hat{\theta}_A}{\hat{\sigma}_{S/A}}$ ile hesaplanır. Bir f değeri 0.10 ise küçük, 0.25 ise orta, 0.40 ise büyük etki olarak yorumlanır.

9.2.1.1.4 Etki büyüklüğü hesaplamaları üzerine

Yukarıda verilen örneklerde tanıtımı kolaylaştırmak adına her bir alt sınıfta yer alan gözlem sayısı eşit kabul edilmiştir. Fakat pratikte alt sınıfların katılımcı sayısı genellikle eşit değildir. Aynı zamanda genellikle faktör sayısı birden fazladır. Bunlara ilave olarak, tasarı içerisinde faktörlerin manipüle edilmiş olması veya ölçülmüş (measured) olması etki büyüklüğü hesaplarını etkiler. Manipule edilen faktörler için rassal (random) ölçülen faktörler için sabit (fixed) faktör isimlendirilmesi de yaygındır. Örneğin TCA puanları rasgele seçilen 10 ilde yaşayan erkek ve kadınlar için hesaplansın. İk faktörlü bu tasarıda iller manipüle edlien faktör, cinsiyet ise ölçülen faktördür.

ezANOVA fonksiyonu (Lawrence (2016)) Bakeman (2005) tarafından bir araya getirilen genelleştirilmiş eta-kare (generalized eta-squared) formüllerini kullanarak etki büyüklüğü hesaplar. Bakeman (2005) çalışmasında Olejnik and Algina (2003) tarafından tanımlanan genelleştirilmiş eta kareyi kullanır. Etki büyüklüğünü _R paketi *ezANOVA* fonksiyonu ile hesaplamak isteyen kullanıcılar *observed* argümanını incelemeli ve ölçülmüş faktörleri bu argüman ile belirtmelidir. Etki büyüklüğünü R paketleri yerine kod yazarak hesaplamak isteyen araştırmacılar Olejnik and Algina (2003) tarafından verilen formülleri kullanabilir.

9.2.1.2 Alt sınıf ortalamalarının kıyaslanması (Testing specific contrasts)

Varyans analizi sonrasında veya varyans analizi yerine, ortalamalar ile oluşturulmuş farklı kıyaslamalar test edilebilir. Bu çerçevede kıyas, ağırlıklandırılmış aritmetik ortalamaların toplamdır ve ağırlıkların toplamı sıfırdır. İki sınıf kıyas vardır, ikili kıyaslar var karmaşık kıyaslar. Konuyu tanıtım amaşlı, tek faktörün olduğu bir tasarı düşünelim. Bu tasarıda faktöre ait 3 alt sınıf olsun, bir kontrol grubu ve iki farklı müdahele grubu, μ_1, μ_2, ve μ_3. İkili kıyaslama da bir alt grubun ağırlığı -1, diğer bir alt sınıfın ağırlığı -1 ve üçüncü alt sınıfın 0 olur. Örneğin müdahele gruplarının bir biri ile kıyaslanması için $(0)\mu_1+(1)\mu_2+(-1)\mu_3$ kullanılabilir. Kontrol grubunun diğer iki müdahele grubunun ortalaması ile kıyaslanması karmaşık kıyaslamaya örnektir ve $(-1)\mu_1 + (.5)\mu_2 + (.5)\mu_3$ kullanılabilir. Ortalamar arasında fark yoktur boş hipotezini test etmek için;

$$t = \frac{\sum_{j=1}^{J}(w_j\bar{Y})}{\sqrt{MS_{S/A}\sum_{j=1}^{J}(\frac{w_j^2}{n_j})}}$$

9.2.1.3 Bütün ikili kıyaslamaların test edilmesi

Bütün ikili kıyaslamaların yapılabileceği bir kaç farklı prosedür vardır. Birden çok kıyaslamanın yapılacağı durumlarda birinci tip hata oranı kontrol edilmelidir. Bu kontrol birinci tip hata oranını belirlenen bir değerde (örneğin .05) veya daha altında tutmak demektir. En çok kullanılan kontrol yöntemlerinden ikisi (a) kıyaslama bazında (per comparison) hata oranı ve (b) ortak hata oranıdır (familywise). Kıyaslama bazında kritik değer olarak $\pm t_{(1-\alpha 2),N-J}$ kullanıldığında birinci tip hata oranı α'dır. Ortak hata oranı en az bir

kıyaslama için birinci tip hata yapılma oranıdır. Eğer bütün ikili kıyaslamalar sıfıra eşitse, ortak hata oranı α ve $[J(J-1)2]\alpha$ arasındadır. Örneğin 3 alt sınıfı olan bir faktör için yapılacak tüm ikili karşılaştırmalarda birinci tip hata üst limiti 3α 'dır. Ortak hata oranını kontrol etmek için birden fazla prosedür vardır. Bu prosdürlerin R ile tamamlanması oldukça kolaydır ve ilerleyen bölümlerde gösterimi yapılmıştır.

9.2.1.3.1 Trend analizleri

Eklenecek

9.2.1.4 Varsayımlar: tek-yönlü bağlı olmayan gözlemler varyans analizi

Tek-yönlü bağlı olmayan gözlemler varyans analizi varsayımları bağlı olmayan gruplar t testi varsayımları ile benzerdir.

1. Yanıtların bağımsızlığı (independence) her alt sınıfta yer alan puanlar birbirinden bağımsız olmalıdır. Yanıtların bağımsızlığını tehdit eden durumlardan biri aynı grup içeresinde yer alan bireylerin birbirlerinin yanıtlarını etkilemesidir. Eğer bir alt sınıfın içerisinde yanıtların bağımsılığını etkileyen gruplaşmalar varsa çözümleme yöntemi değiştirilmelidir. Örneğin çalışma kapsamında ulaşılan katılımcılar sınıf okul şirket gibi kümelerin içerisinde yer alıyor ve bu kümelere müdahil olmak katılımcıların yanıtlarını etkiliyor ise çok düzeyli modeller kullanılabilir. Eğer katılımcılar bir kümeden etkilenmiyorsa fakat faktöre ait alt sınıflar eşleme (matching) yöntemi ile oluşturulmuşsa oluşan bu bağlılığı göz önünde bulundurmak için rassal blok varyans analizi kullanılabilir.
2. Normallik. Her alt sınıfa ait puanların normal bir dağılımdan geldiği varsayılır. Eğer dağılımların uzun kuyrukları var ise muhtemelen istatistiksel güç azalacaktır. Eğer alt sınıflara ait gözlem sayısı eşit ise normal dağılımdan kopmalar birinci tip hata oranını büyük ölçüde değiştirmez. Bu durum normallikten büyük çapta kopmalar ve küçük örneklem sayıları için geçerli değildir.
3. Eş varyanslılık: Varyans homojenliği olarak da bilinir. J farklı alt sınıfa ait puanların J farklı evrenden geldiğini fakat J farklı evrenin eşit varyansa sahip olduğu varsayımıdır. Alt sınıflara ait gözlem sayısı eşit ve yeterince büyük değil ise, bu varsayımın ihlali birinci tip hata oranını etkiler. Bu varsayımlar sadece özet olarak değinilmiştir. Eğer yanıtların bağımsızlığı zedelenmedi ise, her alt sınıfta eşit sayıda ve en az 20 gözlem var ise ve puanların daağılımı yaklaşık olarak normal ise varyans analizi sonuçları geçerlidir. Diğer durumlarda alternatif analizler kullanılmalıdır. Eğer dirençli analizler ve geleneksel varyans analizi bütün testler için aynı sonuçları veriyorsa geleneksel analizlerin rapor edilmesi araştırmacılar arası iletişimi kolaylaştırmak adına tercih edilebilir.Varyans analizi çerçevesinde dirençli analiz yöntemleri Wilcox (2012) tarafından detaylı olarak ele alınmıştır.

9.2.1.5 R betiği: tek-yönlü bağlı olmayan gözlemler varyans analizi

Gösterim amaçlı, dataWBT'den (2.3) Kocaeli ilinde yaşayan katılımcılar seçilmiştir. TCA puanları bağımsız değişkeni, eğitim durumu 7 alt sınıflı faktörü oluşturur. Bu alt sınıflar diplomasız, ilkokul, ortaokul, lise, meslek lisesi, önlisans ve lisanstır. Fakat diplomasız katılımcı sadece bir kişi olduğu için bu katılımcı ilkokul alt sınıfına aktarılmıştır. [1]

Basamak 1: Veri setini hazırla ve betimsel istatistikleri rapor et

```
# csv yükle
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

# URL sil
rm(urlfile)
```

[1] Bu katılımcı analizlerden çıkarılsa da sonuçlar değişmiyor

```
#Kocaeli'yi seç
# sıralı silme uygula (listwise deletion)
dataWBT_KOCAELI=na.omit(dataWBT[dataWBT$city=="KOCAELI",
                                c("id","gen_att","education")])


#diplomasız katılımcıyı ilkokul alt sınıfına al
library(car)
dataWBT_KOCAELI$eduNEW <- recode(dataWBT_KOCAELI$education,
                                 "'None'='Primary School (5 years)'")

#kozmetik, faktör etiketini kısalt
dataWBT_KOCAELI$eduNEW <- recode(dataWBT_KOCAELI$eduNEW,
                                 "'High School (Lycee)'=
                                 'High School (Lycee) (4 years)'")

dataWBT_KOCAELI$eduNEW <- recode(dataWBT_KOCAELI$eduNEW,
                                 "'Vocational School'=
                                 'Vocational High School (4 years)'")
# faktör alt sınıflarını görmek için
#table(dataWBT_KOCAELI$eduNEW)

##kozmetik, alt sınıfları sırala
#levels(dataWBT_KOCAELI$eduNEW)
dataWBT_KOCAELI$eduNEW = factor(dataWBT_KOCAELI$eduNEW,
                         levels(dataWBT_KOCAELI$eduNEW)[c(4,3,1,6,2,5)])

# hangi katılımcı diplomasız
#which(dataWBT_KOCAELI$education=="None")

#boş alt sınıfları kaldır
dataWBT_KOCAELI$eduNEW=droplevels(dataWBT_KOCAELI$eduNEW)


#betimsel
library(psych)
desc1BW=data.frame(with(dataWBT_KOCAELI,
                        describeBy(gen_att, eduNEW,mat=T,digits = 2)),
                        row.names=NULL)


#istenilenleri seç
# Table 1
desc1BW[,c(2,4,5,6,7,13,14)]
##                                 group1   n mean   sd median  skew kurtosis
## 1             Primary School (5 years)  70 2.11 0.41    2.2 -0.19     0.81
## 2 Junior High/ Middle School (8 years)  94 2.08 0.52    2.1 -0.35    -0.37
## 3        High School (Lycee) (4 years) 158 1.84 0.58    2.0  0.29     0.64
## 4     Vocational High School (4 years)  74 2.04 0.50    2.0 -0.14     0.41
## 5          Higher education of 2 years 112 1.80 0.53    1.8  0.28    -0.36
## 6    University - Undergraduate degree  62 1.78 0.53    1.8  0.06    -0.63

# kaydet
```

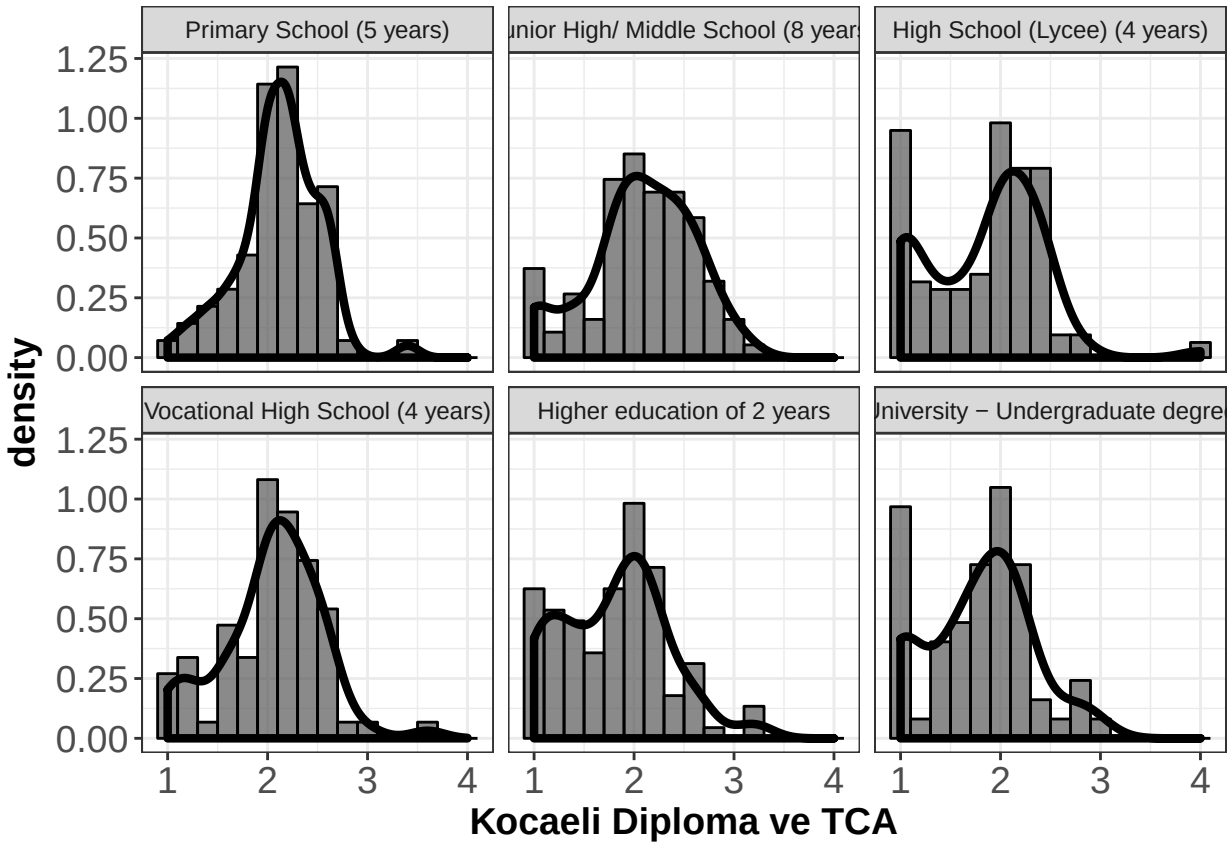


Figure 9.1: Diploma ve TCA

```
#write.csv(desc1BW,file="onewayB_ANOVA_desc.csv")
#write.csv2(desc1BW,file="onewayB_ANOVA_desc.csv")
```

Basamak 2 : Varsayım kontrolü

```
require(ggplot2)
ggplot(dataWBT_KOCAELI, aes(x = gen_att)) +
  geom_histogram(aes(y = ..density..),col="black",binwidth = 0.2,alpha=0.7) +
  geom_density(size=1.5) +
  theme_bw()+labs(x = "Kocaeli Diploma ve TCA")+ facet_wrap(~ eduNEW)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="bold"))
```

Normallikten kopmalar büyük ölçüde değil.

```
require(ggplot2)
ggplot(dataWBT_KOCAELI, aes(eduNEW,gen_att)) +
  geom_boxplot() +
  labs(x = "Education",y="Kocaeli Diploma ve TCA")+coord_flip()
```

Varyans homojenliği sorgulanabilir fakat büyük çaplı bir farklılık yok.

Basamak 3 : varyans analizi

Gösterimin kolaylığı açısından varsayımların ihlal edilmediğini düşünelim. *ezANOVA* fonksiyonu (Lawrence

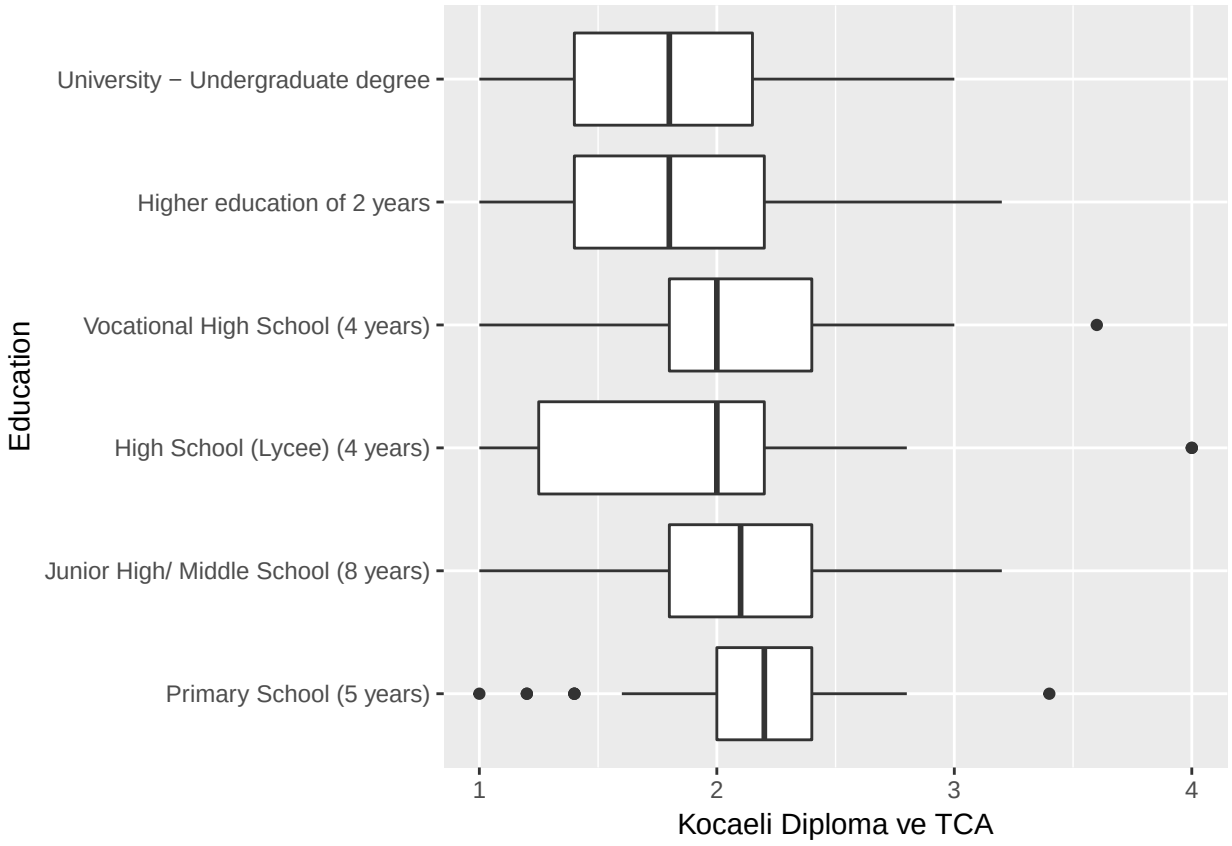


Figure 9.2: Diploma ve TCA

(2016)) F testini, levene testini ve etki büyüklüğünü rapor eder. Etki büyüklüğü hesabı modele göre değişir (Tablo 1 Bakeman (2005) veya Olejnik and Algina (2003)). Bu örnekte, Levene testi *alt gruplar için varyanslar eşittir* boş hipotezinin terkedilmesini destekliyor.

```
#ez kütüphanesini aktif hale getir
library(ez)

#katılımcı kimliğini belirten id değişkeni faktör olmazsa uyarı verir

dataWBT_KOCAELI$id=as.factor(dataWBT_KOCAELI$id)

# kozmetik, virgülden sonra kaç rakam gösterilsin?
options(digits = 3)


#birinci yol, ezANOVA fonksiyonu

alternative1 = ezANOVA(
    data = dataWBT_KOCAELI,
    wid=id, dv = gen_att, between = eduNEW,observed=eduNEW)
## Warning: Data is unbalanced (unequal N per group). Make sure you specified
## a well-considered value for the type argument to ezANOVA().

alternative1
## $ANOVA
##   Effect DFn DFd    F        p p<.05    ges
## 1 eduNEW   5 564 7.27 1.31e-06     * 0.0605
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd  SSn  SSd   F      p p<.05
## 1   5 564 1.35 63.5 2.4 0.0361     *



# kritik F değeri
qf(.95,5,564)
## [1] 2.23
```

```
ez fonksiyonu uyarısı hakkında;
#Warning: Data is unbalanced (unequal N per group). Make sure you specified
#a well-considered value for the type argument to ezANOVA().

bu fonksiyon toplam kareleri 3 farklı şekilde hesaplayabilir.
Tek yönlü varyans analizinde her 3 yöntem de aynı sonucu verir.
Dolayısıyla bu uyarı göz ardı edilebilir.
```

R Core Team (2016b) paketinde yer alan *lm* fonksiyonu ile aynı sonuçlar elde edilebilir.

```
# ikinci yol, lm fonksiyonu
alternative2=lm(gen_att~eduNEW,data=dataWBT_KOCAELI)

#Tablo 2
anova(alternative2)
## Analysis of Variance Table
```

```
##
## Response: gen_att
##            Df Sum Sq Mean Sq F value  Pr(>F)
## eduNEW      5   10.1   2.026    7.27 1.3e-06 ***
## Residuals 564  157.2   0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Core Team (2016b) paketinde yer alan *aov* fonksiyonu da kullanılabilir.

```
#üçüncü yol, aov fonksiyonu
alternative3=aov(gen_att~eduNEW,data=dataWBT_KOCAELI)
summary(alternative3)
##              Df Sum Sq Mean Sq F value  Pr(>F)
## eduNEW        5   10.1   2.026    7.27 1.3e-06 ***
## Residuals   564  157.2   0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

pairwise.t.test fonksiyonu ikili kıyaslama için oldukça kullanışlıdır. Hangi ortak hata kontrol prosedürünü kullanacağınızı belirledikten sonra *p.adjust.method* argümanını kullanabilirsiniz. Örneğin *p.adjust.method* =*"Holm"* ile Holm (1979) tarafından verilen prosedür uygulanabilir. Toplamda 6 farklı prosedür seçilebilir, detaylar için inceleyiniz; *?p.adjust*

```
# ikili kıyaslamalar
# Tablo 3
with(dataWBT_KOCAELI, pairwise.t.test(gen_att,eduNEW,p.adjust.method ="holm"))
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  gen_att and eduNEW
##
##                                     Primary School (5 years)
## Junior High/ Middle School (8 years) 1.000
## High School (Lycee) (4 years)        0.004
## Vocational High School (4 years)     1.000
## Higher education of 2 years          0.001
## University - Undergraduate degree    0.004
##                                     Junior High/ Middle School (8 years)
## Junior High/ Middle School (8 years) -
## High School (Lycee) (4 years)        0.005
## Vocational High School (4 years)     1.000
## Higher education of 2 years          0.002
## University - Undergraduate degree    0.006
##                                     High School (Lycee) (4 years)
## Junior High/ Middle School (8 years) -
## High School (Lycee) (4 years)        -
## Vocational High School (4 years)     0.044
## Higher education of 2 years          1.000
## University - Undergraduate degree    1.000
##                                     Vocational High School (4 years)
## Junior High/ Middle School (8 years) -
## High School (Lycee) (4 years)        -
## Vocational High School (4 years)     -
```

```
## Higher education of 2 years          0.018
## University - Undergraduate degree    0.036
##                                      Higher education of 2 years
## Junior High/ Middle School (8 years) -
## High School (Lycee) (4 years)        -
## Vocational High School (4 years)     -
## Higher education of 2 years          -
## University - Undergraduate degree    1.000
##
## P value adjustment method: holm
```

9.2.1.6 Dirençli tahminleme yöntemi: tek-yönlü bağlı olmayan gözlemler varyans analizi

Wilcox (2012) tarafından bir araya toplanan dirençli prosedürlerden bir tanesi Mair and Wilcox (2016) paketi ile kullanılabilecek *t1way* fonksiyonu ile tamamlanabilir. Kırpılmış ortalamalar için farklı-varyanslı(heteroscedastic) ve tek yönlü ANOVA yöntemini kullanan bu fonksiyonun detayları için inceleyiniz ;*?t1way*

```
library(WRS2)

#t1way
# 20% kırpılmış
t1way(gen_att~eduNEW,data=dataWBT_KOCAELI,tr=.2,nboot=5000)
## Call:
## t1way(formula = gen_att ~ eduNEW, data = dataWBT_KOCAELI, tr = 0.2,
##     nboot = 5000)
##
## Test statistic: 7.57
## Degrees of Freedom 1: 5
## Degrees of Freedom 2: 144
## p-value: 0
##
## Explanatory measure of effect size: 0.29

# 10% kırpılmış
t1way(gen_att~eduNEW,data=dataWBT_KOCAELI,tr=.1,nboot=5000)
## Call:
## t1way(formula = gen_att ~ eduNEW, data = dataWBT_KOCAELI, tr = 0.1,
##     nboot = 5000)
##
## Test statistic: 9.54
## Degrees of Freedom 1: 5
## Degrees of Freedom 2: 188
## p-value: 0
##
## Explanatory measure of effect size: 0.3

# 5% kırpılmış
t1way(gen_att~eduNEW,data=dataWBT_KOCAELI,tr=.05,nboot=5000)
## Call:
## t1way(formula = gen_att ~ eduNEW, data = dataWBT_KOCAELI, tr = 0.05,
##     nboot = 5000)
##
```

```
## Test statistic: 9.41
## Degrees of Freedom 1: 5
## Degrees of Freedom 2: 212
## p-value: 0
##
## Explanatory measure of effect size: 0.31

## heteroscedastic ikili kıyaslamalar

#alt sınıfların sıralanışı
lincon(gen_att~eduNEW,data=dataWBT_KOCAELI,tr=.1)[[2]]
## [1] "Higher education of 2 years"
## [2] "Junior High/ Middle School (8 years)"
## [3] "University - Undergraduate degree"
## [4] "Vocational High School (4 years)"
## [5] "High School (Lycee) (4 years)"
## [6] "Primary School (5 years)"
#ikili kıyaslamalar
round(lincon(gen_att~eduNEW,data=dataWBT_KOCAELI,tr=.1)[[1]][,c(1,2,6)],3)
##       Group Group p.value
##  [1,]     1     2   0.701
##  [2,]     1     3   0.000
##  [3,]     1     4   0.360
##  [4,]     1     5   0.000
##  [5,]     1     6   0.000
##  [6,]     2     3   0.000
##  [7,]     2     4   0.597
##  [8,]     2     5   0.000
##  [9,]     2     6   0.000
## [10,]     3     4   0.004
## [11,]     3     5   0.460
## [12,]     3     6   0.467
## [13,]     4     5   0.001
## [14,]     4     6   0.003
## [15,]     5     6   0.911
```

9.2.1.7 Örnek rapor: tek-yönlü bağlı olmayan gözlemler varyans analizi

Gösterim amaçlı seçtiğimiz dataWBT alt kümesi ile (Kocaeli şehri) tamamlanan geleneksel ANOVA ve dirençli ANOVA , aynı zamanda ikili karşılaştırma testleri aynı sonuçları vermiştir. Varsayımların ihlalleri büyük çapta olmadığı için bu sonuçlar şaşırtıcı değildir. Bu gibi geleneksel ve dirençli yöntemlerin bütün hipotez testleri için aynı karara götürdüğü durumlarda, geleneksel yöntemlerin raporlanması tercih edilebilir.

TCA puanlarının eğitim durumuna göre değişip değişmediğini test etmek amaçlı varyans çözümlemesi yapılmıştır. Tablo 1 ile bütün eğitim durumları için aritmetik ortalama, standart sapma, örenklem çarpıklığı ve örneklem basıklığı değerleri verilmiştir. Varyans analizi eğitim durumunun TCA puanları üzerinde etkisi olduğu hipotezini desteklemiştir, $F(5,564) = 7.27$, $p < .001$, $\hat{\eta}^2_G = .06$. Bu analizler için ANOVA tablosu Tablo 2 ile verilmiştir. İkili kıyaslamalar ortak hata oranını Holm (1979) tarafından verilen prosedüre uygun olarak tamamlanmış sonuçlar Tablo 3 ile verilmiştir. 15 farklı ikili kıyaslamada, ortalamaların farkı 9 kıyaslama için istatistiksel olarak anlamlı bulunmuştur (tespit edilen farklılıklar detaylı olarak açıklanabilir.) Model varsayımları kontrol edilmiş, büyük çaplı bir ihlal tespit edilmemiş olmasına rağmen dirençli yöntemlerden kırpılmış ortalamalar için farklı-varyanslı ANOVA (Mair and Wilcox (2016)) prosedürü ile sonuçlar karşılaştırılmış ve bir farklılık olmadığı görülmüştür.

9.2.1.8 Kayıp data teknikleri: tek-yönlü bağlı olmayan gözlemler varyans analizi

To be added

9.2.1.9 İstatistiksel güç hesapları: tek-yönlü bağlı olmayan gözlemler varyans analizi

To be added

9.2.2 İki faktörlü bağlı olmayan gözlemler varyans analizi

Bu başlık altında iki bağlı olmayan faktörlü varyans analizi ele alınmıştır. Faktörlerden ilki J alt sınıfa sahip A faktörü, ikincisi K alt sınıfa sahip B faktörü olarak düşünülmüştür. Bu durumda JK farklı alt sınıf kombinasyonu oluşur. Her bir gözlemin bir ve yalnız bir alt sınıf kombinasyonunda yer alması ve alt sınıfların bir birine eşlenmesi söz konusu olmadığından, bu tasarı bağlı olmayan gözlemler varyans çözümlemesine uygundur.En bait halinde, her iki faktör sadece 2 alt sınıfa sahiptir. Örneğin TCA puanlarına ait varyans cinsiyet ve yüksek öğretim durumuna göre çözümlenirse aşağıda yer alan tablo oluşur.

	Lise ve altı	Yüksek öğretim	
Kadın	μ_{11}	μ_{12}	$\mu_{1\cdot}$
Erkek	μ_{21}	μ_{22}	$\mu_{2\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	

Bu tasarıda hipotez testlerinde kullanılan artitmetik ortalamar $\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$, satır ortalamaları $\mu_{1\cdot}$, $\mu_{2\cdot}$ ve sütun ortalamaları $\mu_{\cdot 1}$, $\mu_{\cdot 2}$ parametreleri ile gösterilmiştir. Satır veya sütun ortalamalarının genel adı marjinal ortalamadır (kenar veya köşe ortalamarı da denilebilir).

Faktörler arasında etkileşim (interaction) olup olmadığına yönelik kurulacak boş hipotez $H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$. Bu etkileşim aynı zamanda iki basit etkinin de karşılaştırılmasıdır, ($\mu_{21}\mu_{11}$ ve $\mu_{22}\mu_{12}$) ve $H_0 : \mu_{21}\mu_{11} = \mu_{22} - \mu_{12}$ boş hipotezi ile de gösterileblir. Bu boş hipotezlerden biri doğru ise diğeri de doğru, biri yanlış ise diğeri de yanlıştır.

Etkileşim Bu tasarıda ilk test edilen hipotez etkilesşim hipotezidir. Fakat etkileşimi tanımlamadan önce basit etkileri tanımlamak gerekir. Basit etki tek bir satırda veya tek bir sütunda yer alan ortalamarın farkıdır. Kullandığımız örnekte iki çeşit basit etki vardır, cinsiyetin basit etkisi ve yüksek öğretim durmunun basit etkisi. Her bir basit etkinin de iki çeşidi vardır; cinsiyetin yüksek öğretimliler üzerine basit etkisi (μ_{12} ve μ_{22}), cinsiyetin yüksek öğretim mezunu olmayanlar üzerine etkisi (μ_{11} ve μ_{21}). Yüksek öğretimin kadınlar üzerine etkisi (μ_{11} ve μ_{12}) ve yüksek öğretimin erkekler üzerine etkisi (μ_{21} versus μ_{22}).

Asıl etkiler marjinal ortalamar ile tanımlanan etkilerdir. Cinsiyetin asıl etkisi $\mu_{1\cdot} - \mu_{2\cdot}$ şeklinde gösterilir ve $H_0 : \mu_{1\cdot} - \mu_{2\cdot} = 0$ boş hipotezi üzerinden test edilir. Yüksek öğretim etkisi de $H_0 : \mu_{\cdot 1} - \mu_{\cdot 2} = 0$ boş hipotezi üzerinden test edilir.

Etkileşim istatistiksel olarak anlamlı olduğunda:

1. Eğer basit etkilerin yönü aynı değil ise asıl etkiyi yorumlamak yanıltıcı olur.
2. Eğer basit etkilerin yönü aynı ise asıl etkiyi yorumlamanın yanıltıcı olup olmadığına araştırmacı karar verir.

Etkileşim istatistiksel olarak anlamlı ve asıl etkileri yorumlamak yanıltıcı ise araştırmacı hücre bazında aritmetik ortalamaları yorumlamalıdır.

Eşitlik İki bağlı olmayan faktör varyans analizi için çözümleme modeli $Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \epsilon_{ij}$, olarak verilebilir. Y_{ijk} birinci faktörün j alt sınıfı, ikinci faktörün k alt sınıfında yer alan i katılımcısının puanını; μ

genel ortalamayı; α_j ilk faktöre ait j alt sınıfının etkisini; β_k ikinci faktöre ait k alt sınıfının etkisini; $\alpha\beta_{jk}$ etkileşimi ve ϵ_{ij} hata terimini temsil eder.

SV	df	F
A	$J-1$	$\frac{MS_A}{MS_{S/AB}}$
B	$K-1$	$\frac{MS_B}{MS_{S/AB}}$
AB	$(J-1)(K-1)$	$\frac{MS_{AB}}{MS_{S/AB}}$
S/AB	$N-JK$	
Total	$N-1$	

9.2.2.1 R betiği: İki faktörlü bağlı olmayan gözlemler varyans analizi

Gösterim amaçlı dataWBTde yer alan Kayseri ili katılımcıları seçilmiştir. TCA puanlarına ait varyans cinsiyet ve yüksek öğretim durumuna göre ayrıştırılmıştır.

Basamak 1 Veriyi hazırla ve betimsel istatistikleri raporla

```
# CSV yükle
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#URL sil
rm(urlfile)

#Kayseri ilini seç
# sıralı silme uygula
dataWBT_Kayseri=na.omit(dataWBT[dataWBT$city=="KAYSERI",c("id","gen_att","higher_ed","gender")])

# Yüksek öğretim etiketlerini değiştir
dataWBT_Kayseri$HEF=droplevels(factor(dataWBT_Kayseri$higher_ed,
                    levels = c(0,1),
                    labels = c("non-college", "college")))

#table(dataWBT_Kayseri$gender)
#table(dataWBT_Kayseri$HEF)

#boş alt sınıfları düşür
dataWBT_Kayseri$gender=droplevels(dataWBT_Kayseri$gender)

with(dataWBT_Kayseri,
     table(gender,HEF))
##         HEF
## gender   non-college college
##   Female          99      50
##   Male            67      36


#  kozmetik, virgülden sonra kaç rakam gösterilsin?
options(digits = 3)

#betimsel analizler
library(doBy)
```

```
library(moments)
desc2BW=as.matrix(summaryBy(gen_att~HEF+gender, data = dataWBT_Kayseri,
          FUN = function(x) { c(n = sum(!is.na(x)),
                                mean = mean(x,na.rm=T), sdv = sd(x,na.rm=T),
                               skw=moments::skewness(x,na.rm=T),
                               krt=moments::kurtosis(x,na.rm=T)) } ))
# Tablo 4
desc2BW
##   HEF           gender   gen_att.n gen_att.mean gen_att.sdv gen_att.skw
## 1 "non-college" "Female" "99"      "1.93"       "0.424"     "-0.548"
## 2 "non-college" "Male"   "67"      "2.32"       "0.419"     "-0.191"
## 3 "college"     "Female" "50"      "1.80"       "0.346"     " 0.263"
## 4 "college"     "Male"   "36"      "2.13"       "0.543"     " 0.159"
##   gen_att.krt
## 1 "2.51"
## 2 "3.18"
## 3 "1.94"
## 4 "2.25"
#write.csv(desc2BW,file="twowayB_ANOVA_betimsel.csv")
```

Basamak 2: Varsayım kontrolü

```
require(ggplot2)
ggplot(dataWBT_Kayseri, aes(x = gen_att)) +
  geom_histogram(aes(y = ..density..),col="black",binwidth = 0.2,alpha=0.7) +
  geom_density(size=1.5) +
  theme_bw()+labs(x = "Kayseri Diploma ve TCA")+ facet_wrap(~ HEF+gender)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="bold"))
```

Normallikten kopmalar büyük ölçüde değil.

```
require(ggplot2)
ggplot(dataWBT_Kayseri, aes(x=gender, y=gen_att))+
  geom_boxplot()+
  facet_grid(.~HEF)+
  labs(x = "Gender",y="Kayseri Diploma ve TCA")
```

Varyanslar benzer görünüyor.

Basamak 3: Varyans analizi

ezANOVA fonksiyonu (Lawrence (2016)) F testini, Levene testini ve etki büyüklüğünü rapor eder. Etki büyüklüğü hesabı kullanılan modele göre (Bakeman (2005) veya Olejnik and Algina (2003)) ve toplam kareler hesaplama yöntemine göre değişir. *type* argümanı hangi tip toplam kareler hesabı kullanılacağını belirler.

```
library(ez)
#katılımcı kimliğini belirten id değişkeni faktör olmazsa uyarı verir

dataWBT_Kayseri$id=as.factor(dataWBT_Kayseri$id)

#birinci yol ezANOVA
alternative1 = ezANOVA(
    data = dataWBT_Kayseri,
```

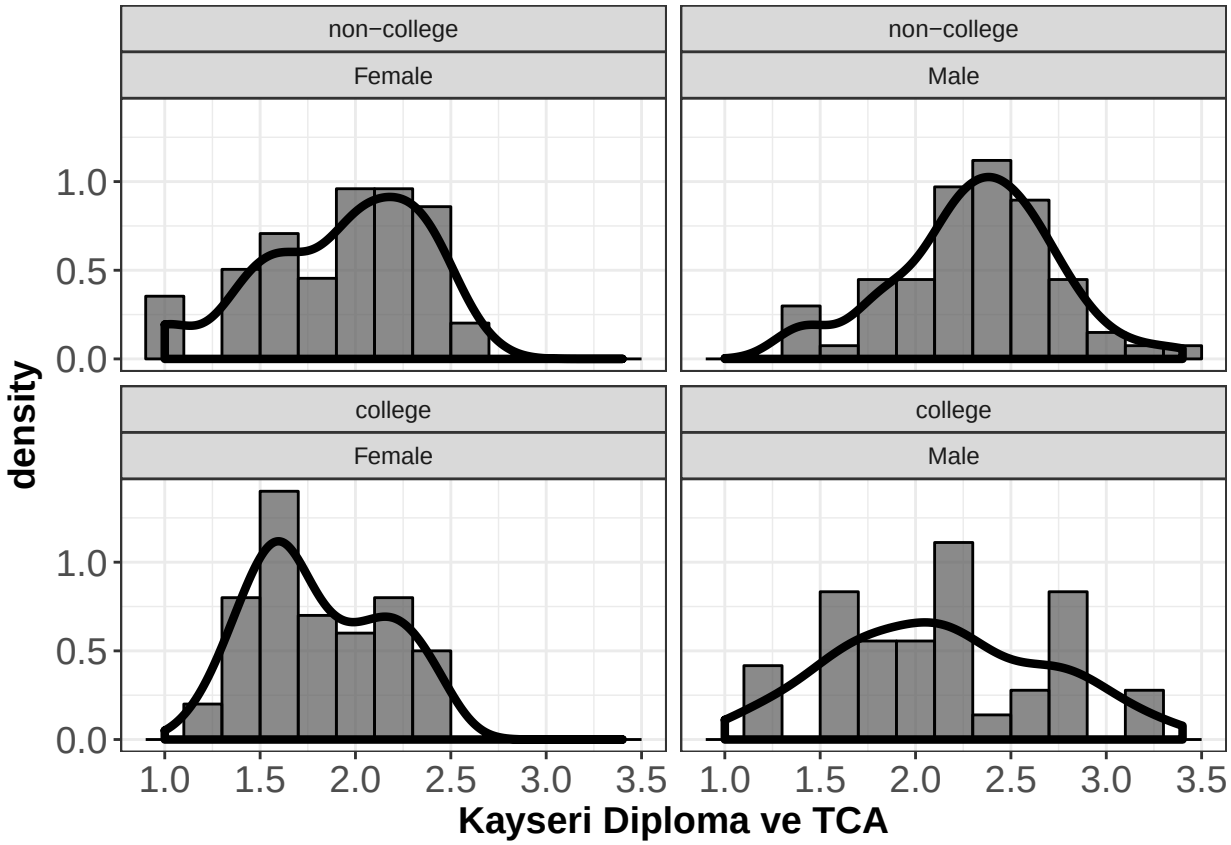


Figure 9.3: Kayseri Diploma ve TCA

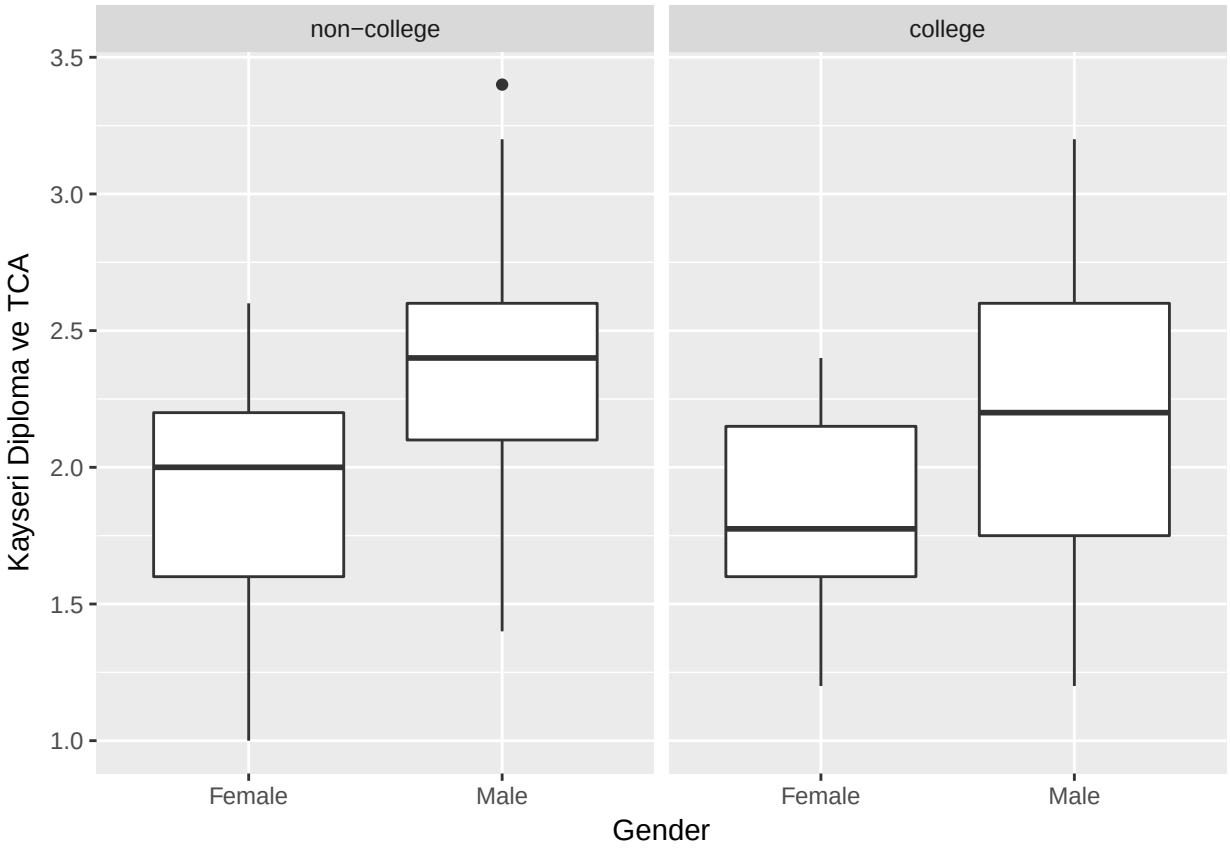


Figure 9.4: Kayseri Diploma ve TCA

```
    wid=id, dv = gen_att, between = .(HEF,gender),observed=.(HEF,gender),type=2)
## Warning: Data is unbalanced (unequal N per group). Make sure you specified
## a well-considered value for the type argument to ezANOVA().

alternative1
## $ANOVA
##       Effect DFn DFd      F        p p<.05      ges
## 1        HEF   1 248  6.739 9.99e-03     * 0.022436
## 2     gender   1 248 45.389 1.12e-10     * 0.151106
## 3 HEF:gender   1 248  0.251 6.17e-01       0.000837
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd   SSn  SSd    F      p p<.05
## 1   3 248 0.469 17.5 2.22 0.0867


# Tip III toplam kareler
# alternative1b = ezANOVA(
#     data = dataWBT_Kayseri,
#     wid=id, dv = gen_att, between = HEF+gender,type=3)
#
# alternative1b

# kritik F değeri
qf(.95,1,248)
## [1] 3.88
```

9.2.2.2 Dirençli tahminleme yöntemi: iki-yönlü bağlı olmayan gözlemler varyans analizi

Wilcox (2012) tarafından bir araya toplanan dirençli prosedürlerden bir tanesi Mair and Wilcox (2016) paketi ile kullanılabilecek *t2way* fonksiyonu ile tamamlanabilir. Kırpılmış ortalamalar için farklı-varyanslı(heteroscedastic) ve iki yönlü ANOVA yöntemini kullanan bu fonksiyonun detayları için inceleyiniz ;*?t2way*

```
library(WRS2)

#t2way
# 20% kırpılmış
t2way(gen_att~HEF*gender,data=dataWBT_Kayseri,tr=.2)
## Call:
## t2way(formula = gen_att ~ HEF * gender, data = dataWBT_Kayseri,
##     tr = 0.2)
##
##              value p.value
## HEF         7.1310   0.011
## gender     20.2039   0.001
## HEF:gender  0.0855   0.772

# 10% kırpılmış
t2way(gen_att~HEF*gender,data=dataWBT_Kayseri,tr=.1)
## Call:
## t2way(formula = gen_att ~ HEF * gender, data = dataWBT_Kayseri,
##     tr = 0.1)
```

```
##
##              value p.value
## HEF         8.4235   0.005
## gender     33.1599   0.001
## HEF:gender  0.0361   0.850

# 5% kırpılmış
t2way(gen_att~HEF*gender,data=dataWBT_Kayseri,tr=.05)
## Call:
## t2way(formula = gen_att ~ HEF * gender, data = dataWBT_Kayseri,
##     tr = 0.05)
##
##              value p.value
## HEF          6.169   0.015
## gender      29.838   0.001
## HEF:gender   0.164   0.687
```

9.2.2.3 Örnek rapor: iki-yönlü bağlı olmayan gözlemler varyans analizi

Gösterim amaçlı seçtiğimiz dataWBT alt kümesi ile (Kayseri şehri) tamamlanan geleneksel ANOVA ve dirençli ANOVA aynı sonuçları vermiştir. Varsayımların ihlalleri büyük çapta olmadığı için bu sonuçlar şaşırtıcı değildir. Bu gibi geleneksel ve dirençli yöntemlerin aynı karara götürdüğü durumlarda, geleneksel yöntemlerin raporlanması tercih edilebilir.

Tablo 4 Kayseri ilinde yaşayan katılımcılara ait TCA puanlarını cisiyet ve yüksek öğretim faktörlerine göre raporlamıştır. 2x2 varyans analizi sonuçları raporlanmıştır. F testleri alfa=0.05 ile tamamlanmıştır. Cinsiyet etkisi istatistiksel olarak anlamlı bulunmuştur $F(1, 248) = 45.39, p < .001$. Yüksek öğretim etkisi anlamlı bulunmuştur, $F(1, 248) = 6.24, p = .013$. İki faktör arasında etkileşimin mevcut olduğuna dair kanıt bulunamamıştır,$F(1, 248) = 0.25, p = .617$. *ezANOVA* (Lawrence (2016)) fonksiyonu cinsiyet faktörü için 0.15, yüksek öğretim faktörü için 0.02 genelleştirilmiş eta kare ($\hat{\eta}^2_G$) değeri hesaplamıştır. Tablo 5 ANOVA sonuçlarını bildirir.

9.2.2.4 Takviye çözümlemeler (Follow-ups): iki-yönlü bağlı olmayan gözlemler varyans analizi

To be added.

9.2.2.4.1 İkili kıyaslamalar: iki-yönlü bağlı olmayan gözlemler varyans analizi

To be added.

9.2.2.4.2 Karmaşık kıyaslamalar: iki-yönlü bağlı olmayan gözlemler varyans analizi

To be added.

9.2.2.5 Kayıp veri teknikleri: iki-yönlü bağlı olmayan gözlemler varyans analizi

To be added

9.2.2.6 İstatiksel güç hesapları: iki-yönlü bağlı olmayan gözlemler varyans analizi

To be added

9.3 Bağlı gözlemler varyans analizi

Aynı katılımcıya ait puanlar birden fazla faktör alt sınıfında yer alıyorsa veya aynı katılımcı belirli zaman aralıkları ile tekrar gözlemleniyorsa (repeated measures) bağlı gözlemler varyans çözümlemesi kullanılabilir. Blokların kullanıldığı tasarılarda da kullanılabilir. Bağlı olmayan gözlemler varyans analizi ile karşılaştırıldığında , bu yöntemin varyansın artmasına sebep olabilecek katılımcı farklılıklarını ortadan kaldırabilir. Geride bırakılan bu birey kaynaklı varyans fazlalığı genellikle hatayı azalttığından, istatistiksel gücün artmasını sağlar. Bir diğer ifade ile, örneklem büyüklüğü sabit tutulduğunda, bu yöntem ile farklılıkları tespit edebilme olasılığı daha yüksektir. Bununla beraber bağlı gözlemler her zaman uygun değildir. Örneğin 3 farklı öğretim metodunun karşılaştırılması için aynı birey birden fazla programa müdahil olduğunda, programların etkisi birbirine karışabileceğinden bağlı gözlemler kullanmak uygun değildir.

9.3.1 Tek-yönlü bağlı gözlemler varyans analizi

9.4 Eklemesiz (non-additive) model için eşitlik;

$$Y_{ij} = \mu + \eta_i + \alpha_j + (\eta\alpha)_{ij} + \epsilon_{ij} \tag{9.1}$$

i bireyleri, i=1,...,n; j faktöre ait alt sınıfları, j=1,...,P temsil eder. Y puanları; μ genel ortalamayı; η_i bireye ait ortalamanın genel ortalamadan farkını; α_j j alt sınıfının genel ortalamadan farkını; $(\eta\alpha)_ij$ etkileşimi; ve ϵ_{ij} hata terimini temsil eder. $(\eta\alpha)_ij$ ve ϵ_{ij} aynı alt indise sahip olduğundan etkileri birbirine karışır. Genellikle ilgi α_j üzerinedir ve $H_0 : \mu_1 = \mu_2 = \cdots = \mu_P$ boş hipotezi test edilir. Alternatif hipotez, en az bir aritmetik ortalamanın farklı olduğunu belirtir. Tek-yönlü bağlı gözlemler varyans analizi tablosu;

SV	df	F
Subjects (S)	$n-1$	
Waves (A)	$P-1$	$\frac{MS_A}{MS_{SA}}$
SA	$(n-1)(P-1)$	
Total	$nP-1$	

Not: Eklemesizlik Eşitlik (9.1) içerisinde yer alan $(\eta\alpha)_ij$ parametresinin 0 olmaması durumudur. Bu gerçekçi bir durumdur, çünkü bu parametrenin sıfır olması faktöre ait alt sınıfın bütün bireyleri eşit şekilde etkilemesi demektir. Tekrarlanan ölçümlerde bireylerin zaman içerisinde tamamen aynı puansal değişimi göstermesi anlamına gelir.

Tablo 9.9 bir deneye ait verileri gösterir. Bu deneyde bireylerin tükettiği alkol dozu artırılmış ve reaksiyon zamanları ölçülmüştür.

Alt sınıflara ait ortalama, standart sapma ve bireye ait ortalama aşağıda verilmiştir. Bireye ait ortalama dört alt sınıfın ortalamasıdır.

```
# kozmetik virgülden sonraki basamak sayısı
options(digits = 2)

#bireylerin ortalaması
apply(owadata,1, mean)
## [1] 3.2 4.0 4.4 4.8 5.4 6.0 6.2 7.6

#alt sınıfların ortalaması
apply(owadata[,-1],2, mean)
## Alkolyok    ikioz   dortoz   altioz
##      2.8      3.5      6.2      9.0
```

Table 9.9: Orijinal Alkol Verisi

id	Alkolyok	ikioz	dortoz	altioz
1	1	2	5	7
2	2	3	5	8
3	2	3	6	8
4	2	3	6	9
5	3	4	6	9
6	3	4	7	10
7	3	4	7	10
8	6	5	8	11

Table 9.10: Orijinal Alkol Verisi Korelasyon Katsayilari

	Alkolyok	ikioz	dortoz	altioz
Alkolyok	1.00	0.93	0.88	0.88
ikioz	0.93	1.00	0.89	0.94
dortoz	0.88	0.89	1.00	0.95
altioz	0.88	0.94	0.95	1.00

```
#alt sınıfların standart sapması
apply(owadata[,-1],2,sd)
## Alkolyok    ikioz   dortoz   altioz
##     1.49     0.93     1.04     1.31
```

Her katılımcının reaksiyon zamanları 4 farklı doz sonrasında da ölçüldüğünden, alkol oranı bağlı gözlem faktörüdür. Her doz alt sınıf çifti için korelasyon hesaplanabilir. Tablo 9.2 ile verilen bu korelasyonlar oldukça yüksektir. Her alt sınıf çifti için kovaryans da hesaplanabilir;

$$Cov_{pp'} = S_p S_{p'} r_{pp'}$$

p ve p' alkol faktörüne ait iki farklı alt sınıfı temsil eder. İlk iki alt sınıf için korelasyon $r_{02} = 0.93$ kovaryans ise $Cov_{02} = 1.5 * 0.9 * 0.93 = 1.26$ dir.

P = bağlı-gözlemler faktörü alt sınıf sayısı, örneğimizde P=4 ;

$\bar{C}$ = ortalama kovaryans; örneğimizde $\bar{C} = 1.26$.

F istatistiği

$$F_W = \frac{MS_A}{MS_{SA}} = \frac{MS_A}{MS_{S/A} - \bar{C}}$$

MS_A ve $MS_{S/A}$ bağlı olmayan faktör analizinde olduğu gibi hesaplanır. W harfi F test istatistiğinin bağlı gözlemler (within) için hesaplandığını gösterir. Kritik değer $F_{\alpha,P-1,(P-1)(n-1)}$ ile hesaplanır.

Bağlı olmayan gözlemler için $F_B = MS_A/MS_{S/A}$ iken bağlı gözlemlerde hesaplanan F_W korelasyonları dikkate alır.Korelasyon sıfır değil ise aynı veriye uygulandığında $F_W \geq F_B$ olduğu görülür.

9.4.0.1 Varsayımlar: Tek-yönlü bağlı gözlemler varyans analizi

Küresellik (Sphericity) Kovaryans örüntüsü hakkında bir varsayımdır. Küresellik sağlanırsa her bir tekrarlanan ölçüm çifti farkı için hesaplanan varyans aynıdır.

Örnek kovaryans matrisi;

	Y_1	Y_2	Y_3
Y_1	10	7.5	10
Y_2	7.5	15	12.5
Y_3	10	12.5	20

Küresellik mevcut;

$Y_p - Y_{p'}$	$\sigma_p^2 + \sigma_{p'}^2 - 2\sigma_{pp'}$
$Y_1 - Y_2$	10+15-2(7.5)=10
$Y_1 - Y_3$	10+20-2(10)=10
$Y_2 - Y_3$	15+20-2(12.5)=10

Box epsilon değeri — küreselliğin ne kadar zedelendiğini ölçer

$$\frac{1}{P-1} \leq \epsilon \leq 1$$

ϵ parametresinin tahminlerinden iki tanesi Greenhouse-Geisser ($\hat{\epsilon}$) ve Huynh-Feldtdir ($\tilde{\epsilon}$) . $\tilde{\epsilon}$ 1'den büyük olabilir; bu durumda $\tilde{\epsilon}$ 1'e eşitlenir.

Küresellik varsayımı ile kritik değer $F_{alpha,(P-1),(n-1)(P-1)}$.

Küresellik varsayımı ihlal edildi ise yaklaşık kritik değer $F_{alpha,\epsilon(P-1),\epsilon(n-1)(P-1)}$.

hataların normal dağılımı Eşitlik (9.1) içerisinde ϵ_{ij} ile temsil edilen hataların ortalaması sıfır olan bir normal dağılımdan geldiği varsayılır.

η_i **normal dağılımı** Eşitlik (9.1) içerisinde η_i ile temsil edilen değerlerin ortalaması sıfır olan bir normal dağılımdan geldiği varsayılır.

Listelenen bu varsayımlar, tekrarlanan ölçümlerin çokdeğişkenli normal dağılımdan (multivariate normal distribution) geldiği anlamındadır.

9.4.0.1.1 Eklemesizlik ve Küresellik arasındaki ilişki

Varasyımlar η_i ve ϵ_{ij} üzerinden tanımlanabilse dahi daha basit bir varsayım tanımı *verilerin çokdeğişkenli normal dağılımdan gelmesi ve küreselliği sağlaması* olarak da yapılabilir. Eğer çokdeğişkenli normal dağılım varsayımı gerçekci ise ve varyanslar ve kovaryanslar eşit ise (bileşik simetri, compound symmetry) hesaplanacak F testi geceçerlidir.

Eğer eklemelilik (additivity) ve eş varyanslılık mevcut ise bileşik simetri sağlanmış olur. Fakat bileşik simetri küresellik varsayımına nazaran gerçekleşmesi daha zor bir varsayımdır. Verilerin çokdeğişkenli norma varsayımdan geldiği durumlarda küresellik varsayımının ihlal edilmemesi F testinin geçerli olması için yeterlidir. Dolayısı ile küresellik varsayımını kontrol etmek eklemelilik varsayımını kontrol etmekten daha önemlidir. Bununla birlikte, küresellik varsayımının ihlal edilmesi durumunda kritik değer düzeltme prosedürleri mevcut olduğu için eklemelilik varsayımını kontrol etmek gereksizdir.

9.4.0.2 R betiği: Tek-yönlü bağlı gözlemler varyans analizi

Gösterim amaçlı Daunic et al. (2012) tarafından toplanan verilerden bir alt küme seçilmiştir. Seçilen sınıfta 17 öğrenci mevcuttur. Bağımlı değişken problem çözme bilgisidir. Öğrencilerin 1 sene arayla problem çözme bilgisi ölçülmüştür. Yüksek puanlar bilginin arttığını gösterir.

Basamak 1 Veriyi hazırla

```
# datayı gir
PSdata=data.frame(id=factor(1:17),
                  wave1=c(20,19,13,10,16,12,16,11,11,14,13,17,16,12,12,16,16),
                  wave2=c(28,27,18,17,29,18,26,21,15,26,28,23,29,18,26,21,22),
                  wave3=c(21,24,14,8,23,15,21,15,12,21,23,17,26,18,14,18,19))
```

Betimsel analizleri raporla

```
# kozmetik, basamak sayısını belirle
options(digits = 3)

#veriyi uzun formata çevir
#head(PSdata)
library(tidyr)
data_long = gather(PSdata, wave, PrbSol, wave1:wave3, factor_key=TRUE)

#betimsel analizler
library(doBy)
library(moments)
desc1W=as.matrix(summaryBy(PrbSol~wave, data = data_long,
          FUN = function(x) { c(n = sum(!is.na(x)),
                                mean = mean(x,na.rm=T), sdv = sd(x,na.rm=T),
                                skw=moments::skewness(x,na.rm=T),
                                krt=moments::kurtosis(x,na.rm=T)) } ))
# Tablo 6
desc1W
##   wave    PrbSol.n PrbSol.mean PrbSol.sdv PrbSol.skw PrbSol.krt
## 1 "wave1" "17"     "14.4"      "2.91"     " 0.311"   "2.10"
## 2 "wave2" "17"     "23.1"      "4.67"     "-0.224"   "1.64"
## 3 "wave3" "17"     "18.2"      "4.77"     "-0.315"   "2.45"
#write.csv(desc1W,file="onewayW_ANOVA_desc.csv")
#write.csv2(desc1W,file="onewayW_ANOVA_desc.csv")
```

Kovaryans matrisi

```
#  Tablo 7
cov(PSdata[,-1])
##       wave1 wave2 wave3
## wave1  8.49  8.85  9.87
## wave2  8.85 21.81 18.49
## wave3  9.87 18.49 22.78
```

Basamak 2 Varsayım kontrolü

```
ggplot(data_long, aes(x=wave, y=PrbSol))+
  geom_boxplot()+
  labs(x = "Wave",y="Problem çözme bilgisi")
```

Küresellik varsayımını test etmek için *ezANOVA* tarafından verilen Mauchy testi kullanılmıştır.

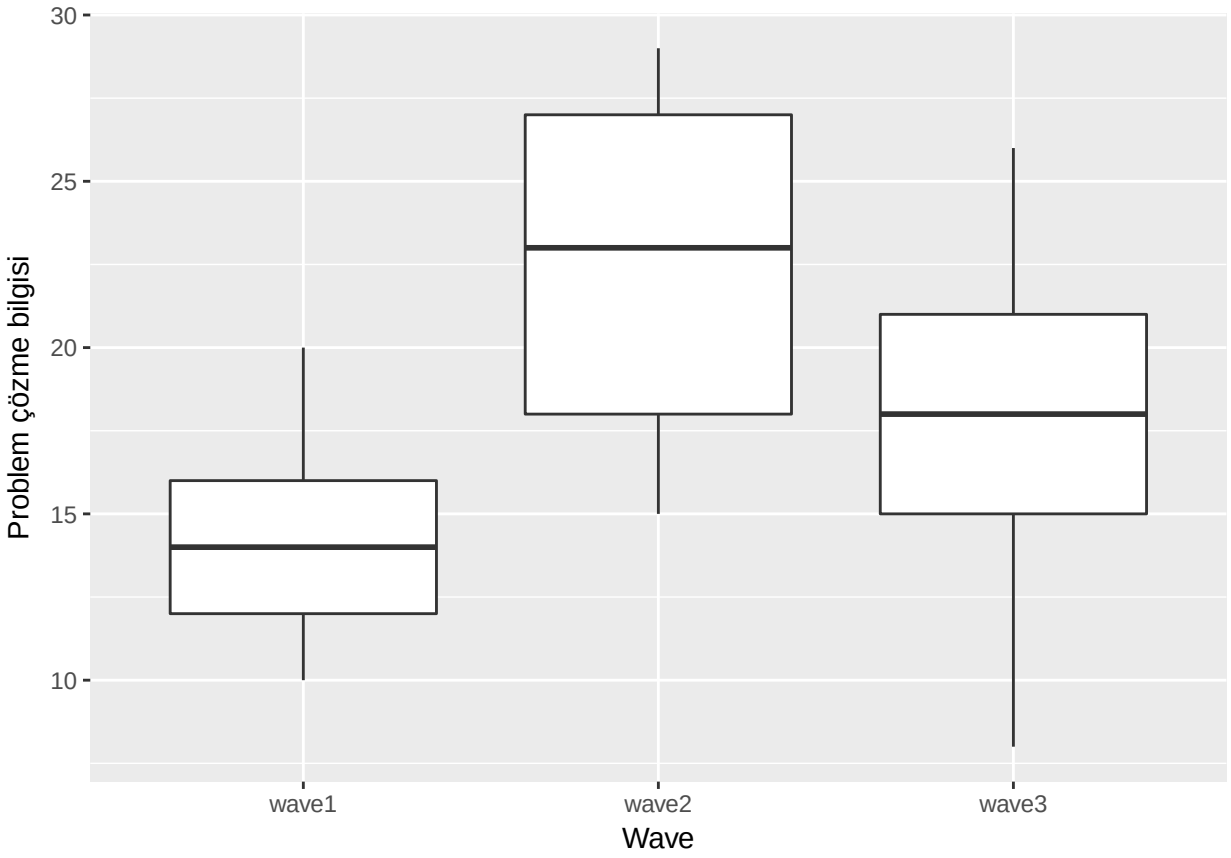


Figure 9.5: Problem çözme bilgisi

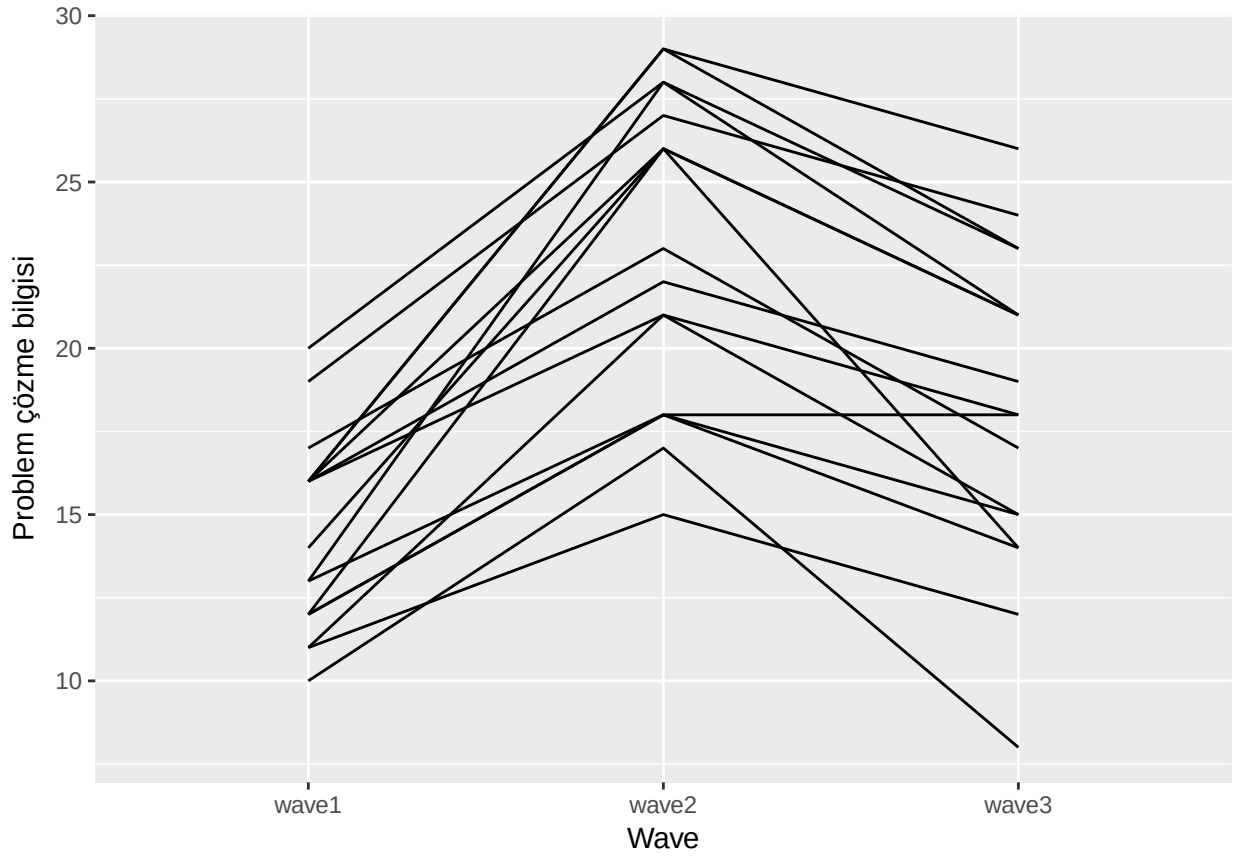


Figure 9.6: Problem çözme bilgisi çizgi grafigi

```
require(ggplot2)
ggplot(data_long, aes(x=wave, y=PrbSol, group=id))+
geom_line() + labs(x = "Wave",y="Problem çözme bilgisi")
```

Bu grafik, $\eta\beta_{ij}$'nin sıfır olmayacağı şeklinde yorumlanabilir. *asbio* paketinde (Aho (2016)) yer alan tukey.add.test fonksiyonu H_0 *: asıl etkiler ve bloklar eklemeli ilerler* boş hipotezini test etmek için kullanılabilir. Fakat daha önce beirtildiği gibi bu varsayımın ihlali, küresellik varsayımı ihlal edilmediği veya düzeltmesi yapıldığı sürece, öenmsizdir.

```
library(asbio)
with(data_long,tukey.add.test(PrbSol,wave,id))
##
## Tukey's one df test for additivity
## F = 5.943   Denom df = 31    p-value = 0.021

# eğer eklemelilik mevcut ise  rassal bloklar tasarısı kullanılabilir
#additive=with(data_long,lm(PrbSol~id+wave))
#anova(additive)
```

Tukey eklemelilik testi boş hipotezin terkedilebileceğini dolayısıyla eklemesiz modelin daha uygun olduğunu göstermiştir.

Basamak 3: Varyans analizi (küresellik ve hataların normal dağılımı varsayımı kontrolleri ile birlikte).

```
library(ez)
#birinci yol ezANOVA fonksiyonu

alternative1 = ezANOVA(
    data = data_long,
    wid=id, dv = PrbSol, within = wave,
    detailed = T,return_aov=T)

alternative1
## $ANOVA
##        Effect DFn DFd   SSn SSd     F        p p<.05   ges
## 1 (Intercept)   1  16 17510 680 412.0 7.62e-13     * 0.954
## 2        wave   2  32   647 169  61.2 1.16e-11     * 0.433
##
## $`Mauchly's Test for Sphericity`
##   Effect     W     p p<.05
## 2   wave 0.918 0.526
##
## $`Sphericity Corrections`
##   Effect   GGe    p[GG] p[GG]<.05  HFe    p[HF] p[HF]<.05
## 2   wave 0.924 6.17e-11         * 1.04 1.16e-11         *
##
## $aov
##
## Call:
## aov(formula = formula(aov_formula), data = data)
##
## Grand Mean: 18.5
##
## Stratum 1: id
##
## Terms:
##                 Residuals
## Sum of Squares        680
## Deg. of Freedom        16
##
## Residual standard error: 6.52
##
## Stratum 2: id:wave
##
## Terms:
##                  wave Residuals
## Sum of Squares   647       169
## Deg. of Freedom    2        32
##
## Residual standard error: 2.3
## Estimated effects may be unbalanced
```

```
PrbSolres=sort(alternative1$aov$id$residuals)
qqnorm(PrbSolres);qqline(PrbSolres)
```

Hataların dağılımı normal sayılabilir.

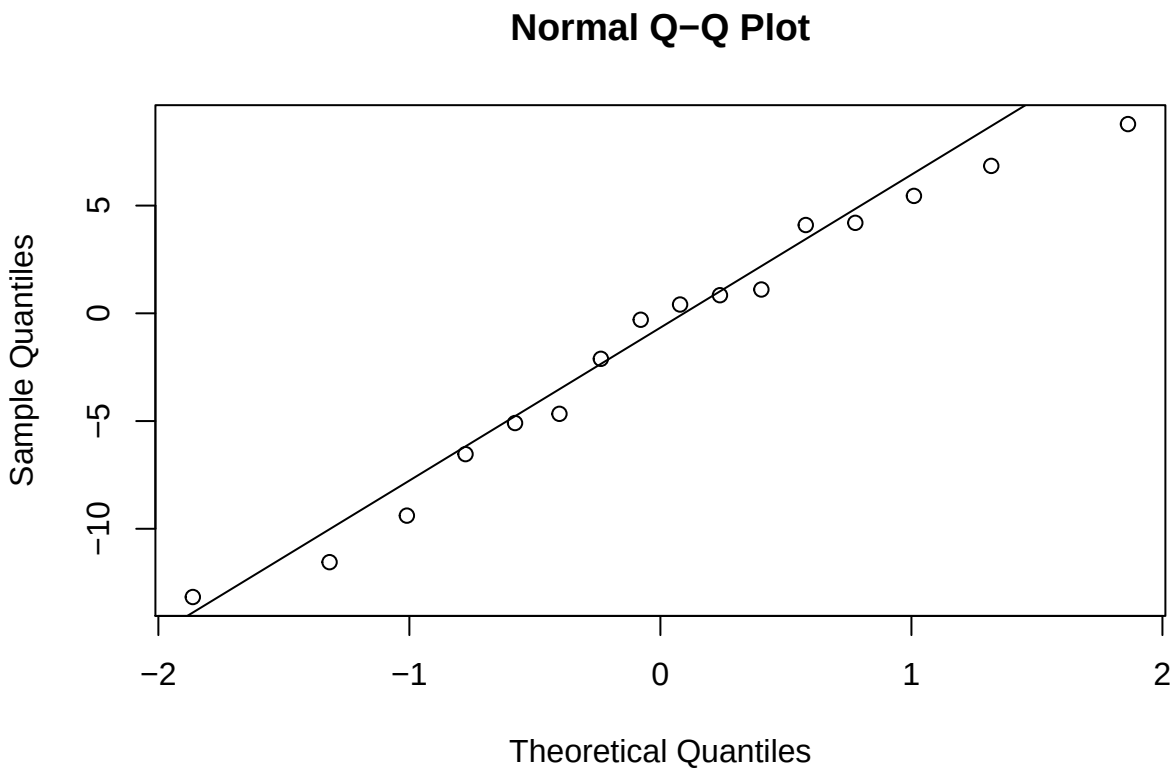


Figure 9.7: Problem çözme bilgisi hata terimleri

```
# ikinci yol aov fonksiyonu
summary(aov(PrbSol ~ wave + Error(id/wave), data=data_long))
##
## Error: id
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 16    680    42.5
##
## Error: id:wave
##           Df Sum Sq Mean Sq F value  Pr(>F)
## wave       2    647     324    61.2 1.2e-11 ***
## Residuals 32    169       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.4.0.3 Dirençli tahminleme yöntemi: tek-yönlü bağlı gözlemler varyans analizi

Wilcox (2012) tarafından bir araya toplanan dirençli prosedürlerden bir tanesi Mair and Wilcox (2016) paketi ile kullanılabilecek *rmanova* fonksiyonu ile tamamlanabilir. Kırpılmış ortalamalar için farklı-varyanslı(heteroscedastic) ve tek yönlü tekrarlanan ölçümler ANOVA yöntemini kullanan bu fonksiyonun detayları için inceleyiniz ;*?rmanova*

```
library(WRS2)

#rmanova
# 20% kırpılmış
with(data_long,rmanova(PrbSol,wave,id,tr=.20))
## Call:
## rmanova(y = PrbSol, groups = wave, blocks = id, tr = 0.2)
##
## Test statistic: 34.9
## Degrees of Freedom 1: 1.9
## Degrees of Freedom 2: 19
## p-value: 0
```

9.4.0.4 Example writeup tek-yönlü bağlı gözlemler varyans analizi

Her bir ölçme durumu için betimleyici istatistikler Tablo 6 ile verilmiştir. Kovaryans matrisi tablo 7 ile verilmiştir. Tek-yönlü bağlı gözlemler varyans analizi raporlanmıştır. F testi alfa=0.05 ile tamamlanmıştır. Varsayım ihlali tespit edilmemiştir ve ölçme durumları arasında anlamlı bir fark bulunmuştur. $F(2, 32) = 61.2, p < .001$, genelleştirilmiş eta kare değeri ($\hat{\eta}^2_G$) 0.43 olarak hesaplanmıştır.

9.4.0.5 Takviye çözümlemeler: Tek-yönlü bağlı gözlemler varyans analizi

Eklenecek

9.4.0.6 Kayıp veri teknikleri: Tek-yönlü bağlı gözlemler varyans analizi

Eklenecek

9.4.0.7 İstatistiksel güç: Tek-yönlü bağlı gözlemler varyans analizi

Eklenecek

9.5 Karma tasarı (Mixed design)

Eklenecek

Chapter 10

Korelasyon

Değişkenlerin birbiri ile olan ilişkilerini açıklamak çoğu araştırmacının ilgisini çekmiştir. İki değişkene ait çarpımlar toplamı (sum of cross products) $S_{XY} = \sum(X - \bar{X})(Y - \bar{Y})$ iki değişken arasındaki ilişki hakkında sınırlı olsa da bilgi verebilir. Örneğin Şekil 10.1 X ve Y değişkeni arasındaki ilişkiyi gösterir ve çarpımlar toplamı sıfırdır.

```
##        x     y deviationX deviationY crossPRODUCT
## 1   1.00  0.00       0.93       0.00         0.00
## 2   0.90  0.43       0.83       0.43         0.36
## 3   0.62  0.78       0.56       0.78         0.44
## 4   0.22  0.97       0.16       0.97         0.15
## 5  -0.22  0.97      -0.29       0.97        -0.28
## 6  -0.62  0.78      -0.69       0.78        -0.54
## 7  -0.90  0.43      -0.97       0.43        -0.42
## 8  -1.00  0.00      -1.07       0.00         0.00
## 9  -0.90 -0.43      -0.97      -0.43         0.42
## 10 -0.62 -0.78      -0.69      -0.78         0.54
## 11 -0.22 -0.97      -0.29      -0.97         0.28
## 12  0.22 -0.97       0.16      -0.97        -0.15
## 13  0.62 -0.78       0.56      -0.78        -0.44
## 14  0.90 -0.43       0.83      -0.43        -0.36
## 15  1.00  0.00       0.93       0.00         0.00
```

İki değişken arasındaki kovaryans ise $Cov_{XY} = S_{XY}/n - 1$ ile hesaplanabilir. Fakat kovaryans ölçülen değişkenlerin skalasına bağımlıdır. Bir diğer ifade ile, değişkenlerin sayısal değerleri arttıkça kovaryans artar. Bu durum kovaryans yorumunu zorlaştırır. Bir korelasyon katsayısı ise genellikle -1 ve 1 arasındadır ve sınırları olduğu için yorumlaması daha kolaydır.

10.1 Pearson korelasyon katsayısı

Pearson 1986 yılında bir korelasyon katsayısı hesaplama yöntemi tanıtmıştır. Bu katsayı -1 ile +1 arasında değişir ve $Cov_{XY}/S_X S_Y$ ile hesaplanabilir. Bu katsayı iki değişken arasındaki doğrusal ilişkiyi ölçer. Şekil 10.1 aralarındaki korelasyonun sıfır olduğu iki değişken ile çizilmiştir. Aslında şekil içerisindeki X ve Y bir biri ile ilişkisiz değillerdir çünkü şekil kabaca bir çemberdir. X ve Y beraber bir çember oluşturabilecek ilişkiye sahip oldukları halde , ilişki doğrusal olmadığından, aralarındaki korelasyon sıfırdır. Şekil 10.2 aralarında ilişki olan değişkenlere örnekler gösterir,(A) kusursuz pozitif korelasyon, +1, (B) 0.7 pozitif korelasyon, (C) 0 korelasyon, (D) -0.4 korelasyon ve (E) kusursuz negatif korelasyon,-1.

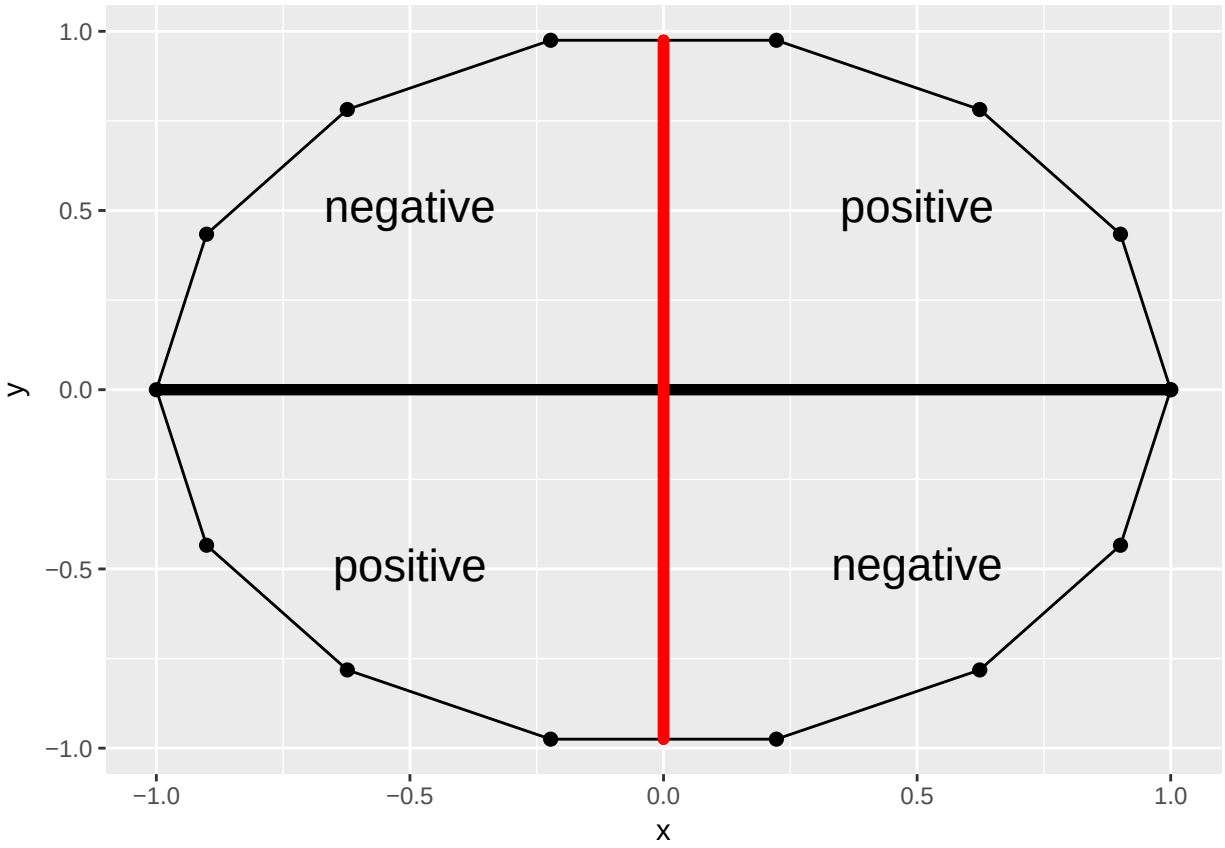


Figure 10.1: Çarpimlar toplami=0

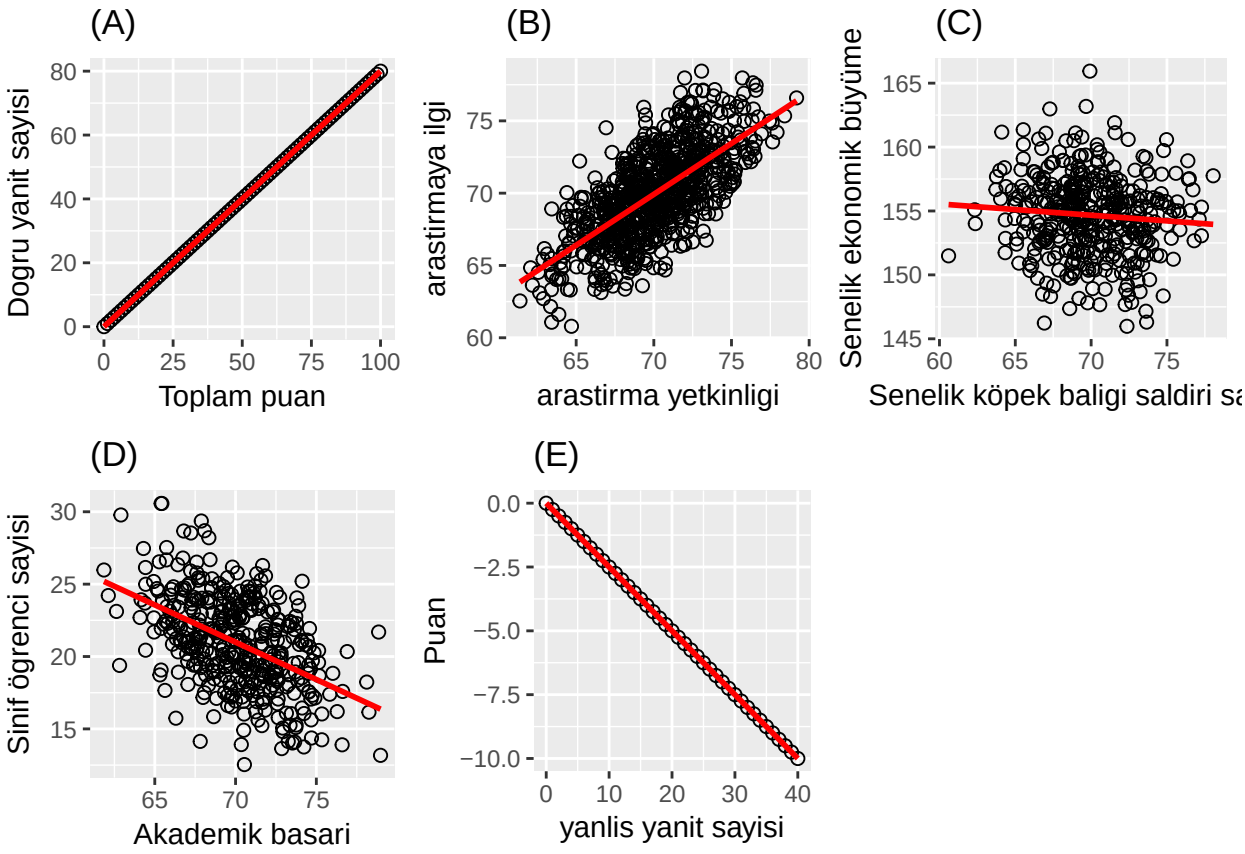


Figure 10.2: Correlation examples

10.1.1 Pearson korelasyon katsayısının evren bazında yorumu

Örneklemden gelen bilgi (r) , evren (ρ) düzeyinde çıkarım yapmak zere kullanılabilir.

z transformasyonu . İkili normallik (bivariate normality) varsayımı ve en az 10 örneklem ile z transformasyonu kullanılarak evrene ait parametre hakkında yorum yapılabilir (Myers et al. (2013)). Transformasyon formülü

$$z_r = \frac{1}{2} ln \left(\frac{1+r}{1-r} \right)$$

Standart hata

$$\sigma_r = \frac{1}{\sqrt{n-3}}$$

Güven aralığı $z_r \pm z_{\alpha/2}\sigma_r$. Korelasyon katsayısı yorumunu kolaylaştırması amacı ile ters transformasyon $r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$.

$H_0 : \rho = 0$ boş hipotezi normal bir dağılımın uygun olduğu varsayımı ile sınanabilir;

$$z = \frac{z_r - z_{\rho_{null}}}{\frac{1}{\sqrt{n-3}}}$$

t dağılımı da $H_0 : \rho = 0$ boş hipotezini test etmek için kullanılabilir.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Bu istatistik $n-2$ serbestlik derecesine sahip t dağılımını takip eder.

10.1.2 R betiği: Pearson korelasyon katsayısı

Gösterim amaçlı dataWBT (2.3) içerisinde yer alan Bayburt ilçesi seçilmiştir. TCA puanları ile kişi başı senelik gelir arasındaki korelasyon incelenmiştir.

```
# CSV yükle
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#URL sil
rm(urlfile)

#Bayburt ilini seç
#  sıralı silme uygula (listwise deletion)
dataWBT_Bayburt=dataWBT[dataWBT$city=="BAYBURT",]
#hist(dataWBT_Bayburt$income_per_member)
```

İkili normal dağılım @ref(fig:testbivarnorm zerinden incelenebilir. *rgl* (Adler and Murdoch (2017)) paketi ile oluşturulan bu grafik interaktiftir, fare ile inceleyiniz.

```
## wgl
##   3
```

İkili normallik varsayımı gerçekci görünmüyor. Karşılaştırmanız amacı ile 10.4 korelasyonun 0.7 olduğu bir ikili normal dağılımı gösterir. Varsayım ihlalinin sonuçları etkileyebileceğini göz ardı ederek, gösterim amaçlı, eldeki veri ile $H_0 : \rho = 0$ boş hipotezi $H_1 : \rho \neq 0$ alternatif hipotezine karşı test edilmiştir.Saçılım grafiği Şekil 10.5 ile verilmiştir.

Figure 10.3: Bayburt TCA ve Gelir

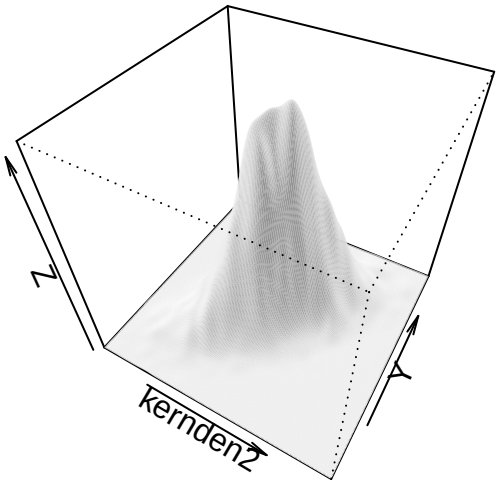


Figure 10.4: Ikili normal dagilim

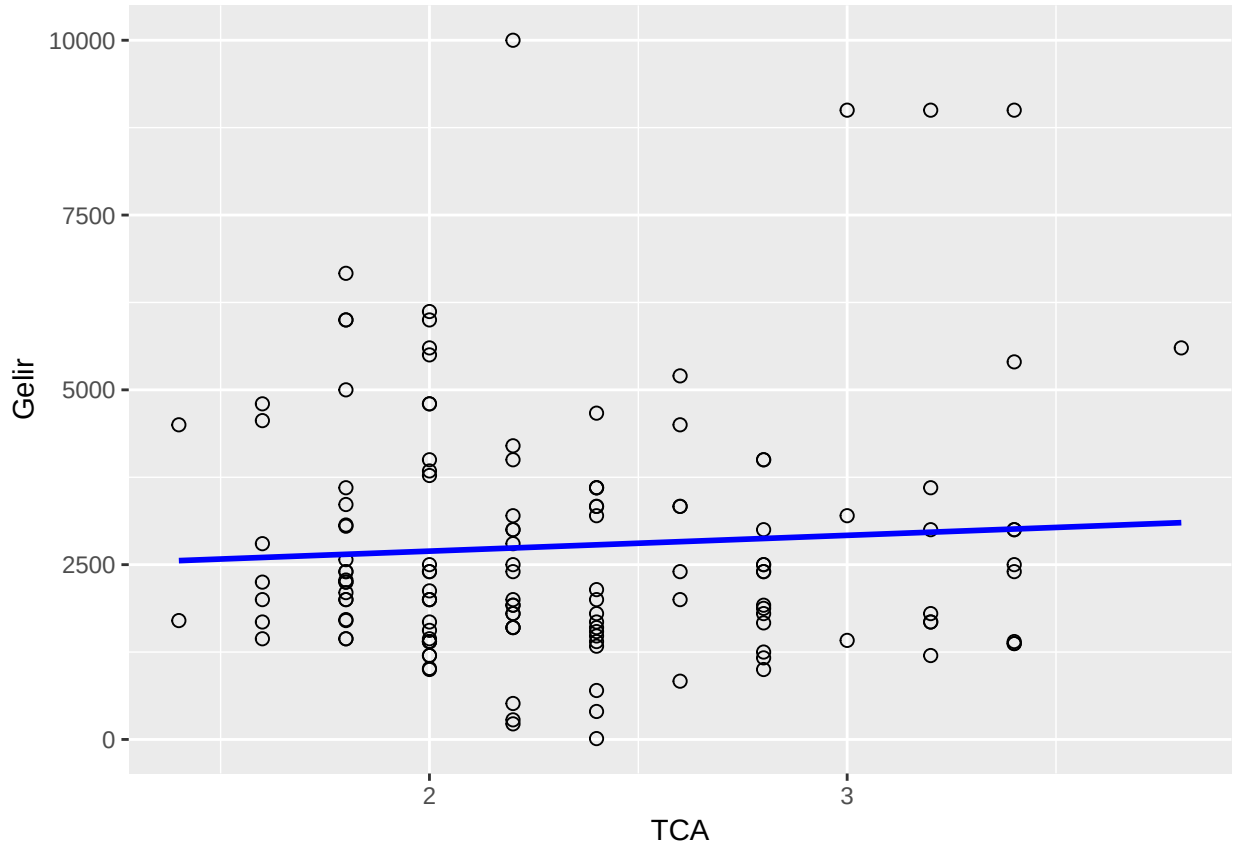


Figure 10.5: Bayburt TCA ve gelir

Bu iki değişken arasındaki korelasyon *stats* paketinde (R Core Team (2016b)) yer alan *cor* fonksiyonu ile hesaplanabilir. Aynı paket içerisinde yer alan *cor.test* fonksiyonu t testi sonuçlarını ve z transformasyonu ile hesaplanmış güven aralığı hesaplarını rapor eder.

```
#?cor komutu ile use = "complete.obs" argümanının ikili silme kullandığını görebilirsiniz

with(dataWBT_Bayburt,cor(gen_att,income_per_member,
                         use = "complete.obs",method="pearson"))
## [1] 0.0664

with(dataWBT_Bayburt,cor.test(gen_att,income_per_member,
         alternative = "two.sided",
         method="pearson",
         conf.level = 0.95,
         na.action="na.omit"))
##
##  Pearson's product-moment correlation
##
## data:  gen_att and income_per_member
## t = 0.8, df = 100, p-value = 0.4
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.102  0.232
## sample estimates:
##    cor
## 0.0664
```

Eğer kendiniz hesaplamak isterseniz, $H_0 : \rho = 0$ ve $H_0 : \rho \neq 0$;

```
sample_r=0.06641641
r0=0        #boş hipotez
sample_n=137       # örneklem sayısı
zr=(0.5)*log((1+sample_r)/(1-sample_r))  # z transformasyonu
z0=(0.5)*log((1+r0)/(1-r0))  # z transformasyonu
sigmar=1/(sqrt(sample_n-3))

#z istatistiği
(zr-z0)/sigmar
## [1] 0.77


ll=zr-(qnorm(0.975)*sigmar)  # alt limit


ul=zr+(qnorm(0.975)*sigmar)  # üst limit


(exp(2*ll)-1)/(exp(2*ll)+1)  #ters transform
## [1] -0.102
(exp(2*ul)-1)/(exp(2*ul)+1)  #ters transform
## [1] 0.232


t=sample_r*(sqrt((sample_n-2)/(1-sample_r^2)))
qt(c(.025, .975), df=(sample_n-2))
```

```
## [1] -1.98  1.98
p.value = 2*pt(-abs(t), df=sample_n-2)
p.value
## [1] 0.441
```

Yüzdeli bootstrap yönetimi varsayım ihlallerine dirençli bir yöntem olabilir (Myers et al. (2013)).

```
#Normallik varsayımı olmadan bootstrap ile %95 güven aralığı hesabı
set.seed(31012017)
B=5000       #  bootstraps tekrarı
alpha=0.05   # alfa

#TCA ve gelir
originaldata=dataWBT_Bayburt2

# id ekle
originaldata$id=1:nrow(originaldata)

output=c()
for (i in 1:B){
  #sample rows
  bs_rows=sample(originaldata$id,replace=T,size=nrow(originaldata))
  bs_sample=originaldata[bs_rows,]
  output[i]=cor(bs_sample$gen_att,bs_sample$income_per_member)
  }
output=sort(output)

## Yönsüz
# alt limit
output[as.integer(B*alpha/2)]
## [1] -0.138

# d yıldız üst
output[B-as.integer(B*alpha/2)+1]
## [1] 0.252
```

Yüzdeli bootstrap dışında alternatif dirençli yöntemler mevcuttur. Wilcox (2012) dirençli korelasyon katsayısı hesaplama yöntemlerini bir araya toplamıştır. WRS2 paketi *pbcor* ve *wincor* fonksiyonları incelenebilir.

```
#  WRS2 paketi
library(WRS2)
pbcor(dataWBT_Bayburt2$gen_att,dataWBT_Bayburt2$income_per_member,beta=.2)
## Call:
## pbcor(x = dataWBT_Bayburt2$gen_att, y = dataWBT_Bayburt2$income_per_member,
##     beta = 0.2)
##
## Robust correlation coefficient: -0.0351
## Test statistic: -0.407
## p-value: 0.684

wincor(dataWBT_Bayburt2$gen_att,dataWBT_Bayburt2$income_per_member,tr=.2)
## Call:
## wincor(x = dataWBT_Bayburt2$gen_att, y = dataWBT_Bayburt2$income_per_member,
##     tr = 0.2)
##
```

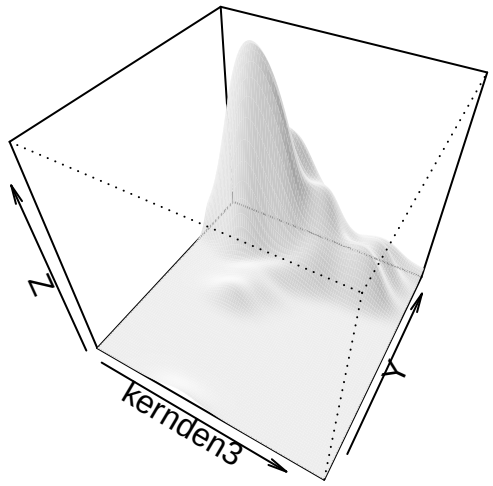


Figure 10.6: Transform edilmis Gelir

```
## Robust correlation coefficient: -0.0197
## Test statistic: -0.229
## p-value: 0.82
```

Rapor örneği: Bayburt ilinde yaşayan katılımcılar göz önünde bulundurularak, TCA puanları ile gelir düzeyi değişkenleri arasında korelasyon olmadığı boş hioptezi test edilmiştir. Pearson korelasyon katsayısı $r = .066(p = .44)$ ve bu katsayı için %95 güvern aralığı [-.10,.23] olarak hesaplanmıştır.İki değişken arasında korelasyonun sıfır olduğunu ileri süre boş hipotez terkedilmemiştir. Aynı çıkarıma 5000 tekrarlı boostrap yöntemi ile de ulaşılmıştır (%95 güven aralığı [-.138 to .252]).

Not: Farklı işaret Pearson katsayısı istatistiksel olarak sıfırdan farklı değildirfakat işareti pozitiftir (.066). WRS paketinde yer alan fonksiyonlar da korelasyonun sıfırdan farklı olmadığı sonucuna ulaşmıştır. Fakat hesaplanan korelasyon negatiftir. Gelir değişkenine ait dağılımın çarpık olduğu ikili normallik dağılım grafiğinde de göze çarpmaktadır. World Bank araştırma takımı bu değişkenin çarpık dağılım göstermesi sebebi ile değişkeni trabnsform ederek analiz etmişlerdir (Hirshleifer et al. (2016)). Benzer transformasyonu kullanırsak;

```
with(dataWBT_Bayburt2,cor.test(gen_att,incomeTC,
         alternative = "two.sided",
         method="pearson",
         conf.level = 0.95,
         na.action="na.omit"))
##
```

```
##  Pearson's product-moment correlation
##
## data:  gen_att and incomeTC
## t = -0.009, df = 100, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.169  0.167
## sample estimates:
##      cor
## -0.00081
```

Transformasyon sonucunda gelir değişkeni daha normal bir dağılım göstermiştir ve Pearson korelasyon katsaysının işareti negatiftir.

10.2 Spearman rho ve Kendall tau

Verilerin sıralı veri olması durumunda veya sürekli değişkenlerde aykırı değerlerin etkisi azaltılmak istenildiğinde Spearman rho veya Kendall tau kullanılabilir. Burada bahsedilen aykırı değerlerin etkisinin azaltılması durumu Pearson korelasyona kıyasla geçerlidir. Aykırı değerlere karşı rho ve tauya nazaran daha iyi koruma sağlayan dirençli yöntemler mevcuttur, Wilcox (2012).

10.2.1 R betiği: Spearman's rho ve Kendall's tau

Pearson korelasyon katsayısı hesaplama örneğinde kullanılan TCA puanları ve gelir ilişkisi için Spearman rho hesaplaması;

```
with(dataWBT_Bayburt,cor.test(gen_att,income_per_member,
         alternative = "two.sided",
         method="spearman",
         conf.level = 0.95,
         na.action="na.omit",
         exact=FALSE))
##
##  Spearman's rank correlation rho
##
## data:  gen_att and income_per_member
## S = 5e+05, p-value = 0.6
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##     rho
## -0.0508
```

Sıralama verisinde eş-sıra (ties) durumu var ise *cor.test* fonksiyonu Spearman rho hesaplanışında düzeltme yapar fakat p değeri hesaplamaz. Eğer *exact=FALSE* argümanı kullanılırsa t dağılımı üzerinden p değeri hesaplanabilir. Field et al. (2012) eş-sıralılık durumunun çok olması durumunda Kendall tau hesaplanmasını önerir.

```
#use ?cor to see use="complete.obs" is doing casewise deletion
with(dataWBT_Bayburt,cor.test(gen_att,income_per_member,
         alternative = "two.sided",
         method="kendall",
         conf.level = 0.95,
```

```
         na.action="na.omit",
         exact=FALSE))
##
##  Kendall's rank correlation tau
##
## data:  gen_att and income_per_member
## z = -0.6, p-value = 0.5
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##     tau
## -0.0373
```

exact=FALSE argümanı *method="kendall"* ile kullanılınca normal dağılım üzerinden p değeri hesaplar.

Gelir değişkeninde bulunan aykırı değerlerin etkisini azaltmak üzere Spearman ve Kendall korelasyon katsayıları hesaplanmıştır. Spearman rho değeri $r_S = -.051(p = .56)$ ve Kendall tau değeri $\tau = -.037, p = .54$ olarak bulunmuştur.

10.3 R betiği: Çift Serili ve Nokta-Çift Serili Korelasyonlar

Çift serili korelasyon bir sürekli değişken ve örtük bir sürekli değişkeni yansıtan ikili değişken (dichotomous) arasındaki korelasyonu hesaplamak için kullanılabilir. Örneğin öğrencilerin bir soruya verdiği doğru veya yanlış yanıt değişkeni ile toplam puanlar arasındaki ilişki çift serili korelasyon ile hesaplanabilir.

Gösterim amaçlı veri setinde yer alan birinci TCA sorusunu ikili veri olarak kodlayalım[1]. Bu ikili değişken ile 2.,3.,4.,5. ve 6. soruların ortalaması olan TCA puanları arasındaki ilişki *psych* paketinde (Revelle (2016)) yer alan *biserial* fonksiyonu ile hesaplanabilir.

```
dataWBT_Bayburt$binitem1=ifelse(dataWBT_Bayburt$item1==4,1,0)
require(psych)
with(dataWBT_Bayburt,biserial(gen_att,binitem1))
##       [,1]
## [1,] 0.317
```

Nokta çift serili korelasyon ise sürekli bir değişken ve ikili bir değişken arasındaki ilişkiyi ölçmek için kullanılabilir. *cor.test* fonksiyonu ve *method="pearson"* argumanı ile nokta çift serili korelasyon hesaplanabilir. TCA puanları ve cinsiyet arasındaki nokta-çift serili korelasyon;

```
dataWBT_Kayseri=dataWBT[dataWBT$city=="KAYSERI",]
dataWBT_Kayseri$genderNUM=ifelse(dataWBT_Kayseri$gender=="Female",1,0)
with(dataWBT_Kayseri,cor.test(gen_att,genderNUM,
         alternative = "two.sided",
         method="pearson",
         conf.level = 0.95,
         na.action="na.omit"))
##
##  Pearson's product-moment correlation
##
## data:  gen_att and genderNUM
## t = -7, df = 200, p-value = 2e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.487 -0.277
```

[1] World Bank araştırma takımı bu soruyu ikili değişken olarak yeniden kodlayarak analiz etmiştir

```
## sample estimates:
##    cor
## -0.387
```

10.4 R betiği: Phi korelasyon katsayısı

Değişkenler ikili ise aralarındaki ilişki phi(ϕ) korelasyon katsayısı ile ölçülebilir. Gösterim amaçlı cinsiyet ve maaş durumu ilişkisi incelenmiştir. Maaş durumu katılımcının son 12 ay boyunca maaş alıp almadığını gösterir. *psych* paketinde yer alan *phi* fonksiyonu 2x2 frekans tablosu üzerinden hesaplama yapabilir.

```
dataWBT_Kayseri=dataWBT[dataWBT$city=="KAYSERI",]
table(dataWBT_Kayseri$gender,dataWBT_Kayseri$wage01)
##
##           No Yes
##   Female  52  97
##   Male    49  54
##   Unknown  0   0

genderWAGE=matrix(c(52,49,97,54),ncol=2)
library(psych)
phi(genderWAGE)
## [1] -0.13
```

10.5 Tetrakorik ve polikorik korelasyon katsayısı

Aralarında korelasyon belirlenmesi istenen değişkenlerin doğasının sürekli ve normal dağılımlı olduğu hâlde iki kategorili (dikatom) olarak gözlenmiş olması durumunda, bir diğer ifade ile, dikatom değişkenlerin örtük sürekli değişkenleri yansıtıyor olması durumunda, tetrakorik (rt) korelasyon katsayısı kullanılır. Bu noktada araştırmacı, karşılaştığı dikatom değişkenlerin doğası gereği sürekli olup olmadığına karar vermesi ve kararını savunabilmesi gerekir. Örneğin öğrencilerin doğru-yanlış sorularına verdiği yanıtlar (0: yanlış, 1: doğru) sürekli bir değişkenin yansıması olarak düşünülebilir. Diğer bir örnek ankete katılan bireylerin yaşlarının 30 yaş altı ve 30 yaş üstü olarak toplanmış olmasıdır. Gösterim amaçlı, veri setinde yer alan üçüncü ve altıncı TCA sorularını ikili veri olarak kodlayarak ve psych paketinde yaralan tetrachoric fonksiyonunu kullanılarak tetrakorik korelasyon katsayısı hesaplanabilir. Hesaplanan korelasyon 0.07'dir.

```
# 3. ve 6. sorular dikatom yapılsın
# Dünya Bankası araştırma grubu tarafından kullanılan yöntem,
# eğer yanıt 1 (strongly disagree) veya 2 (disagree) ise 1, değilse 0.
dataWBT_Kayseri$Bitem3=ifelse(dataWBT_Kayseri$item3==1|dataWBT_Kayseri$item3==2,1,0)
dataWBT_Kayseri$Bitem6=ifelse(dataWBT_Kayseri$item6==1|dataWBT_Kayseri$item6==2,1,0)
require(psych)
tetrachoric(as.matrix(dataWBT_Kayseri[,c("Bitem3","Bitem6")]))
## Call: tetrachoric(x = as.matrix(dataWBT_Kayseri[, c("Bitem3", "Bitem6")]))
## tetrachoric correlation
##        Bitm3 Bitm6
## Bitem3 1.00
## Bitem6 0.07  1.00
##
##  with tau of
## Bitem3 Bitem6
##  -0.23   0.54
```

Polikorik korelasyon katsayısı eldeki değişkenlerin sıralı (ordinal) kategorik olduğu durumlarda hesaplanır. Tetrakorik korelasyonda olduğu gibi, polikorik korelasyon hesaplanırken de kategorik değişkenlerin sürekli bir değişeni yansıttığı varsayımı yapılır. Bu iki korelasyon katsayısı örtük sürekli korelasyonlar (latent continuous correlation) olarak tek bir terim altında düşünülebilir (Uebersax (2015)). Tetrakorik ve polikorik korelasyonların hesaplanmasında kullanılan yöntemler kapalı ve açık form olarak düşünülebilir. Kapalı form nispeten daha basittir ve formüllerle basitçe ifade edilebilir. Fakat kapalı formlar genellikle yaklaşık değerler hesaplar. Açık form yöntemlerden kasıt iteratif prosedürlerdir ve oldukça karmaşık olabilirler, fakat kapalı formlara kıyasla daha doğru sonuçlar vermesi beklenir. Kategorik değişkenler için yürütülen faktör analizlerinde bu iki tür korelasyon birçok yazılım tarafından kullanılmaktadır. Tetrakorik veya polikorik korelasyon hesaplayacak olan araştırmacıların kullandığı yöntemi ve yazılımı açık olarak ifade etmesi doğru bir yaklaşım olacaktır, çünkü farklı yazılımlar ve farklı yöntemler kullanıldığında sonuçlar arasında farklılık görülebilir. Sonuçların birbirinden neden farklı olabileceğini daha detaylı araştırmak isteyen okuyucular Olsson (1979) tarafından kaleme alınan makaleyi inceleyebilirler. Bu noktada okuyucular psych paketinde yer alan polychoric fonksiyonunun argümanlarını dikkatlice incelemek isteyebilirler. Gösterim amaçlı veri setinde yer alan üçüncü ve altıncı TCA sorular için polikorik korelasyon hesaplanmış ve 0.16 bulunmuştur.

```
require(psych)
polychoric(as.matrix(dataWBT_Kayseri[,c("item3","item6")]))
## Call: polychoric(x = as.matrix(dataWBT_Kayseri[, c("item3", "item6")]))
## Polychoric correlations
##       item3 item6
## item3 1.00
## item6 0.16  1.00
##
##  with tau of
##           1     2    3
## item3 -0.72  0.23 1.30
## item6 -1.37 -0.54 0.82
```

10.6 Korelasyon katsayısı hakkında dikkat edilmesi gerekenler

Sebep-Sonuç Bir korelasyon katsayısı sebep-sonuç belirtmez. Sebep-sonuç ilişkisi kurulmak istense dahi dört farklı senaryo mevcuttur, (a) X Y'yi etkiler, (b) Y X'i etkiler, (c) X ve Y bir veya daha fazla ortak sebebin sonucudur, (d) X ve Y farklı sebeplerle ortaya çıkmıştır fakat bu farklı sebepler ilişkilidir.

Büyüklük Bir korelasyon katsayısının büyük veya küçük oluşu konuya göre değişir. Birbirine benzemesi üzerine tasarlanmış iki matematik sınavından sonra hesaplanan korelasyon 0.6 ise bu küçük bir korelasyon olarak düşünülebilir çünkü paralel formlar korelasyonunun en az .70 olması beklenir. Fakat ALES puanları ile yüksek lisans not ortalaması arasında hesaplanacak bir 0.6 korelasyon oldukça büyüktür çünkü alan yazında bu değer genellikle .1 ve .3 arasındadır.

Aykırı değerler Veri setinde yer alan aykırı değerler korelasyon katsayılarını etkiler.

Güvenirlik X veya Y ölçme hatası (measurement error) içeriyorsa hesaplanan korelasyon aşağı yönde etkilenir. Düzeltme yapmak için

$$r_{T_xT_y} = \frac{r_{xy}}{\sqrt{(r_{xx}r_{yy})}}$$

kullanılabilir, r_{xx} ve r_{yy} X ve Y için güvenirlik katsayılarıdır.

- *Bu düzeltme ne zaman kullanılmaz:* Gerçek hayatta yeri olacak bir karar verilecekse bu düzeltme yapılmamalıdır. Kararlar gözlemlenen veriler üzerine verilmelidir.
- *Bu düzeltme ne zaman kullanılır:* Teorilerin geliştirilmesi aşamasında düzeltme kullanılabilir.

Varyans Korelasyon katsayısı değişkenlerin varyansından etkilenir. Varyansın yapay olarak küçültülmesi korelasyonun da küçülmesine sebep olur. Varyansın yapay olarak küçülmesine örnekler;

- Sürekli verilerin kategorize edilmesi
- Ranj sınırlılığı
- Taban ve Tavan etkisi (Floor and Ceiling Effects)

Chapter 11

Çoklu Doğrusal Regresyon , Kısa Tanıtım

Bilimsel ilerleme bilginin güvenilir şekilde bir çalışmadan diğerine aktarılmasını gerektirir. Galileo'nun 350 sene önce dile getirdiği gibi bu aktarma keskinliği olan formal bir lisan ile yapılmalıdır. Pearl (2009)

Bu alıntıda yer alan *formal lisanlardan* biri matematiksel eşitliklerdir. Örneğin Galton 19. yüzyılda anne ve yavru bezelye tanelerinin büyüklüğü arasındaki ilişkiyi matematiksel eşitlikler ile açıklamaya çalışmıştır. Galton'un çalışmaları Pearson'ın çalışmalarına öncü olmuş ve ortaya regresyon fikri çıkmıştır.[1].

Web of Science veri tabanında ,sadece 2016 yılı içerisinde, 60 binden fazla bilimsel makalenin özünde regresyon kelimesi yer almıştır. Alan yazın oldukça geniştir. Regresyon moellerinin bu kadar sık kullanılmasının sebebi , değişkenlerin arasındaki ilişkilerin sıradan bir korelasyon ile açıklanamayacak kadar karmaşık olmasıdır. Bir kitap bölümünde regresyonun bütün alt başlıkları ile ele alınması gerçekci değildir. Bu bölümde oldukça basit bir çoklu doğrusal regresyon modeli tanıtılmıştır.

11.1 Matrisler ve En Küçük Kareler Yöntemi

Regresyon modeli işlem basamaklarını matrisler ve en küçük kareler yöntemi (OLS) ile göstermenin iki avantajı vardır, (a) sürecin basamakları kolayca takip edilebilir ve (b) regresyon çözümlemesi konusunda daha üst düzey modelleri çalışmak isteyenler için sağlam bir temel oluşturabilir. Burdan sonraki bölümler iki farklı veri üzerinden devam edecektir. İlk veri seti sadece 12 katılımcıdan oluşur ve sentetik veri seti olarak isimlendirilmiştir. İkinci veri seti daha çok katılımcı içerir, gerçekci bir veri setini temsil edebilir ve simulasyon verisi olarak isimlendirilmiştir.

Araştırmacının bir bağımlı değişken (yordanan) ve iki farklı bağımsız değişken (yordayıcı) arasındaki ilişkiyi incelemek istediğini varsayalım. Bütün değişkenlerin sürekli değişken olduğunu düşünelim. Bu durumda regresyon modeli;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

i katılımcıları, i=1,...,n; Y bağımlı değişkeni; X_1 ve X_2 bağımsız değişkenleri; β regresyon katsayılarını ve ϵ hata terimini temsil eder. Bu model matris eşitliği olarak da yazılabilir.

$$Y = X\beta + \epsilon$$

[1] http://ww2.amstat.org/publications/jse/v9n3/stanton.html

Bu genel eşitlikte bütün bağımsız değişkenler X matrisi ile ve bütün regresyon katsayıları da β matrisi ile temsil edilir. Araştırmacının veri seti şu şekilde olsun

id	Y	X1	X2
ind 1	8	0	3
ind 2	4	-2	1
ind 3	6	6	3
ind 4	6	-2	0
ind 5	5	5	0
ind 6	9	4	2
ind 7	7	3	3
ind 8	-6	-4	-5
ind 9	-8	-4	-6
ind 10	-1	-3	0
ind 11	0	-2	-2
ind 12	5	-1	1

Bu sentetik veri setinde sadece 12 katılımcı vardır. Araştırmacı bu veri setinden oluşturacağı 2 farklı matris ile, β tahmini olan $\hat{\beta}$ matrisini hesaplayabilir.

$$Y = \begin{bmatrix} 8 \\ 4 \\ 6 \\ 6 \\ 5 \\ 9 \\ 7 \\ -6 \\ -8 \\ -1 \\ 0 \\ 5 \end{bmatrix}, X = \begin{bmatrix} 1 & 0 & 3 \\ 1 & -2 & 1 \\ 1 & 6 & 3 \\ 1 & -3 & 0 \\ 1 & 5 & 0 \\ 1 & 4 & 2 \\ 1 & 3 & 3 \\ 1 & -4 & -5 \\ 1 & -4 & -6 \\ 1 & -3 & 0 \\ 1 & -2 & -2 \\ 1 & -1 & 1 \end{bmatrix}$$

OLS yöntemi ile $\hat{\beta}$ kolayca hesaplanabilir

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{11.1}$$

Bu hesaplamaları R ile yapalım;

```
Y=matrix(c(8,4,6,6,5,9,7,-6,-8,-1,0,5),ncol=1)
X=matrix(cbind(rep(1,12),
               c(0,-2,6,-2,5,4,3,-4,-4,-3,-2,-1),
               c(3,1,3,0,0,2,3,-5,-6,0,-2,1)),ncol=3)

solve(t(X)%*%X)%*%t(X)%*%Y
##       [,1]
## [1,] 2.917
## [2,] 0.199
## [3,] 1.552
```

Regresyon eşitliği;

$$\hat{Y}_i = 2.9167 + 0.1989X_{i1} + 1.5519X_{i2}$$

$\hat{Y}_i$ *i.* sıradaki birey için tahmin edilen değerdir. Eşitlik (11.1) hata kareleri toplamını minimize etmek üzere türetilmiştir: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Y'Y - \beta'X'X\beta$. Bu yöntem en iyi doğrusal yansız kestiricilerdir (Best Linear Unbiased Estimates).

Örnekte yer alan her iki bağımsız değişkeninde aritmetik ortalaması sıfırdır. X_1 ve X_2 sıfır iken, $(\hat{Y}_i)$ 2.92 olarak hesaplanır. Bir diğer ifade ile, her iki bağımsız değişken için ortalamada bulunan bireylerin bağımlı değişken ölçümü 2.92 olarak tahmin edilmiştir. X_2 değişkeni kontrol edildiğinde, X_1 değişkeninde gerçekleşecek 1 birim artışın bağımlı değişken için 0.2 birim artışa yol açacağı tahmin edilmiştir. Benzer şekilde, X_1 değişkeni kontrol edildiğinde, X_2 değişkeninde gerçekleşecek 1 birim artışın bağımlı değişken için 1.55 birim artışa yol açacağı tahmin edilmiştir. Çoklu regresyon modellerinde "kontrol edildiğinde (ceteris paribus)" ifadesi gereklidir. Hesaplanan katsayılar, .20 ve 1.55, araştırmacıyı değişkenler arasındaki ilişki hakkında bilgilendirir. Tabi bu noktada araştırmacının "1 birim artış" ifadesinin tam olarak ne anlama geldiğini bilmesi gerekir.

11.1.1 A) "Esasen bütün modeller yanlıştır, fakat bir kısmı işe yarardır."

Bu çıkarım Box and Draper (1987) tarafından yapılmıştır. Araştırmacı, araştırma sorusu doğrultusunda inşaa edeceği modelde yer alan değişkenleri nasıl seçtiğine yönelik ikna edici argümanlar sunmalıdır. Eğer önemli bir değişken modelin dışında bırakıldı ise kestirilen regresyon katsayıları muhtemelen geçersizdir.

Aşağıdaki durumu düşünelim

```
#sentetik veri setinde X2 göz ardı edilsin
X2omitted=matrix(cbind(rep(1,12),c(0,-2,6,-2,5,4,3,-4,-4,-3,-2,-1)),ncol=2)
solve(t(X2omitted)%*%X2omitted)%*%t(X2omitted)%*%Y
##      [,1]
## [1,] 2.92
## [2,] 1.09
```

Sentetik veri setinde X_1 ve X_2 arasındaki korelasyon 0.68, Y ve X_2 arasındakş korelasyon ise 0.93'tür.Eğer araştırmacı X_2 değişkenine modelde yer vermezse X_1 için hesaplanan katsayı 1.09 olur. =.20 ile kıyaslandığında bu büyük bir değişikliktir. Bir diğer ifade ile, modelde yer alan bağımsız değişkenler ile ve bağımlı değişken ile ilişkili olduğu halde modelde yer verilmeyen bir değişken var ise hesaplanan regresyon katsayıları yanıltıcıdır. [2]

Dışarıda kalan değişken probleminin yanında bir regresyon modelinin geçerliği örneklem seçimine ve örnekleme çerçevesinin (sampling frame) model ile yansıtabilmesine de bağlıdır. Örneğin örnekleme çerçevesinde ağrılıklandırma kullanıldı ise çözümleme esnasında bu ağırlıklar göz ardı edilmemelidir.

11.1.2 B) Bağımlı değişken ve bağımsız değişkenler arasındaki ilişkinin kuvveti

Bağımlı değişkene ait kareler toplamı (total sum of squares) iki ana parçaya ayrıştıtılabilir, *modele ait kareler toplamı* ve *hataya ait kareler toplamı*. Modele ait kareler toplamının değişkene ait kareler toplamına oranı R^2 ile gösterilir ve çoklu belirlilik katsayısı olarak adlandırılır. R^2 bağımlı değişken ve bağımsız değişkenler arasındaki ilişkinin kuvvetini ölçer.

```
# KT total
n=length(Y)
TotalSS=t(Y)%*%Y-(n*mean(Y)^2)

# KT Model
betahat=solve(t(X)%*%X)%*%t(X)%*%Y
ModelSS=t(betahat)%*%t(X)%*%Y-(n*mean(Y)^2)

ModelSS/TotalSS
##       [,1]
## [1,] 0.879
```

[2] Bu durum gerçek deneysel çalışmaların neden önemli olduğu hakkında ipucu teşkil eder.

Fakat R^2 evren parametresi için yanlı bir kestiricidir.Daha yansız bir kestirici ise düzeltilmiş belirlilik katsayısıdır, R^2_{Adj};

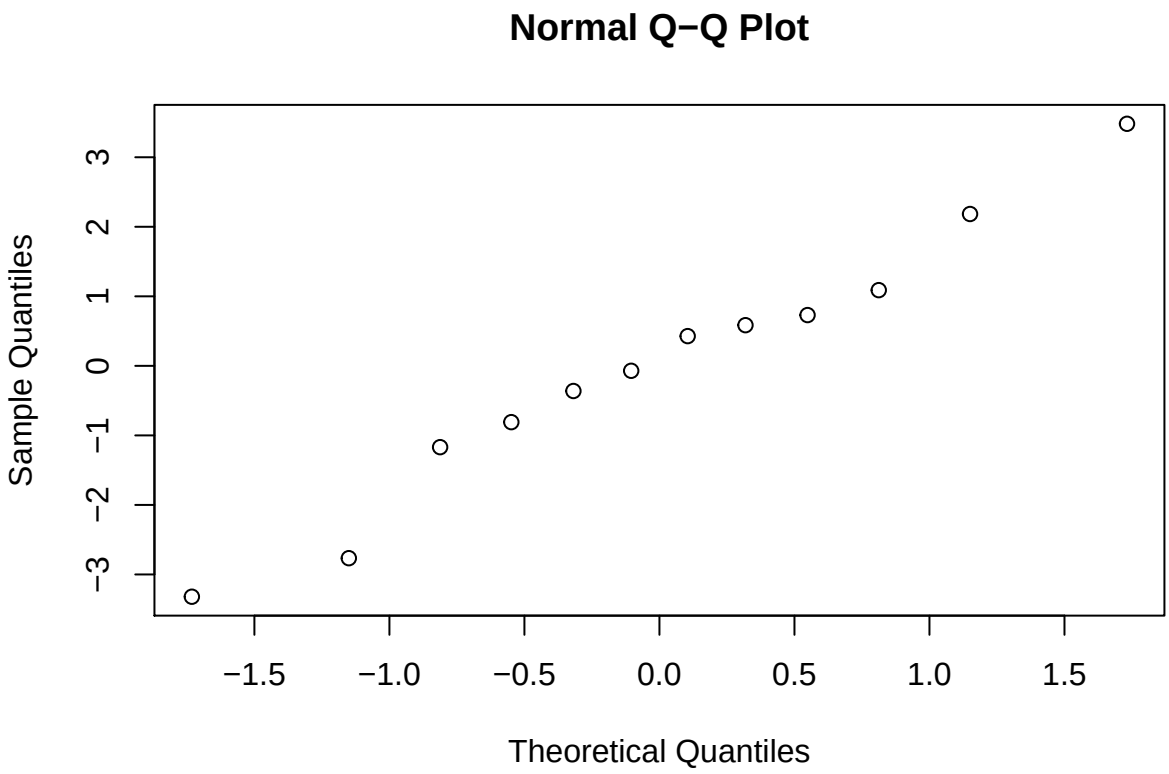
```
Rsquared=ModelSS/TotalSS
#sörneklem
n=12

#bağımsız değişken sayısı
p=2

# sabit (intercept) varsa 1, yoksa 0

int_inc=1


AdjustedRsquared=1-(1-Rsquared)*((n-int_inc)/(n-int_inc-p))
AdjustedRsquared
##       [,1]
## [1,] 0.852
```

R^2 ve R^2_{Adj} kullanışlı katsayılardır ve ne kadar varyansın model ile açıklandığını belirtir. Örneğimiz için $R^2 = .879$ ve $R^2_{Adj} = .852$ birbirine yakın değerlerdir.Eğer $R^2 = .25$ olsa idi R^2_{Adj} 0.08 olurdu. R^2 1 ise model varyansın tamamını açıklamıştır (%100). R^2 ve R^2_{Adj} aynı bağımlı değişkeni açıklayan farklı bağımsız değişken gruplarını karşılaştırmak için de kullanılır. R^2'in yorumu korelasyon katsayısı yorumunda olduğu gibi konu alanına göre değişir. 0.7 R^2 değerinin çok güçlü veya çok zayıf adledilebileceği durumlar olabilir.

11.1.3 C) Artıklar ve etkili gözlemler

Artıklar modelin uyumsuzluğu hakkında bilgi verebilir. Artıkların incelenmesi ile yordayıcı ve yordanan değişklenlerin arasındaki ilişkinin doğrusallığı görsel olarak kontrol edilebilir. Artıkların dağılımsal özelliklerini incelemek örneklemden evrene yorum yapmak için gerekli olabilir. Örneğin istatistiksel anlamlılık testlerinde ve güven aralığı hesaplamaları normal dağılım varsayımına dayandırılmış ise artıkların standart normal dağılım olasılığı grafiğinde (QQ grafiği) düz bir çizgiyi takip etmesi gerekir.

```
#tahmin edilen değerler
Yhat=X%*%betahat
residuals=Y-Yhat
residuals
##          [,1]
##  [1,]  0.4276
##  [2,] -0.0708
##  [3,] -2.7658
##  [4,]  3.4811
##  [5,]  1.0888
##  [6,]  2.1839
##  [7,] -1.1691
##  [8,] -0.3615
##  [9,] -0.8096
## [10,] -3.3199
## [11,]  0.5850
## [12,]  0.7303
qqnorm(residuals)
```



Genellikle artıklar üç farklı yaklaşımdan biri ile incelenir;

- Ham artıklar, $Y_i - \hat{Y}_i$. Ham artıklar Y ile aynı skaladadır.

- Standardize artıklar:Stanardize artıklar ham artıkların kkendi standart sapmlarına böünmesi ile hesaplanır. z puanı skalasındadırlar (ortalama=0, standart sapma=1). Normal dağılım varsayımı yapıldığında, aykırı değerlerin gözden geçirilmek üzere seçilmesi aşamasında mutlak standardize artıklar için kriter olarak 2 kullanılabilir. Bir diğer ifade ile, mutlak ğeri 2 den büyük olan standardize artıklar aykırı değer tespitinde incelenmelidir. Bununla birlikte bir z dağılımında değerlerin %5'inin +/-2 kriterinin dışında yer alır.

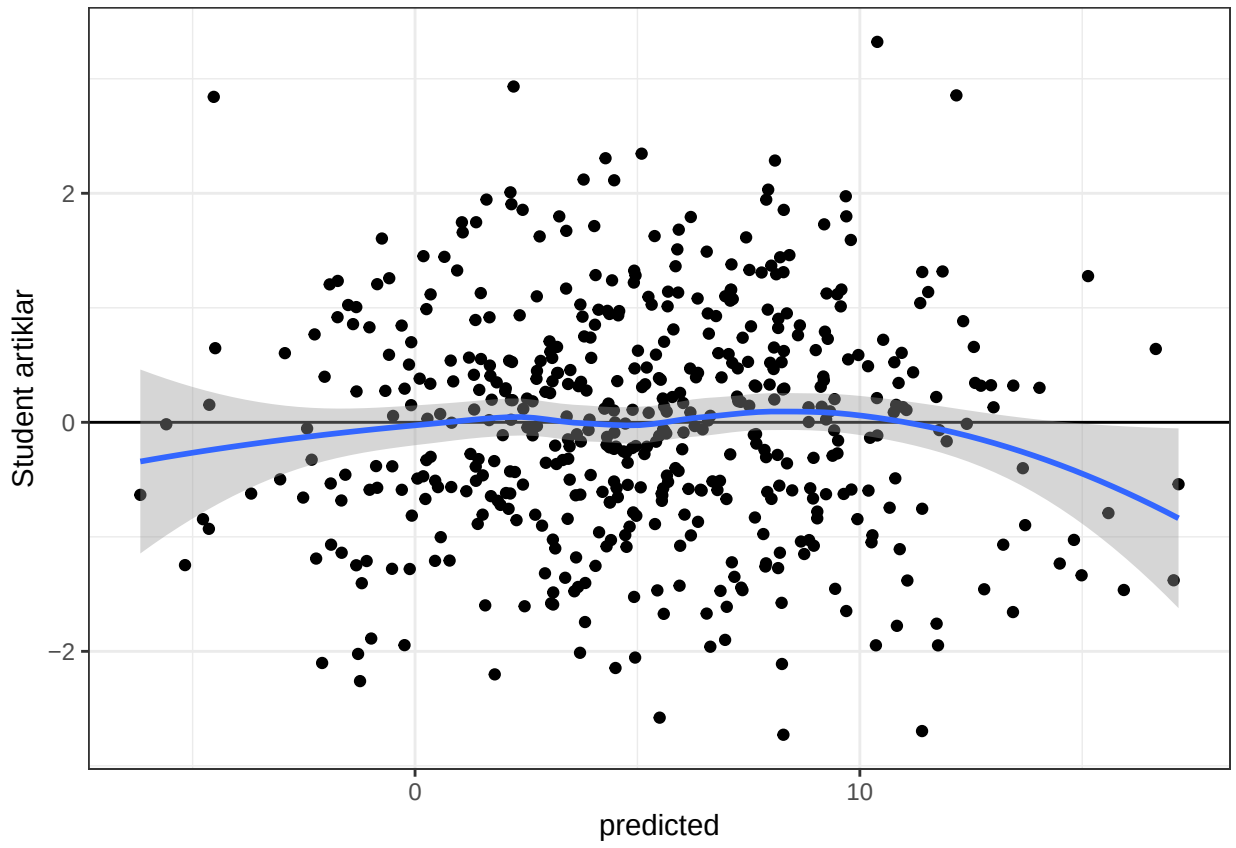
- Student artıklar: Ham artıkların tahmin edilen standart hataya bölünmesi ile hesaplanır.

Artıklar incelenirken bu üç yaklaşımda genellikle aynı sonucu verir. Rawlings et al. (1998) aykırı değerlerin tespitinde student artıkların ve $t_{n-p'-1}$ serbestlik derecesi ile bir t dağılımının kullanımını tavsiye eder. Buradaki p' modelde yer alan katsayıları temsil eder (örneğimiz için sabit+iki yordayıcı =3)

Artık değerleri tahmin edilen değerler ile karşılaştıran saçılım grafikleri doğrusallık hakkında fikir verebilir. Aşağıda 500 katılımcı ve 2 yordayıcı değişken için simüle edilmiş veri üzerine kurulmuş bir regresyon modelinden elde edilen artık değer-tahmin edilen değer saçılım grafiği verilmiştir. Doğrusallık ihlal edilmediğinden grafik üzerinde bir örüntü olmamalıdır.

```
#veri oluştur
library(mvtnorm)
sigma <- matrix(c(4,2,2,3), ncol=2)
xx <- rmvnorm(n=500, mean=c(0,0), sigma=sigma)
yy=5+xx[,1]*2+xx[,2]*-3+rnorm(500,0,1.5)
model=lm(yy~xx[,1]+xx[,2])
errors=rstudent(model)
predicted=predict(model)
```

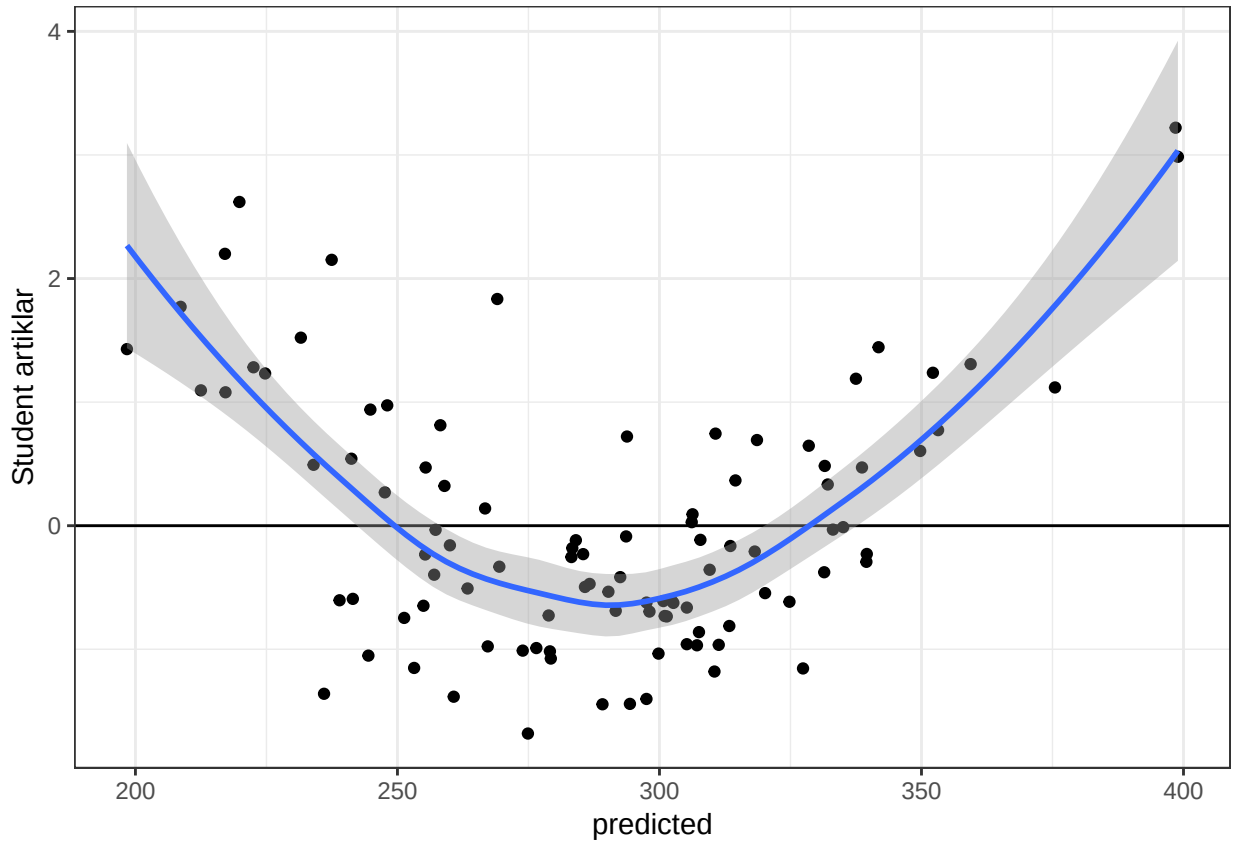
```
library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ ylab("Student artıklar")+
  theme_bw()+stat_smooth()
```



Grafikte yer alan mavi çizginin 0 çizgisine benzer olması doğrusallık ihlalinin olmadığını gösterir.

Gösterim amaçlı simule edilen bir diğer veri setinde Y ve X_2 arasındaki ilişki doğrusal değildir quadratiktir.

```
#veri oluştur
library(mvtnorm)
sigma <- matrix(c(4,2,2,3), ncol=2)
xx <- rmvnorm(n=100, mean=c(10,10), sigma=sigma)
yy=150+(xx[,1]*4)+(xx[,2]*-3)+(xx[,2]^2*1.2)+rnorm(100,0,3)
model=lm(yy~xx[,1]+xx[,2])
errors=rstudent(model)
predicted=predict(model)


library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ylab("Student artıklar")+
  theme_bw()+stat_smooth()
```



Bu grafik ilişkinin doğrusal olmadığını gösterse dahi sorunun kaynağını bulmada yardımcı değildir. Araştırmacı artık değerleri yordayıcılar ile karşılaştırarak sorunun kaynağını bulabilir. Bu konuya tekrar değinilecektir.

Artık değer grafikleri herhangi bir sorun belirtmese dahi sıradan olmayan artık değerler dikkatlice incelenmelidir. Bir artık değerin sıradışı (örneğin ortalamadan 3,4 veya 5 standart sapma farklı oluşu) olup olmadığına araştırmacı karar verir. Fakat bir gözlemin veri setinden çıkarılması iaraştırmacı tarafından detaylı olarak rapor edilmeli ve gerekçelendirilimelidir.

```
#veri oluştur
set.seed(04022017)
library(mvtnorm)
sigma <- matrix(c(4,2,2,3), ncol=2)
xx <- rmvnorm(n=100, mean=c(10,10), sigma=sigma)
yy=(xx[,1]*4)+(xx[,2]*-3)+rnorm(100,0,3)
tempdata=data.frame(yy,xx,id=1:100)
model=lm(yy~X1+X2,data=tempdata)
tempdata$SUTresiduals=rstudent(model)

# kaç tane artık değer kritik değerin üstünde
# alfa=.05
sum(abs(tempdata$SUTresiduals)>qt(c(.975), df=100-3-1))
## [1] 8

#hangi gözlemler?
tempdata[which(abs(tempdata$SUTresiduals)>qt(c(.975), df=100-3-1)),]
##       yy    X1    X2 id SUTresiduals
```

```
## 13 21.39 11.49 10.29 13         2.02
## 32  8.85 11.96 10.65 32        -2.20
## 43 15.80 11.14  7.56 43        -1.99
## 50  9.21  8.00 10.21 50         2.53
## 51 19.96 10.11  8.97 51         2.02
## 68 25.33 10.96  8.33 68         2.04
## 84  2.03  7.94  7.84 84        -2.03
## 91  5.51 10.74 10.25 91        -2.10
```

Tip I hata oranı 0.05 ile $t_{.975,96}$ kritik değerini kullanırsak yaklaşık olarak $n * .05$ gözleme ait mutlak student ayrık değerin kritik değerden büyük olması beklenir. Son örneğimizde 100 katılımcı olduğu için bu sayı 100*0.05=5'tir, tespit edilen potansiyel aykırı değer sayısı ise 8'dir. Fakat aykırı olma ihtimali olan gözlemler incelendiğinde bir anormallik görülmemiştir. Burada kullanılan $t_{.975,96}$ ihtiyatli bir kritik değerdir,$t_{.99,96}$ değeri de kullanılabilir. Bu yöntemin amacı potansiyel aykırı değerleri tespit edip incelemektir. Veri setine aşina olan araştırmacı hangi gözlemlerin sıradışı olduğunu söyleyebilir.

Araştırmacı sıradışılık farkeder ve gözlemleri veri setinden çıkarmaya karar verirse bunu birer bire yapmalıdır. Belirlenen en sıradışı gözlem çıkarılıp analizler tekrarlanmalıdır. Eğer gözlem veri setinden çıkarılacaksa sebepleri detaylı bir şekilde açıklanmalıdır. Bununla beraber, aykırı değerlere dirençli yöntemlerde kullanılabilir. Güçlü bir gerekçesi olmadığı sürece gözlemlerin veri setinden çıkarılması doğru değildir.

R programlama dili ile etkili gözlemleri tespit etmekte oldukça kolaydır.Etkili gözlem, veri setinden çıkarıldığında sonuçları değiştirebilecek gözlem olarak tanımlanabilir . *influence.measures* fonksiyonu 5 farklı ölçüm hesaplar;

```
summary(influence.measures(model))
## Potentially influential observations of
##   lm(formula = yy ~ X1 + X2, data = tempdata) :
##
##     dfb.1_ dfb.X1 dfb.X2 dffit cov.r   cook.d hat
## 12   0.08  -0.02  -0.08  -0.10  1.12_*  0.00   0.08
## 33   0.09  -0.03  -0.07  -0.11  1.11_*  0.00   0.07
## 41  -0.01  -0.03   0.03  -0.04  1.10_*  0.00   0.06
## 42   0.05  -0.12   0.07   0.13  1.11_*  0.01   0.07
## 50   0.20  -0.40   0.21   0.47  0.88_*  0.07   0.03
## 64  -0.03   0.03   0.00   0.04  1.10_*  0.00   0.06
## 100  0.01   0.13  -0.15  -0.18  1.10_*  0.01   0.07
```

Bu örnekte 12,33,41,42,50,64 ve 100 numaralı gözlemlerin etkili olma potansiyeli vardır. Tabloda görüldüğü gibi bu gözlemlerin hangi kritere göre seçildiği (*) işareti ile gösterilmiştir. Örneğimizde 7 gözlem de *kovaryans oranı* kriterine göre etkin bulunmuştur. Bu katsayı, gözlemlerin, regresyon katsayılarına ait örneklem varyansına etkisini ölçmeye çalışır. $1 + (3p'/n)$ ve $1 - (3p'/n)$ sınırları dışında kalan *kovaryans oranı* katsayısına sahip gözlemler influence.measures fonksiyonu tarafından işaretlenir. Örneğimizde n=100 ve p'=3 olduğundan kritik değerler 1.09 and .91 olarak hesaplanır.

Dfb (DFBETAS) değeri, gözlem çıkarıldığında regresyon katsayılarının ne kadar değişeceği hakkında bilgi vermeye çalışır. Gözlem çıkarıldıktan sonra hesaplanan yeni katsayı ile eski katsayı arasındaki farki yeni katsayının standart hatasına böler. Yani bir t istatistiğidir. Hesaplanan değer $2/\sqrt(n)$ kritik değerinden büyük ise influence.measures fonksiyonu gözlemi işaretler. Örneğimizde kritik değer $2/\sqrt(100) = .2$

dffit, gözlem veri setinden çıkarıldığında o gözlem için yeni tahminin ne ölçüde değişeceği hakkında bilgi vermeye çalışır. $2 * \sqrt{\frac{p'}{n}}$ kritik değerinden büyük olan mutlak dffit değerleri influence.measures fonksiyonu tarafından işaretlenir.

Cook uzaklğı (cook.d) bir gözlemin bütün regresyon katsayıları üzerindeki etkisini aynı anda ölçmeye çalışır. $F_{.5,p',n-p'}$ kritik değerinden büyük olan değerler influence.measures fonksiyonu tarafından işaretlenir. Cook

uzaklığı aynı zamanda belli bir gözlemin veri setinden çıkarılması durumunda geri kalan tahmini değerlerin ($\hat{Y}_i$) ne kadar etkilendiğini ölçmeye çalışır.

Leverage değeri (Hat Diag) bir gözlemin diğer gözlemlerden ne kadar uzak olduğunu ölçmeye çalışır. $2p/n$ kritik değerinden büyük olan leverage değerleri potansiyel etkili gözlemdir ve influence.measures fonksiyonu tarafından işaretlenir.

Potansiyel etkili gözlem olarak işaretlenen gözlemlerin değerlendirilmesi araştırmacının sorumluluğundadır. Daha önce belirtildiği gibi, etkili bir gözlemin veri setinden çıkarılıp çıkarılmaması önemli bir karardır. Veri setinden çıkarılan bir gözlem varsa sebepleri detaylı olarak açıklanmalıdır.

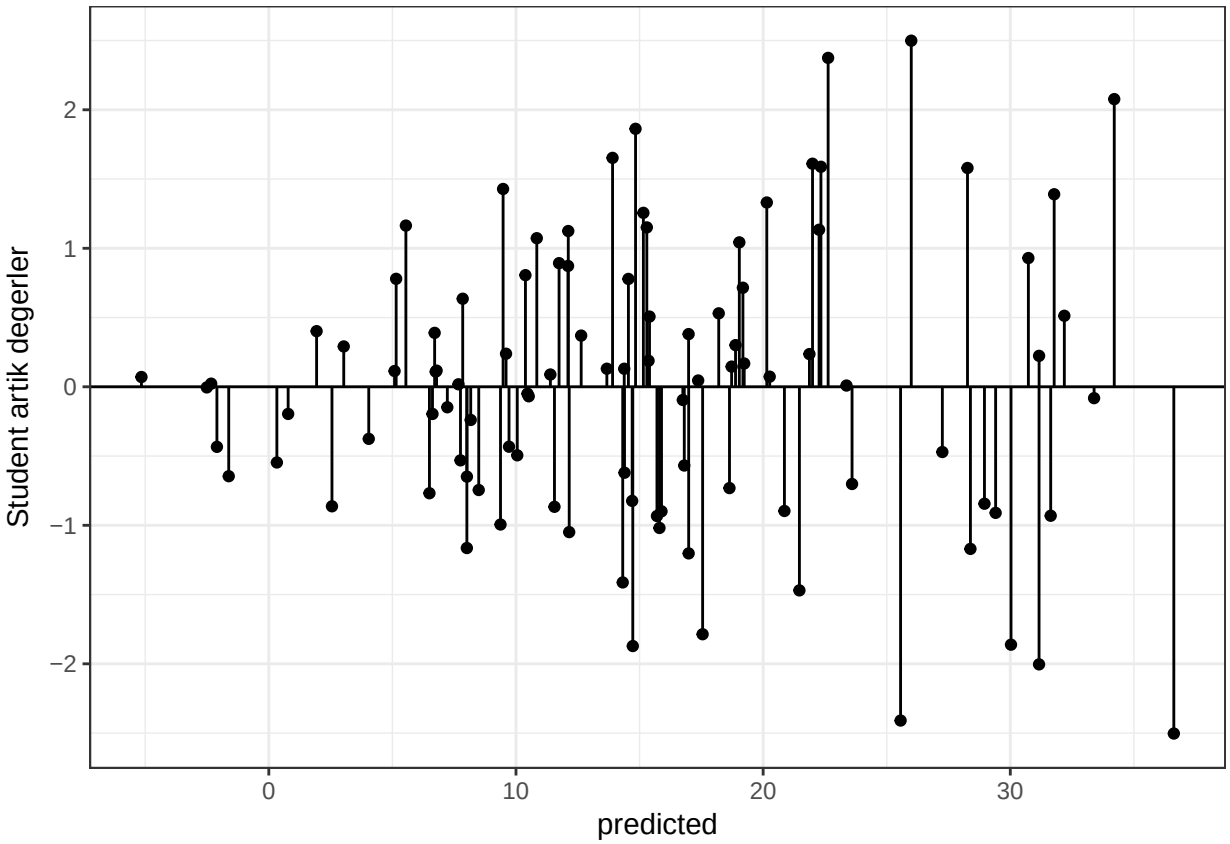
11.1.4 D) *Eş varyanslılık varsayımı*

Bağımlı değişkenin normal dağıldığı varsayımı ve OLS yöntemi ile regresyon katsayılarının β dağılımları hesaplanabilir. Bu sayede regresyon katsayıları için standart hatalar $\hat{\sigma}^2(XX)^{-1}$ ile kestirilebilir. Bu eşitlikteki $\hat{\sigma}^2$ hata terimlerinin varyansıdır.

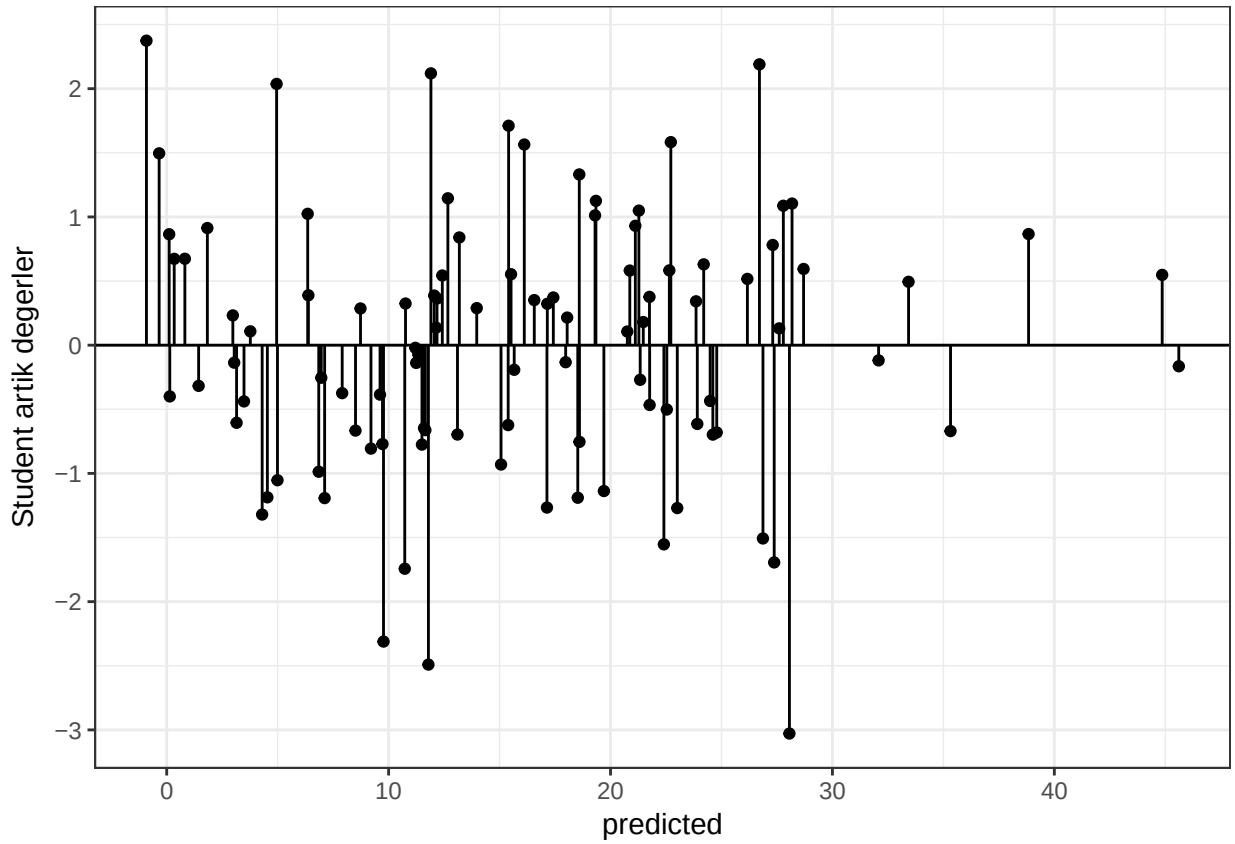
```
#artıklar
s2 <- (t(residuals) %*% residuals)/(nrow(Y)-nrow(betahat))
Var_betahat <- s2[1,1]*solve(t(X)%*%X)
```

$\sigma^2(X'X)^{-1}$ eşitliği homojen varyans varsayımı altında geçerlidir. Bu varsayım, bağımsız değişkenler kontrol edildiğinde bağımlı değişkenin eş varyanslılık göstermesi durumudur. Bir diğer deyişle, bağımlı değişken için yapılmış her bir gözlem aynı miktarda bilgi sağladığı varsayılır (Rawlings et al. (1998)). Bu varsayım ile birlikte regresyon katsayıları $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ toplamını minimize etmek üzere kestirilir. Bu eşitlikte her bir artık değer eşit bir ağırlığa sahiptir. Eğer eş varyanslılık varsayımı ihlal ediliyorsa artık değerlere farklı ağırlık vererek kestirici modifiye edilebilir veya başka bir dirençli kestirici kullanılabilir. Başka bir alternatif olarak bağımlı değişkenin transformasyonu kullanılabilir veya standart hataların tahmin yöntemi değiştirilebilir (bakınız Lumley and Zeileis (2015)). Aksi takdirde $\hat{\beta}$ için standart hatalar, eş varyanslılığın nasıl ihmal edildiğine evren parametresine kıyasla gerekenden büyük veya küçük hesaplanabilir. Eğer küçük hesaplanırsa Tip I hata oranı alfadan daha büyük olabilir. Eğer büyük hesaplanırsa istatistiksel güç yitirilir. Eş varyanslılık hipotezi, artık değerler ile tahmin edilen değerlerin kıyaslandığı bir grafik üzerinden incelenebilir. Eş varyanslılığın zedelendiği bir durum;

```
#veri oluştur
set.seed(03032017)
library(mvtnorm)
sigma <- matrix(c(1,.7,.7,1), ncol=2)
xx <- rmvnorm(n=100, mean=c(1,1), sigma=sigma)
#heteroscedasticity ekle
hts=function(v1,v2){2+.5*v1+.5*v2}
yy=5+xx[,1]*5+xx[,2]*5+rnorm(100,0,hts(xx[,1],xx[,2]))
model=lm(yy~xx[,1]+xx[,2])
#summary(model)
errors=rstudent(model)
predicted=predict(model)

#student artıklar ve tahmin edilen Y
library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ylab("Student artık değerler")+
  geom_segment(mapping=aes(xend = predicted, yend = 0)) +
  theme_bw()
```



Genel olarak $\hat{Y}$ değerleri küçük ise hata varyansı da küçük görülmektedir. Eş varyanslılığın sağlandığı bir durum;



11.1.5 E) Hipotez testi

$H_0 : \beta_1 = ... = \beta_p = 0$ boş hipotezi F dağılımı takip bir istatistik ile test edilebilir. Bu boş hipotez bütün regresyon katsayılarının sıfıra eşit olduğunu öne sürer. Alternatif hipotez ise en az bir regresyon katsayısının sıfırdan farklı olduğunu ileri sürer. $MS_{regression}/MS_{residual}$ istatistiği p ve $n-p'$ serbestlik derecesine sahip bir F dağılımı takip eder. p bağımsız değişken sayısını p' ise regresyon katsayıları sayısını temsil eder (sabit yoksa $p = p'$). Alfa=0.05 ve sentetik data için;

```
# Model KT ve Total KT daha önceden hesaplanmıştır.
dfREG=2  #(p=2, bağımsız değişkenler X1 and X2)
dfRES=9  #(n-p', 12-3)
MSreg=ModelSS/dfREG
MSres=(TotalSS-ModelSS)/dfRES

MSreg/MSres
##      [,1]
## [1,] 32.8

#kritik F
qf(.95,dfREG,dfRES)
## [1] 4.26
1-pf(MSreg/MSres,dfREG,dfRES)
##           [,1]
## [1,] 7.39e-05
```

t-testi ise $H_0 : \beta_X = \beta_{hyp}$ boş hipotezi ve $H_1 : \beta_X \neq \beta_{hyp}$ alternatif hipotezini test etmek için kullanılabilir.

Genellikle $\beta_{hyp} = 0$ kullanılır.

$(b_X - \beta_{hyp})/SE(b_X)$ istatistiği N-p' serbest dağılımına sahip bir t dağılımı takip eder.

```
# X2 regresyon katsayısı 0'dan farklı mı
Bhyp=0  #boş hipotez değeri

# betahat daha önceden hesaplanmıştır
# X2 için hesaplanan katsayı
bx2=betahat[3]

# Var_betahat daha önce hesaplanmıştır
# X2 regresyon katsayısına ait standart hata
se_bx2=sqrt(Var_betahat[3,3])

#t istatistiği
(bx2-Bhyp)/se_bx2
## [1] 5.33

# t kritik değeri
qt(.975,9)
## [1] 2.26

#p değeri
2*(pt(-abs((bx2-Bhyp)/se_bx2),9))
## [1] 0.000478
```

11.1.6 F) Değişken seçimi

Uzak bir bakışaçısından, çoklu regresyonun kullanıldığı iki senaryo vardır.

İlk senaryo: Sosyal bilimci alan yazını dikkatli bir şekilde inceler, araştırma sorusu ile ilgili olan bağımsız değişkenleri belirler, gereken örneklem büyüklüğüne karar verir, veriyi toplar,bütün bağımsız değişkenleri içeren bir model kurar ve sonuçları raporlar

İkinci senaryo: Sosyal bilimcinin oldukça büyük bir veri setine erişimi vardır ve hangi bağımsız değişkenlerin modelde yer alacağına dair önceden bir kararı yoktur. Bu durum, (a) araştırmacının yeni bir teori üzerinde çalıştığı ve bir çok değişkeni ölçtüğü durumlarda durumlarda veya (b) daha önceden toplanmış bir veri seti üzerinde (secondary data) çalıştığı durumlarda görülebilir. Her iki durumda da araştırmacı en iyi tahmin kabiliyeti gösteren değişkenleri seçmek isteyebilir. Bu seçim işlemi için farklı yaklaşımlar mevcuttur, örneğin adım adım seçim (stepwise), eleme ile seçim (backward) veya ekleme ile seçim (forward). Fakat bizim tecrübemize göre tamamen aynı veri setine uygulandığında bile bu yöntemler farklı sonuçlar vermektedir.

R ile oldukça kolay tamamlanabilecek bir diğer değişken seçme yöntemi mümkün olan bütün regresyonları koşmaktır. Tanıtım amacı ile yazılan R kodunu inceleyiniz;

```
#veri oluştur
set.seed(02082017)
library(mvtnorm)
sigma=matrix(c(5.899559,4.277045,3.906341,
               4.277045,5.817412,3.654419,
               3.906341,3.654419,5.642258),ncol=3)
xx <- rmvnorm(n=200, mean=c(0,0,0), sigma=sigma)
yy=5+xx[,1]+xx[,2]*1.5+xx[,3]*2+rnorm(200,0,3)
simdata=data.frame(yy,xx,id=1:200)
```

```
library(leaps)
formula <- formula(paste("yy ~ ",
     paste(names(simdata[2:4]), collapse=" + ")))
allpossreg <- regsubsets(formula,nbest=3,data=simdata)
aprout <- summary(allpossreg)

# str(aprout) u inceleyiniz
# bu fonksiyon R2 ve Düzeltilmiş R2 den başka kriterler de hesaplar

APRtable=with(aprout,round(cbind(which,rsq,adjr2),3))
APRtable=data.frame(APRtable,check.rows = F,row.names = NULL)
APRtable$ppri=rowSums(APRtable[,1:4])
kable(APRtable)
```

X.Intercept.	X1	X2	X3	rsq	adjr2	ppri
1	0	1	0	0.753	0.751	2
1	0	0	1	0.696	0.695	2
1	1	0	0	0.630	0.629	2
1	0	1	1	0.871	0.870	3
1	1	0	1	0.811	0.809	3
1	1	1	0	0.808	0.806	3
1	1	1	1	0.890	0.888	4

Bu tabloya göre sabitin ve sadece X_2 değişkeninin olduğu model $R^2 = .753$ sonucunu vermektedir. Bütün değişkenler eklendiğinde $R^2 = .890$ olur, fakat X_1 in modelde yer almaması R^2 değerini sadece .019 düşürür. Grafiği inceleyiniz

```
require(ggplot2)
ggplot(APRtable, aes(x=ppri-1, y=rsq)) +
    geom_point(shape=1,size=3)+
    scale_y_continuous(breaks = seq(0.5, 1, by = 0.05)) +
    scale_x_continuous(breaks = seq(0, 3, by = 1))+
    theme_bw()+labs(x = "R-squared")+
    theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))

ggplot(APRtable, aes(x=ppri-1, y=adjr2)) +
    geom_point(shape=1,size=3)+
    scale_y_continuous(breaks = seq(0.5, 1, by = 0.05)) +
    scale_x_continuous(breaks = seq(0, 3, by = 1))+
      theme_bw()+labs(x = "Adjusted R-squared")+
    theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

11.1.7 G) Güçlü doğrusal bağlantı sorunu

Bağımsız değişkenlerin birbiri ile çok güçlü olarak ilişkili olması tahmin sürecinde istenmeyen sonuçlara yol açabilir. Bu durum doğrusal bağlantı (collinearity) sorunu olarak bilinir. Standart hatalar doğrusal bağlantı arttıkça artar çünkü bu sorun her bir bağımsız değişkenin tahmin sürecine olacak olumlu katkısını saklar.

Açıklama amacı ile iki bağımsız değişkeninin olduğunu düşünelim, bu değişkenler arasındaki korelasyonun yüksek olduğu durumda iki tür problem oluşabilir, (a) regresyon katsayıları istikrarsız olabilir, aynı evrenden çekilen örneklemler ile çok farklı katsayılar elde edilebilir, (b) istatistiksel olarak anlamlı bir R^2 bulunsa dahi katsayılar istatistiksel olarak sıfırdan farksız olabilir.

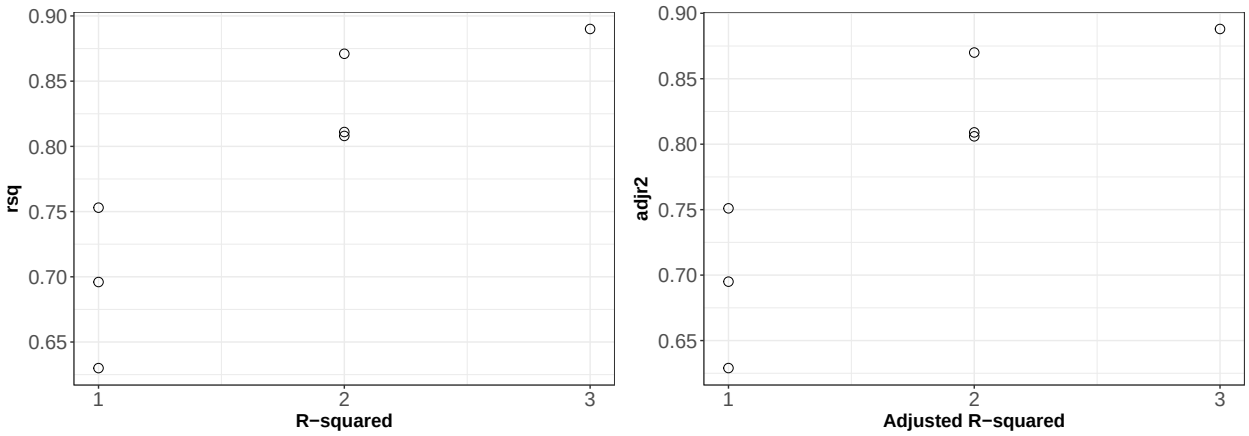


Figure 11.1: Mümkün olan bütün regresyonlar

Varyans şişkinliği faktörü (Variance inflation factor,VIF) çoklu doğrusallık tespitinde kullanılabilir,$VIF_x = \frac{1}{1-R_X^2}$. Bu eşitlikte R_X^2 X değişkeni çıkarıldığında hesaplanan R^2 değeridir. Büyük VIF değerleri potansiyel doğrusallık problemini işaret eder.Telaffuz edilen kritik VIF değerleri 4 ve 10 dur, fakat VIF değerleri dolaylı da olsa örneklem büyüklüğüne ve değişken varyansına göre değişebilir (Obrien (2007)). VIF değerleri büyük ise araştırmacı problemin kaynağını incelemedir. Araştırmacı şu seçeneklerden bir tanesini savunmaya çalışabilir, (a) yüksek korelasyona sahip iki değişkenden birini modelin dışında tutma, (b) yüksek korelasyon gösteren iki değişkeni birleştirme. Karar dikkatli bir şekilde verilmeli ve yeni sonuçlar ile eki sonuçlar kıyaslanmalıdır.

```
#korelasyonları kontrol et.
cor(simdata[,2:4])
##      X1    X2    X3
## X1 1.00 0.730 0.640
## X2 0.73 1.000 0.666
## X3 0.64 0.666 1.000

#en yüksek korelasyon .73,
#çoklu doğrusallık problemi beklenmez

library(car)
vif(lm(yy~X1+X2+X3,data=simdata))
##   X1   X2   X3
## 2.36 2.50 1.98
# VIF değerleri düşük
```

11.1.8 H) Doğrusal olmayan regresyon

Eğer bağımlı değişken bağımsız değişkenlerden biri ile lineer olmayan bir ilişkiye sahip ise, bu durumun görmezden gelinmesi önemli bir değişkenin dışarda bırakılması problemi ile aynıdır. Artık değerlerin incelenmesi lineer olmayan ilişkilerin tespitinde işe yarardır. Kullanılan yöntemlerden biri, bağımsız değişkenin üs kuvvetleri ile oluşturulacak yeni bir değişkenin modele eklenmesidir. Örneğin artık grafiği X_k ve artıklar arasında karesel (quadratic) bir ilişki varsa X_k^2 modele alınarak artık model uyumu tekrar incelenebilir. Bir diğer alternatif bağımsız değişkenin transform edilmesidir.

11.1.9 I) Korelasyon gösteren veya bağımsız olmayan hata terimleri

Hatalar birbiri ile korelasyonlu olmamlıdır, daha kapsayıcı bir ifade ile, hatalar bağımsız olmalıdır. Bağlı olma durumu var ise bu durum göz ardı edilirse regresyon sonuçları geçersizdir. Fakat bu konu bu tanıtım materyalinin kapsamı dışında kalır. Sosyal bilimlerde tekrarlı ölçümler kullanıldığında (boylamsal çalışma) korelasyon gösteren hata terimleri oluşabilir, bu durumda araştırmacılar Örtük gelişim modelleri (latent growth models) veya çok düzeyli modeller (multilevel models) kullanabilir. Eğer hataların bağlı olma durumu bireylerin aynı kümelerden gelişi ise (nested or clustered data) yine çok düzeyli modeller kullanılabilir.

11.1.10 J) Bağımsız değişken üzerine işlemler (Centering and Scaling)

Annenin doğum esnasındaki yaşının çocuğun 10 yaşındaki IQ puanını tahmin etmeye çalışan bir çalışma düşünelim. Bu durumda regresyon sabiti anne yaşının 0 olduğu durumda çocuğun IQ puanını gösterir. Bu anlamlı olarak yorumlanabilir bir katsayı değildir. Eğer anne yaşı ortalama etrafında merkezileştirilirse, bir diğer ifade ile, her annenin yaşından örneklemdeki annelerin ortalama yaşı çıkarılırsa, regresyon sabiti yorumlanabilir. Bu durumda yaş değişkeninde sıfır, yaş ortalamasını temsil eder ve regresyon sabiti de ortalama yaştaki bir annenin çocuğunun 10 yaşındaki tahmini IQ puanını verir. Benzer bir örnek olarak, işe devamsızlık değişkeninin iş stresi değişkeni ile açıklanmaya çalışıldığını düşünelim. İş stresi puanları 10 ila 50 arasında değişiyorsa regresyon sabitinin anlamlı bir yorumu olmaz. Araştırmacı iş stresi puanları ortalama etrafında merkezileştirmeyi veya seçtiği bir değeri (örneğin 40) stres puanlarından çıkararak sabitin anlamlı bir şekilde yorumlanmasını sağlayabilir. Başka bir örnekte gelir değişkeninin sağlık değişkeni üzerine etkisi araştırılsın. Gelir verisinde bir birimın 1 dolar olduğunu ve bu değişkene ait regresyon katsayısının .001 olduğunu düşünelim. Araştırmacı bu durumda gelir değişkenini 1000'e bölerek birimi 1 dolardan 1000 dolara çıkarabilir. Bu durumda regresyon katsayıso .001 değil 1 olacaktır. Bu durum araştırmacının yorum yapmasını kolaylaştırabilir.

11.1.11 K) Standardize edilmiş katsayılar

Yukarıda bahsedilen lineer transformasyonun yanında z transformasyonu da yapılabilir. Her sürekli bağımsız değişken için, değerlerden ortalama çıkarılıp standart sapmaya bölünebilir. Değişkenin doğasına bağlı olmak kaydı ile, standardize edilmiş bir değişkenin yorumu daha anlamlı olabilir; Ham puanlar için yorum: Endişe puanlarındaki bir birim artışın akademik başarı değişkeninde 3 birim azalmaya yol açabileceği tahmin edilmiştir. z-puanları: Motivasyon değişkeninde gerçekleşecek bir standart sapma artışın, başarı değişkeninde 0.25 standart sapma artışla ilişkili olabileceği tahmin edilmiştir.

11.1.12 L) Etkileşimler (Interactions)

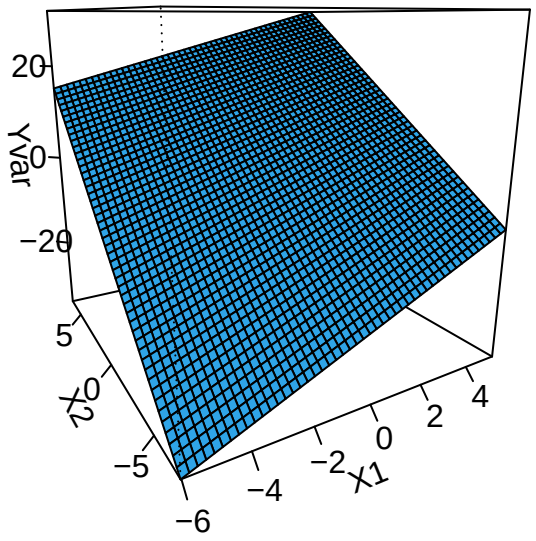
ANOVA bölümünde etkileşim konusu kısaca ele alınmıştı. Değişkenler arası etkileşimin modelde yer alması gerekirken göz ardı edilmesi önemli bir değişkenin dışarıda bırakılması problemidir. Ayrıca iki değişken arasındaki etkileşimin sonuçları etkilemesi durumunda asıl etkilerin yorumlanması yanıltıcı olabilir. Örneğin bir araştırmacının, öğrencilerinin merkezi bir sınavda matematik başarısını (Y) tahmin etmeye çalışsın. Ödevlerden alınan puanlar (X_1) ve öğretmen yapımı sınavlardan alınan puanların (X_2) başarı üzerinde etkili olup olmadığını incelesin. Kuracağı eşitlik $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ olursa araştırmacı Y ve X_1 arasındaki ilişkinin X_2'ye bağlı olmadığını varsayar. Eğer bu varsayım hatalı ise sonuçlar yanıltıcı olabilir. Etkileşimin varlığını kontrol etmek için

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i \tag{11.2}$$

modeli test edilebilir. Burada Y ve X_1 arasındaki ilişki $\beta_1 + \beta_2 X_{i2}$ ile gösterilir, bu Y ve X_1 arasındaki ilişkinin X_2'ye bağlı olduğunu gösterir. Benzer şekilde Y ve X_2 ilişkisi $\beta_2 + \beta_1 X_{i1}$ ile gösterilir. $\beta_1 + \beta_2 X_{i2}$ düşünüldüğünde β_1, Y ve X_1 arasındaki ilişkinin $X_2 = 0$ iken ne olduğunu belirtir. Eğer $X_2 = 0$ anlamlı değil

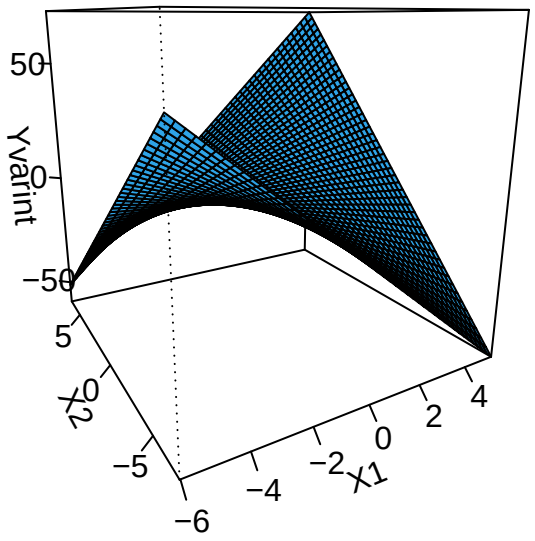
ise β_1 yorumu da anlamlı değildir. Eşitlik (11.2) etkileşimin $\beta_2 X_{i1} X_{i2}$ ile doğru bir şekilde dikkate alındığı varsayımını yapar. Yapar ayrıca bu durum X_2 konrol edilirken Y ve X_1 arasındaki ilişki doğrusal ise geçerlidir. Eşitlik (11.2) ile verilen modelin varsayımları incelenmelidir. R ile çizilebilecek grafikler etkileşimleri tespit etmede yardımcı olabilir. *visreg* (Breheny and Burchett (2016)) paketi ile çizilen grafiklere bakalım;

```
## simule edilmiş veriyi manipüle et
## Yvar: etkileşim yokken  bağımlı değişken
simdata$Yvar=3+simdata$X1*2+simdata$X2*3+rnorm(nrow(simdata),0,5)

library(visreg)

model=lm(Yvar~X1+X2+X1*X2,data=simdata)
visreg2d(model, "X1", "X2", plot.type="persp")
```



Düz yüzey etkileşimin olmadığını işaret eder.

```
## simule edilmiş veriyi manipüle et
## Yvar: etkileşim varken  bağımlı değişken
simdata$Yvarint=3+simdata$X1*1+simdata$X2*2+simdata$X1*simdata$X2*1.5+rnorm(nrow(simdata),0,5)

library(visreg)

model2=lm(Yvarint~X1+X2+X1*X2,data=simdata)
visreg2d(model2, "X1", "X2", plot.type="persp")
```



Yüzey artık düz değil, etkilesim mevcut.

11.1.13 M) Farklı kestiriciler (estimators)

Eklenecek

11.1.14 N) Dirençli yöntemler (Robust regression)

Eklenecek

11.1.15 O) Örneklem büyüklüğü ve istatistiksel güç

Eklenecek

11.1.16 P) Değişkenlerin güvenirliği

Eklenecek

11.1.17 Q) Dğişkenlerin türü

Eklenecek

11.1.18 R) Birden fazla bağımlı değişken

Eklenecek

11.1.19 S) Kayıp veri teknikleri

Eklenecek

Chapter 12

Kullanışlı R betikleri

```
# sayısal veriyi faktöre çevir

temdata[,2:9] <- lapply(temdata[,2:9], as.factor)

# Faktörü sayısal veriye çevir
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}
temdata[,2:5] <- lapply(temdata[,2:5], as.numeric.factor)

# Birden fazla sütun için frekans tablosu

dems=apply(temdata[,5:11], 2, function(x){table(x,temdata$grp)})
library (plyr)
mydems <- ldply (mydems, data.frame)

# Faktpre göre özetle
uncagg=aggregate(. ~ grp, data = temdata, FUN=mean, na.rm=TRUE)

uncaggfaster=temdata[, lapply(.SD, mean,na.rm=T), by = grp]

# maksimum değeri bul
which.max(x)


# R güncellemesi
if(!require(installr)) {
  install.packages("installr"); require(installr)} #load / install+load installr
updateR()


# Kukla değişken oluştur
head(temdata)
for(level in unique(temdata$zp)){
  temdata[paste("dummy", level, sep = "_")] <- ifelse(temdata$zp == level, 1, 0)
}


# noktalı virgül
```

```
x=rnorm(10000,5,10)
mean(x);var(x);sqrt(var(x))

# objeyi sil
y=rnorm(10)
rm(y)

# çalışma alanını boşalt
rm(list=ls())

# Belirlenen dışında sil
rm(list=setdiff(ls(),c("temdata", "temdata2")))


# tam sayı bölümü
7%/%2

# kalan
5%%2

# tanımla ve koş
(count=c(25,12,7,4,6,2,1,0,2))

# tıklama ile csv oku
data=read.csv(file.choose(),header=TRUE,)

#1 den fazla benzer CSV birleştir
filenames <- list.files()
temdata=do.call("rbind", lapply(filenames, read.csv, header = F))
write.table(temdata, file ="temdata.binded.csv" , sep = ",",col.names = F, row.names = F)


#birden çok grafik
layout(matrix(1:9, nc = 3))
sapply(names(temdata)[1:9], function(x) {
  qqnorm(temdata[[x]], main = x)
  qqline(temdata[[x]])
})




#birden fazla grafik için ekranı böl
par(mfrow=c(3,3))


#Çift loop
x=matrix(1:15,3,5)
for(i in seq_len(nrow(x)))
{
  for(j in seq_len(ncol(x)))
  {
    print(x[i,j])
```

```
  }
}



#While loop
count=0
while(count<10){
  print(count)
  count=count+1
}


#kayıp veri
convert -999s to NAs

read.csv("x.csv", na.strings="-999")
temdata[is.na(temdata)] <- 0


# NA lar -99

vector[which(vector== NA)]= (-99)
temdata[is.na(temdata)]= (-99)

# <NA> porblemi (but not NA)
temdata=read.csv("temdata.csv",stringsAsFactors=FALSE)


# grup ortalaması ekle

temdata2=merge(temdata, aggregate(X ~ grp, data = temdata, FUN=mean, na.rm=TRUE),
          by = "grp", suffixes = c("", ".mean"),all=T)


temdata2=merge(temdata, aggregate(cbind(X1 ,X2 ,X3 , X4) ~ grp, data = temdata, FUN=mean,

temdata2=merge(temdata,
          ddply(temdata, c("grp"), function(x) colMeans(x[c("X1" ,"X2","X3" , "X4")])),
          by = "grp", suffixes = c("", ".mean"),all=T)




#ifelse
y=c(1,2,3,4,5,5,5)
y2=ifelse(y==5,NA,y)
y2


temdata <- data.frame (ID=c(2,3,4,5), Hunger =c(415,452,550,318 ))

temdata$newcol<-ifelse(temdata[,2]>=300 & temdata[,2]<400,350,
```

```
                        ifelse(temdata[,2]>=400 &temdata[,2]<500,450,
                               ifelse(temdata[,2]>=500 & temdata[,2]<600,550,NA)))



#if
x=5
y=if(x>6){1}else{0}
y=if(x>6){1} else if(x==5) {99} else {0}


#veri setini B ye göre diz
temdata[order(temdata$B),]

temdata[rev(order(temdata$B)),]


#kombinasyon oluştur
m=c(54,38,51,62,18,31,58,74,35,34)
f=c(41,18,19,39,44,18,58,21,38)

mean(m)
mean(f)

combn(m,8,FUN=mean)
combn(f,8)

min(combn(m,8,FUN=mean))
max(combn(f,8,mean))


# contrasts değiştirme
options('contrasts')
options(contrasts=c('contr.sum','contr.poly'))
options(contrasts=c('contr.treatment','contr.poly'))


# bütün satırlar kayıp ise sil
temdata=temdata[apply(temdata,1,function(x)any(!is.na(x))),]


# grup frekansı ekle
temdata=ddply(temdata, "grp", transform, cellsize = count(grp)[2])


#yeni klasör aç
dir.create("testdir")

#veri setini böl
library(datasets)
head(airquality)
splitdata=split(airquality,airquality$Month)
splitdata
```

```r
str(splitdata)
splitdata[[2]]


x=list(a=1:5, b=rnorm(10))
x
lapply(x,mean)



x=1:4
lapply(x,runif)
lapply(x,runif,min=10, max=20)


x=list(a=matrix(1:4,2,2),b=matrix(1:6,3,2))

lapply(x,function(elt) elt[,1])


# sapply

x=list(a=1:5, b=rnorm(10),c=runif(10))
x
lapply(x,mean)
sapply(x,mean)



#apply

x=matrix(rnorm(200),20,10)
x
apply(x,2,mean)
apply(x,1,sum)

#tapply
x=c(1:10,rnorm(10),runif(10,3,5))
f=gl(3,10)
?gl
h=factor(rep(1:3,each=10))
tapply(x,f,mean)
tapply(x,h,mean)
tapply(x,h,mean,simplify=F)
tapply(x,h,range)




#kayıp veri yüzdeleri
propmiss <- function(temdata) lapply(temdata,function(x) data.frame(nmiss=sum(is.na(x)), n=length(x), pr
propmiss(temdata)
```

```
#büyük harf
temdata$childid=toupper(temdata$childid)



# sütünları aynı anda graikleştir.

plotpdf="C:/Users/Desktop/work/multiplePLOTS.pdf"
pdf(file=plotpdf)
for (i in 7:55){
  muis=round(mean(temdata[,i],na.rm=T),3)
  sdis=round(sd(temdata[,i],na.rm=T),3)
  meansc=c("mean",muis)
  hist(temdata[,i],freq=F,main=names(temdata)[i],xlab=meansc)
  #lines(density(temdata[,i],na.rm=T))
  curve(dnorm(x, mean=muis, sd=sdis), add=TRUE)
  lines(density(temdata[,i],na.rm=T, adjust=2), lty="dotted", col="darkgreen", lwd=2)
  abline(v=muis,col="blue")
  abline(v=muis+3*sdis,col="red")
  abline(v=muis-3*sdis,col="red")
}

dev.off()


# bir üst klasörden oku
dd=read.csv("../temdata.csv")
```

12.1 apaStyle paketi

```
require(pastecs)
require(apaStyle)
library(rJava)
#hata verirse
Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre1.8.0_111') # for 64-bit version

#veri tanımla

apa.descriptives(data = temdataet[,1:5], variables = names(temdataet[,1:5]), report = "", title = "test

example <- data.frame(c("Column 1", "Column 2", "Column 3"), c(3.45, 5.21, 2.64), c(1.23, 1.06, 1.12) )
apa.table(data = example, level1.header = c("Variable", "M", "SD"))


example <- data.frame( c("Column 1", "Column 2", "Column 3"),
                       c(3.45, 5.21, 2.64),
                       c(1.23, 1.06, 1.12),
                       c(8.22, 25.12, 30.27),
                       c("+", "**", "***") )
```

```
apa.table( data = example, level1.header = c("", "Descriptives", "Inferential"),
           level1.colspan = c(1, 2, 1),
           level2.header = c("Variable", "M", "SD", "t-value", "*") )$table
```

Chapter 13

Proje Başvurusu (Seçilen kısımlar)

13.0.1 Amaç / Gerekçe

Hiç bir ticari çıkar gözetmeksizin hazırlanacak bu çalışmaların, R gibi dünyaya kendini ispatlamış ve ücretsiz olan bir programın ülkemizde de yaygınlaşmasına biraz olsun katkı sağlayabileceğini umuyorum. Eğitim alanında çalışan ve çalışacak araştırmacıların R ile tanışmak istediklerinde kullanabileceği dokümanlar ortaya çıkacaktır. Doküman halka açık bir çevrimiçi depoda tutulacaktır. Dokümanların kabul görmesi durumunda, kullanıcılardan gelen istekleri ve önerileri hızlı bir şekilde dokümanlara dahil ederek, ihtiyaçlara cevap verme ihtimali yüksek olan bir kaynağın ortaya çıkmasını umuyorum. R programlama dilini öğrenmiş araştırmacılara vakit kazandırmak, R programlama dilinde istediği gelişmeyi gösteremeyen araştırmacılara yardımcı olmak amacıyla hazırlanacak R tabanlı interaktif web uygulamalarına temel teşkil edecek özgün bir eser ortaya çıkmasını umuyorum. Ülkemizin, diğer her alanda olduğu gibi, eğitim alanında yapılan araştırmalar için kullanılan yazılımlar anlamında da kendi kendine yeter duruma gelebilmesi, diğer ülkelerde yaşanan gelişmeleri daha yakından takip edebilmesi ve yeni gelişmelere öncü olabilmesi için R programlama becerisinin araştırmacılara kazandırılmasının önemli olduğunu düşünüyorum. Her ne kadar öğrenme eğrisi oldukça dik olsa da, programlama becerileri günümüz dünyasında bir zorunluluk haline dönüşmektedir. Stephen Hawking'inde belirttiği gibi, "evrenin gizemini çözmek için ya da 21. yüzyıl dünyasında herhangi bir kariyere sahip olmak için bireylerin programlama becerisine sahip olması gerekmektedir."

13.0.2 Konu/Kapsam

Eğitim alanında yapılan nicel araştırmalardan elde edilen sonuçların sistematik ve nesnel olarak yorumlanabilmesi için istatistiksel yöntemler kullanılmaktadır. Bu prosedürlerin tamamlanması çoğu zaman ücretli yazılımlar ile sağlanır, örneğin SPSS, SAS, Lisrel, AMOS. Ülkemizde üniversiteler bu yazılımlar için maalesef ciddi bütçeler ayırmak zorunda kalmaktadır. Araştırmacıların bu pahalı programlara ulaşmak için zaman zaman yazılım korsanlığına başvurmaları ise daha büyük bir sorun oluşturmaktadır. Eğitim alanında çalışan araştırmacıların bu ücretli (proprietary) yazılımları kullanmanın dışında başka bir alternatifleri açık kaynak (open source) tabanlı programlar olabilir. Bu araştırma önerisi, açık kaynak kodlu ve ücretsiz bir programlama dili olan R'ın eğitim alanında yapılan araştırmalarda daha yaygın kullanılabilmesini sağlamak için hazırlanacak materyallerle ve verilebilecek eğitimlerle ilgilidir.
Robert Muenchen (2011), R'ın yakın gelecekte veri analizi için kullanılan evrensel bir programlama dili olacağını söyleyerek yeni metodların SPSS ve SAS gibi yaygın yazılımlardan daha önce R'da tanıtıldığını vurgulamaktadır. Aynı zamanda Muenchen, kişisel web sitesinde 2014 yılında yayınladığı bir makalede R programının analitik akademik yayınlarda en çok kullanılan üçüncü yazılım paketi olduğunu, ücretsiz olanların içinde ise ilk sırada olduğunu belirtmiştir. Son yıllarda Harvard ve Oxford gibi tanınmış üniversiteler de R programlama dili üzerine çalıştaylar düzenlemektedirler . R programlama dilini öğretmek için hazırlanmış birçok İngilizce kitlesel açık çevrimiçi kurslar da bulunmaktadır . Microsoft, Google ve Ford gibi dünyaca

bilinen şirketler de kullandığı programlar arasında R'a yer vermişlerdir . Data analizi ve istatistiksel modellemenin yanında, işlevsel grafikler çizme, dokümanlar oluşturma, sunum hazırlama ve simülasyon üretme gibi çeşitli amaçlar için kullanılabilen R, başta istatistikçiler olmak üzere, mühendisler, ekonometristler ve sosyal bilimlerde modern ve komplike modellerle çalışan araştırmacıların dikkatini çekmiştir.

13.0.3 Literatür özeti

R programını tanıtım amaçlı birçok yabancı kaynak bulunmaktadır. Kullanıldığı amaca göre oldukça teknik olabilen bu kaynakların içinde sosyal bilimciler için uygun olabilecek kaynaklardan bazıları şunlardır; (a) Field, Miles ve Field (2012), sosyal bilimlerde sıkça kullanılan modelleri (örneğin t-test, korelasyon, ANOVA, regresyon) çoğunluğu psikoloji alanında toplanmış verilerlerin analizi için kullanmıştır, (b) Dalgaard (2008) , Field'a oranla daha teknik bir dil ile yazılmış bu kitap, sosyal bilimcilerin sıkça kullandığı modelleri teorik altyapıları ile birlikte sunmuştur, (c) Everitt ve Hothorn (2011), ölçek geliştirme ve geçerlik çalışmalarında kullanılan faktör analizlerini içeren bir uygulama kitabı yazmıştır. Ücretsiz olan ve birçok gelişmiş analize öncülük eden R programlama dilinin tanıtımı ve yaygınlaştırılması ise Avrupa ve Amerika'ya oranla Türkiye'de neredeyse hiç yapılmamıştır. Türkçe kaynaklar R'ın kısa tanıtımını yapan ve istatistik bölümü öğretim üyeleri tarafından hazırlanmış kitaplar (Sönmez, 2006; Satman, 2010; İlk,2011; Gürsakal, 2014), bir kaç blog yazısı ve konferans bildirileri ile sınırlı kalmıştır. (Baydoğan ve Çetin, 2014; Özdemir, Yıldıztepe ve Binar, 2010). Bununla beraber Google Scholar ve ulusal veri tabanlarında, sosyal bilimcilere yönelik ucretli ya da ucretsiz Türkçe bir R kaynağı bulunamamıştır.

13.0.4 Özgün Değeri

Tanınmış üniversitelerden, tanınmış bilim insanlarından ve büyük şirketlerden saygı gören R programının ülkemizde de yaygınlaşması ekonomik tasarrufun yanı sıra alanın öncülerine yetişme imkanı da sağlayabilir. Kalkınma Bakanlığının yayınladığı çalışma raporunda, Milli Savunma Bakanlığının açık kaynak kodlu yazılımların kullanılması sonucu sadece lisans bedellerinde 2012 yılı ve öncesinde yaklaşık 2 milyon Amerikan doları tasarruf edildiği belirtilmiştir . Bu çalışma, açık kaynak kodlu yazılımların kullanılması doğrultusunda, sosyal bilimler alanında atılacak ilk adımlardan biri olma niteliğindedir. Çalışmanın amacı, eğitim alanındaki araştırmacılara yönelik hazırlanacak olan R materyallerine zemin oluşturabilecek özgün bir tanıtım ve uygulama kılavuzu niteliğinde dokümanlar hazırlamaktır. Hazırlanacak dokümanlar Türkiye'de eğitim alanında toplanmış veriler üzerine inşa edilecektir. Daha sonra Türkçeye çevrilmek üzere İngilizce olarak hazırlanacak tanıtım dökümanı, programın kurulumunu, basit fonksiyonlarını, veri girdisi ve çıktısının nasıl yapılabileceğini, betimsel analizleri, basit grafik çizimlerini, t-test, varyans analizlerini, korelasyon ve regresyon analizlerini kapsayacaktır. Teorik altyapıya kısaca değinilecek, sonuçları raporlama süreci ise daha ayrıntılı verilecektir. Uygulama kılavuzu Türkçe olarak hazırlanacak ve eğitim etkinliklerinin çerçevesini oluşturacaktır. Ortaya çıkan ürünler, şu an çalıştığım kurum öncelikli olmak üzere yurt içinde verilecek çalıştaylar için kullanılabilecektir. Araştırmanın gerçekleşmesi halinde, programın yaygınlaşmasını sağlayabilecek gelişmiş R materyalleri için de zemin oluşacaktır.

13.0.5 Yaygın etki/katma değer

Amaç bölümünde belirtildiği gibi doküman halka açık bir çevrimiçi depoda tutulacak ve üniversitelerin eğitim fakültelerine elektronik posta ile tanıtılacaktır. Dokümanların kabul görmesi durumunda, kullanıcılardan gelen istekleri ve önerileri hızlı bir şekilde dokümanlara dahil ederek, ihtiyaçlara cevap verme ihtimali yüksek olan bir kaynak ortaya çıkabilir. Yine özgün değer bölümünde belirttiğim gibi, tamamen açık kaynak üzerinden çalışan R programının, ülkemizde yaygınlaşması durumunda ithal yazılım gerekliliğini azaltarak tasarruf etme ihtimali mevcuttur.

13.0.6 Genel Tavsiye

1) eksikler var
2) duzeltme 1

Bibliography

Adler, D. and Murdoch, D. (2017). *rgl: 3D Visualization Using OpenGL*. R package version 0.97.0.

Aho, K. (2016). *asbio: A Collection of Statistical Tools for Biologists*. R package version 1.3-4.

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., and Hyndman, R. (2016). *rmarkdown: Dynamic Documents for R*. R package version 1.0.9014.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37(3):379–384.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Box, G. E. P. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley, New York.

Breheny, P. and Burchett, W. (2016). *visreg: Visualization of Regression Models*. R package version 2.3-0.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3):145–153.

Daunic, A. P., Smith, S. W., Garvan, C. W., Barber, B. R., Becker, M. K., Peters, C. D., Taylor, G. G., Van Loan, C. L., Li, W., and Naranjo, A. H. (2012). Reducing developmental risk for emotional/behavioral problems: A randomized controlled trial examining the tools for getting along curriculum. *Journal of School Psychology*, 50(2):149–166.

de Vreeze, J. (2016). *apaStyle: Generate APA Tables for MS Word*. R package version 0.4.

Field, A. P., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Sage, Thousand Oaks, Calif;London;.

Hirshleifer, S., McKenzie, D., Almeida, R., and Ridao-Cano, C. (2016). The impact of vocational training for the unemployed: Experimental evidence from turkey. *The Economic Journal*, 126(597):2115–2146.

Højsgaard, S. and Halekoh, U. (2016). *doBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities*. R package version 4.5-15.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Komsta, L. and Novomestky, F. (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4:863.

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.4-0.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duursma, R., Biancotto, E., Levy, O., Dutang, C., Solymos, P., Engelmann, R., Hecker, M., Steinbeck, F., Borchers, H., Singmann, H., Toal, T., and Ogle, D. (2016). *plotrix: Various Plotting Functions*. R package version 3.6-3.

Lumley, T. and Zeileis, A. (2015). *sandwich: Robust Covariance Matrix Estimators*. R package version 2.3-4.

Mair, P. and Wilcox, R. (2016). *WRS2: A Collection of Robust Statistical Methods*. R package version 0.9-1.

Muenchen, R. A. (2011). *R for SAS and SPSS users*. Springer, New York, 2nd edition.

Myers, J. L., Well, A., Lorch, R. F., and Corporation, E. (2013). *Research design and statistical analysis*. Routledge, New York, 3rd edition.

Obrien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, 41(5).

Olejnik, S. and Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8(4):434–447.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460.

Pearl, J. (2009). *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge;New York, 2nd edition.

R Core Team (2016a). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* R package version 0.8-67.

R Core Team (2016b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer, New York, 2nd edition.

Revelle, W. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.6.9.

RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Sarkar, D. (2016). *lattice: Trellis Graphics for R*. R package version 0.20-34.

Tippmann, S. (2015). Programming tools: adventures with r: a guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis. *Nature*, (7532):109.

Torchiano, M. (2016). *effsize: Efficient Effect Size Computation*. R package version 0.7.0.

Uebersax, J. S. (2015). Introduction to the tetrachoric and polychoric correlation coefficients. *Obtenido de http://www. john-uebersax. com/stat/tetra. htm.[Links]*.

Verzani, J. (2014). *Using R for introductory statistics*. CRC Press Taylor and Francis Group, Boca Raton, second edition.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2016). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.0.

Wickham, H. and Chang, W. (2016). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 2.2.0.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, US, 3rd;3; edition.

Xie, Y. (2016). *bookdown: Authoring Books with R Markdown*. R package version 0.1.