SARP: Sosyal Bilimler R Platformu

Burak AYDIN, James ALGINA, Walter LEITE

2017

Contents

Chapter 1

Kapak

Figure 1.1:

Bu platformun hakları korunmuştur CC0 by Burak AYDIN.

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 inşaa edilmiştir, Bookdown ise R Markdown (Allaire et al., 2016) üzerine inşa edilmiştir. Materyalin hazırlanma aşamasında R Studio (RStudio Team, 2016) kullanılmıştır.

1.1.1 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.

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

2.1 Yazarlar

Akademik çalışmalarımız araştırma tasarıları (research design) ve nicel veri analizine yoğunlaşmıştır. Monte Carlo simulasyonları mutlaka çalışma takvimimizde yer alır. Diğer bir orta noktamız çok düzeyli modellere 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

2.2 Teşekkür

Burada katkı sağlayanların isimlerine yer verilecektir.

[1] sadece Türkçe versiyonunda yer alır

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 4).

Windows için R yükle

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 4) ile gösterilmiştir.

R betik

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 4 ile gösterilmiştir.

R Studio Yükle

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]  0.0429  1.7411 -0.0195  0.0669  1.2783 -2.2039 -1.2294  0.7491
##  [9]  0.6592  0.3697  0.1006 -0.2617
rnorm(12,mean=10,sd=2) #~ N(10,2) 12 gözlem
##  [1]  6.63 10.49 11.11  8.85 12.10  8.74 10.11 15.69  7.99 13.86 11.62
## [12] 11.58
runif(12, min = 10, max = 37)
##  [1] 36.9 27.0 23.6 21.6 27.4 34.2 24.6 18.7 16.5 30.5 23.1 22.3
```

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,] 33.0 28.6 48.9 38.9
## [2,] 39.3 33.8 49.4 50.0
## [3,] 28.5 37.6 38.8 41.6
## [4,] 27.8 43.8 45.1 45.8
A*B    # çarp
##      [,1] [,2] [,3] [,4]
## [1,] 32.0  118  359  336
## [2,] 74.7  167  394  503
## [3,] 76.6  214  306  399
## [4,] 95.0  286  397  477
A%*%B  # matris çarp
##      [,1] [,2] [,3] [,4]
## [1,]  757  903  918  833
## [2,]  876 1021 1058  951
## [3,]  995 1138 1198 1069
## [4,] 1113 1256 1338 1188
t(B)   # çevir
##      [,1] [,2] [,3] [,4]
## [1,] 32.0 37.3 25.5 23.8
## [2,] 23.6 27.8 30.6 35.8
## [3,] 39.9 39.4 27.8 33.1
## [4,] 25.9 36.0 26.6 29.8
```

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] "2017-03-20"
Sys.Date( )-dt
## Time differences in days
## [1]  8328 10378  9606 14242 15551  8172
```

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  a     var        M zayif   1     52   maas 1994-06-01 muhasebe
## 2  t     var        F  orta   3     75   maas 1988-10-20   garson
## 3  z     var        F   iyi   2     39 destek 1990-12-01 temizlik
## 4  k     yok        M   iyi   2     62   <NA> 1978-03-23   garson
## 5  q     yok        F zayif   1     24   <NA> 1974-08-22 muhasebe
## 6  r     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 "a","k","q","r",..: 1 5 6 2 3 4
##  $ 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 bilgisayarınıza 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.5.3.

R Paketi Yükle

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 kaydedebilirsiniz. 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.

CSV Oku

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.

R Studio ile CSV Oku

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.

```r
# 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 hesaplanması olasıdır. 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 olasıdır. 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.

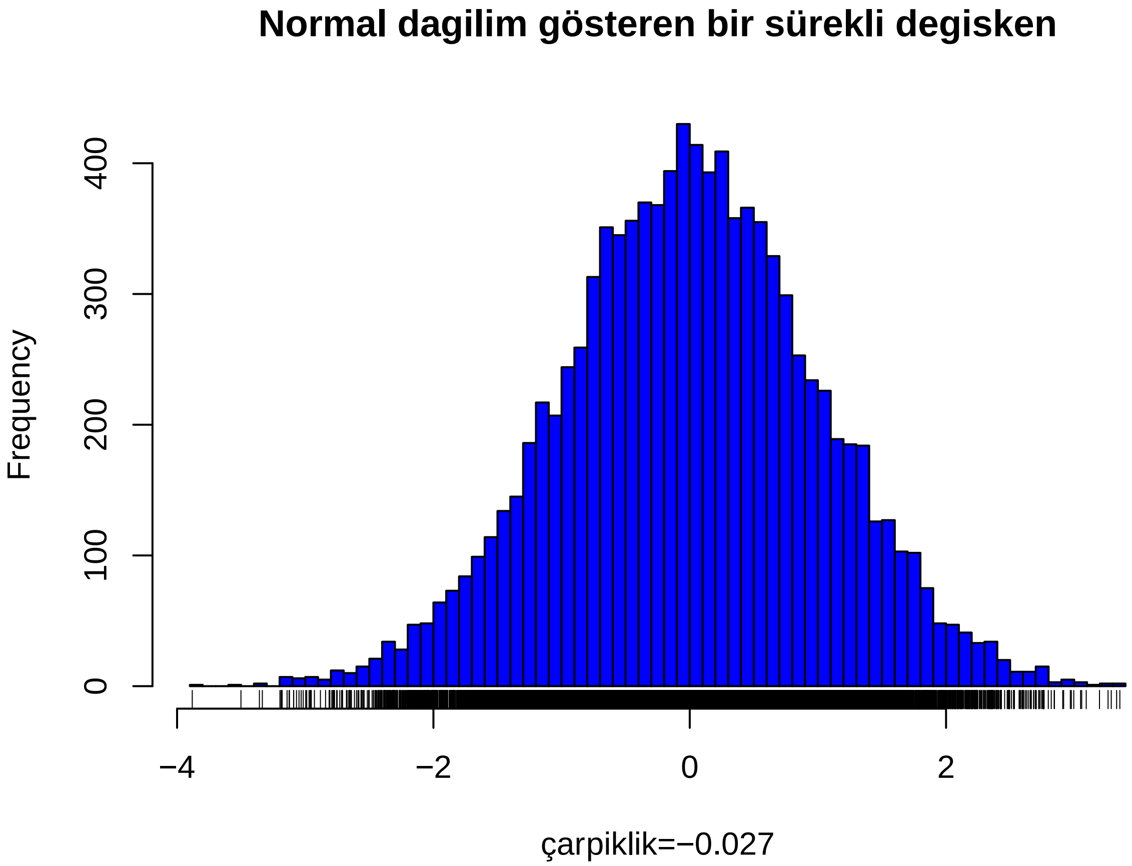


Figure 7.1: Normal dagilim gösteren bir sürekli degisken

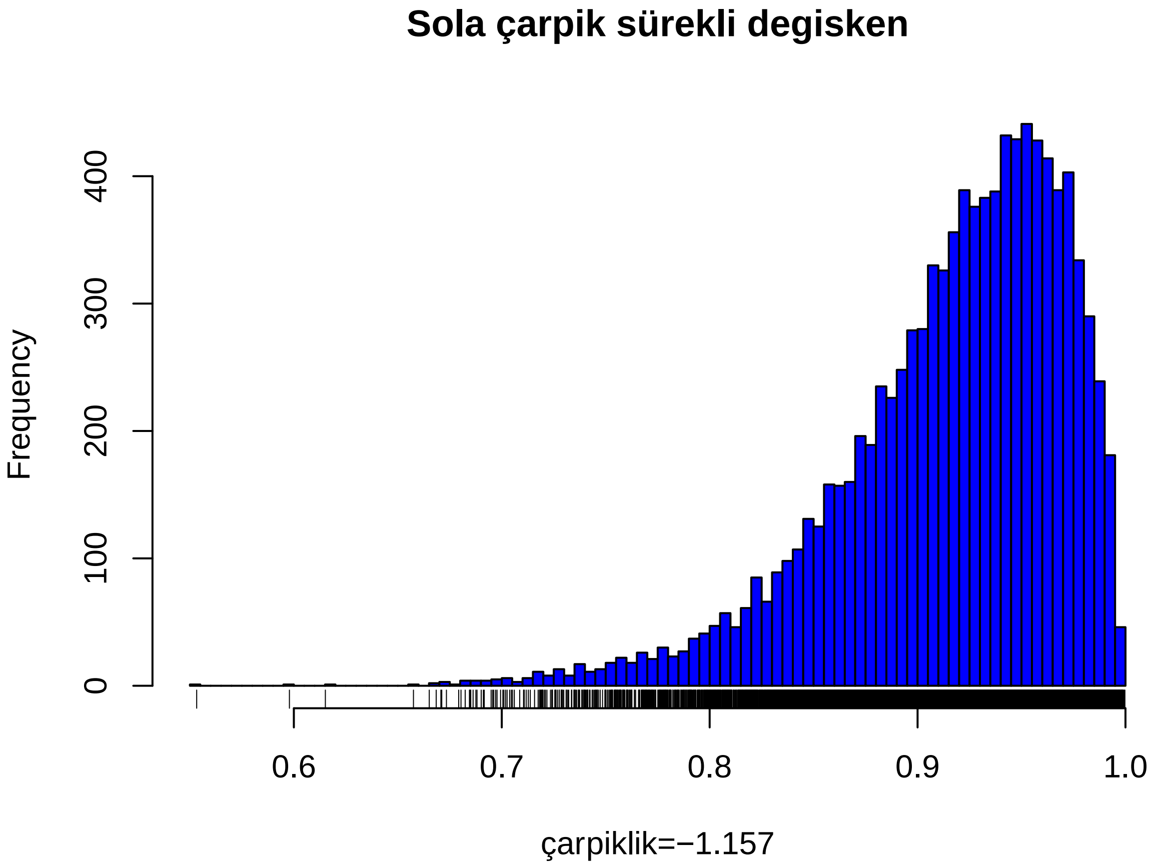


Figure 7.2: Sola çarpik sürekli degisken

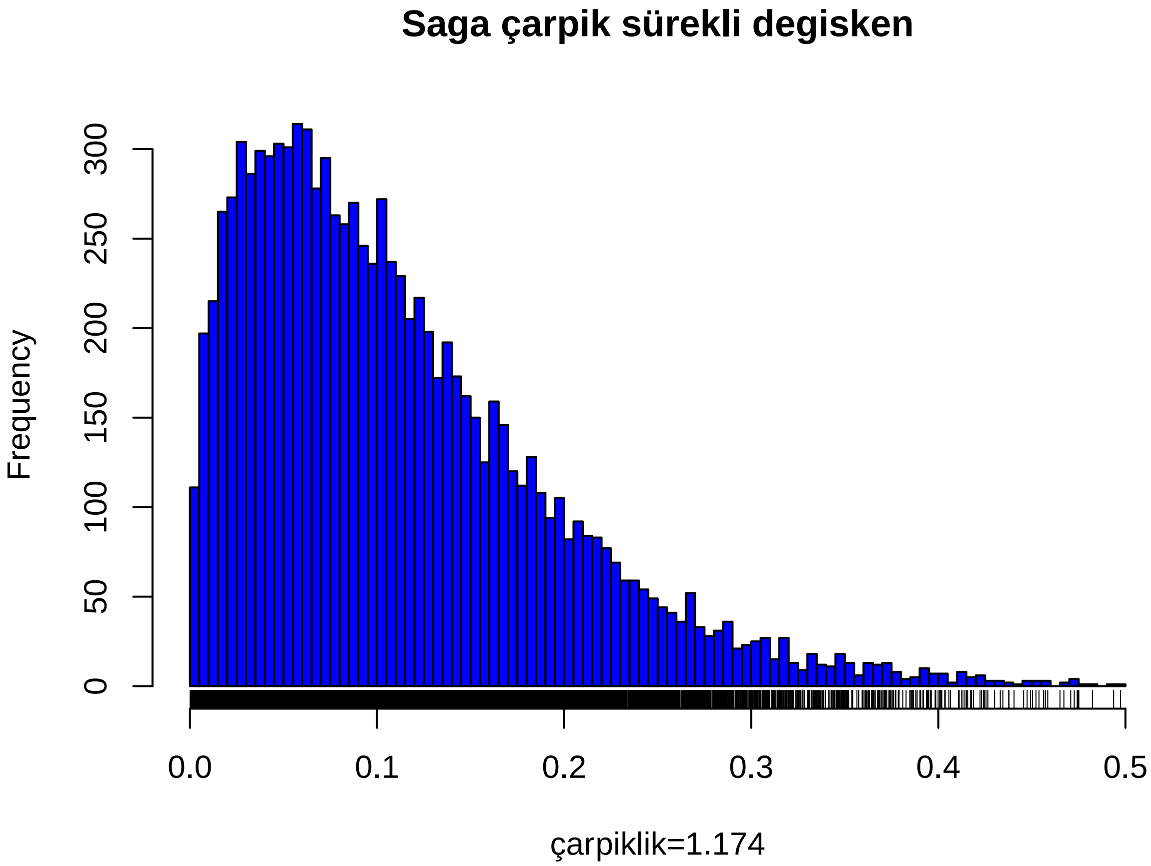


Figure 7.3: Saga çarpik sürekli degisken

7.1.6.1 Basıklık Örnekleri

Normal bir dağılım ve basıklık ölçüsü

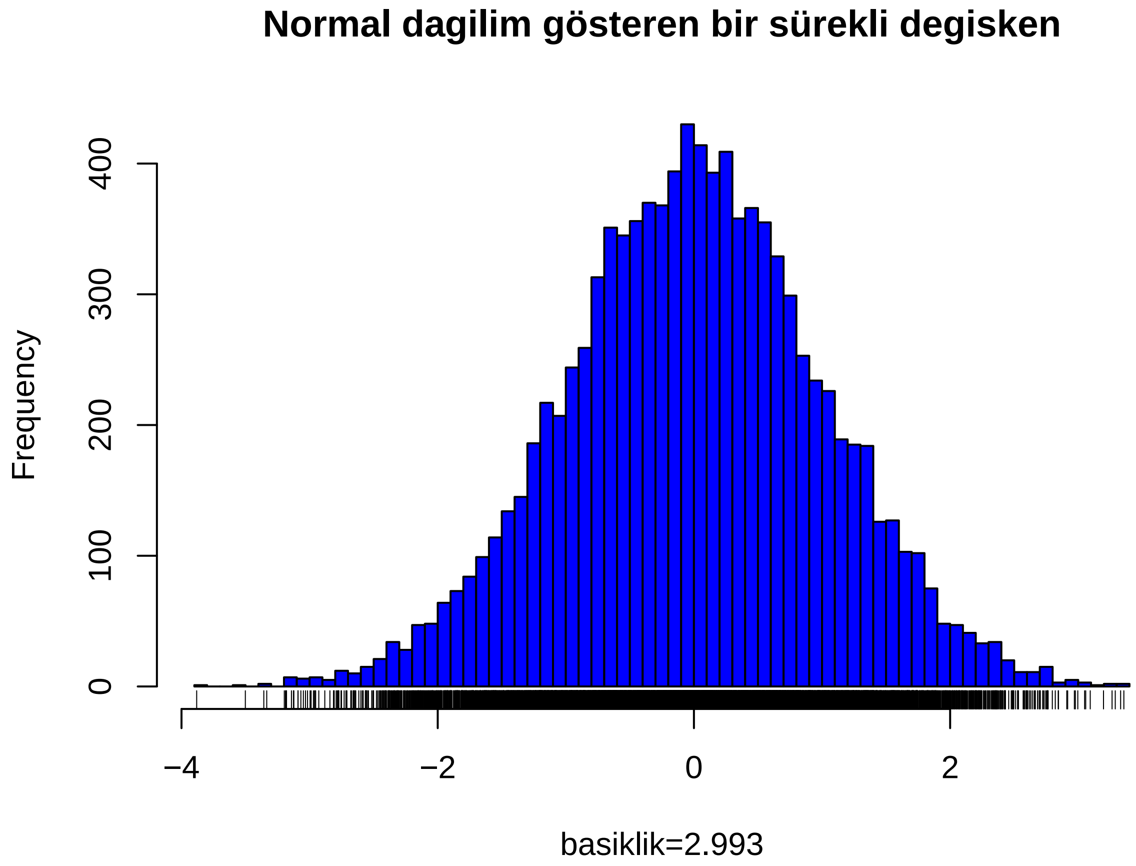


Figure 7.4: Normal dagilim gösteren bir sürekli degisken

Tekdüzey bir dağılım ve basıklık değeri

Beta dağılımı gösteren bir sürekli değişken

7.1.7 Betimleyici İstatisklerin Raporlanması

Library *psych* (Revelle, 2016), *doBy* (Højsgaard and Halekoh, 2016) and *apaStyle* (de Vreeze, 2016) MIGHT be helpful for reporting results but it might need further modifications, for example rows should not be numbered. Following R code outputs descriptive statistics for the gender attitudes average score and age.

```r
# psych package's describe function reports;
# n: the number of available scores
# mean, sd, median, trimmed mean (trim=0.05 5% trimmed)
# median absolute deviation, minimum, maximum, range
# skew and kurtosis-3 (type=2 provide population estimates)
# standard error
library(psych)
```

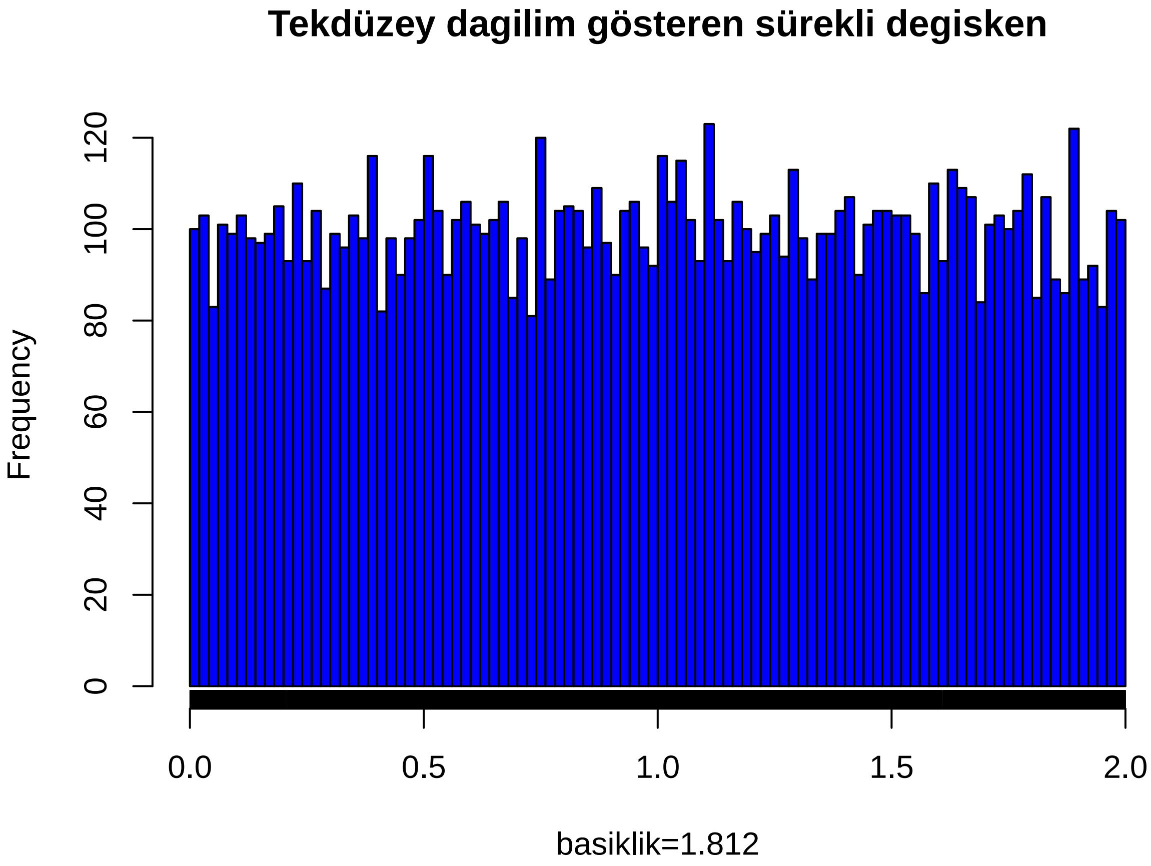


Figure 7.5: Tekdüzey dagilim gösteren sürekli degisken

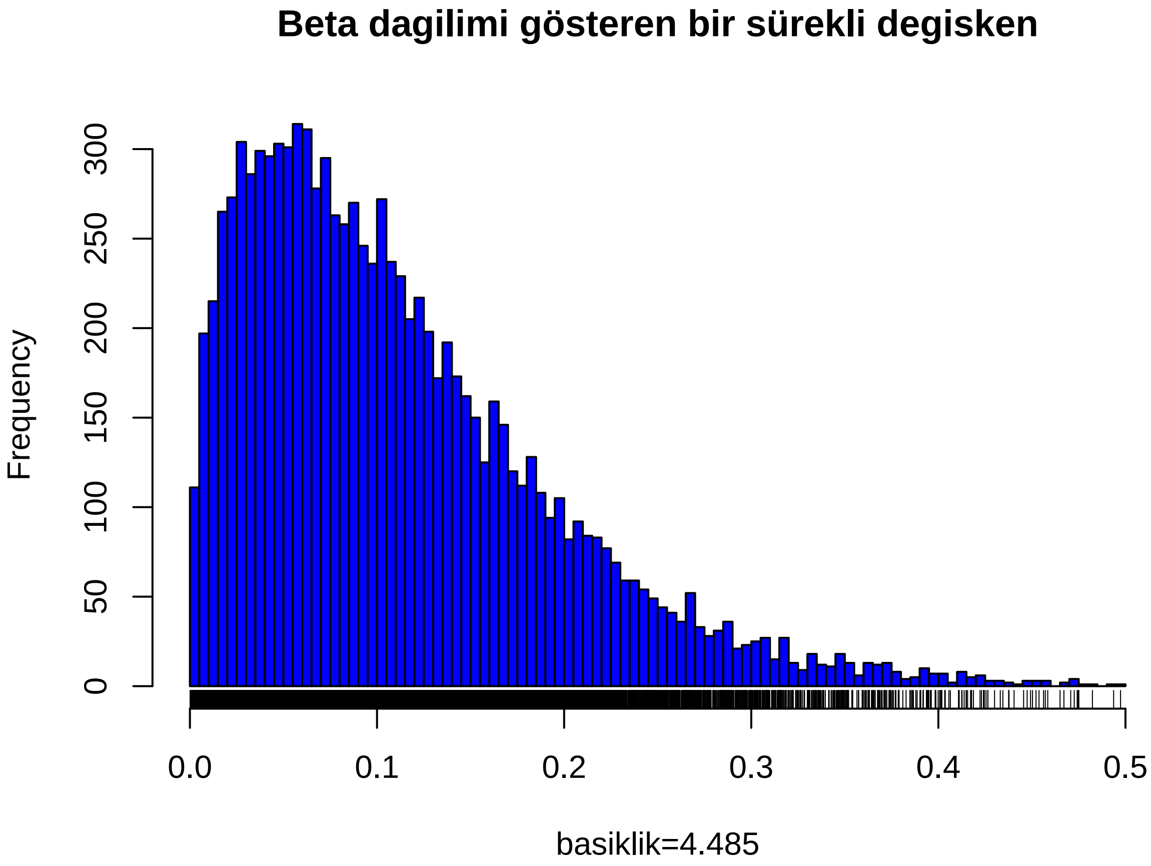


Figure 7.6: Beta dagilimi gösteren bir sürekli degisken

```
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

# export
write.csv(desc1,file="pscyhdesc.csv")

#doBy
# summaryBy is a wrapper, provide variables and functions
# its useful for summary by group
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)) } ))

#set decimals=2 using round function
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
# create APA style table as a word file
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

#if you are receiving Rjava error a quick fix is described here
#https://www.r-statistics.com/2012/08/how-to-load-the-rjava-package-after-the-error-java_home-cannot-be
```

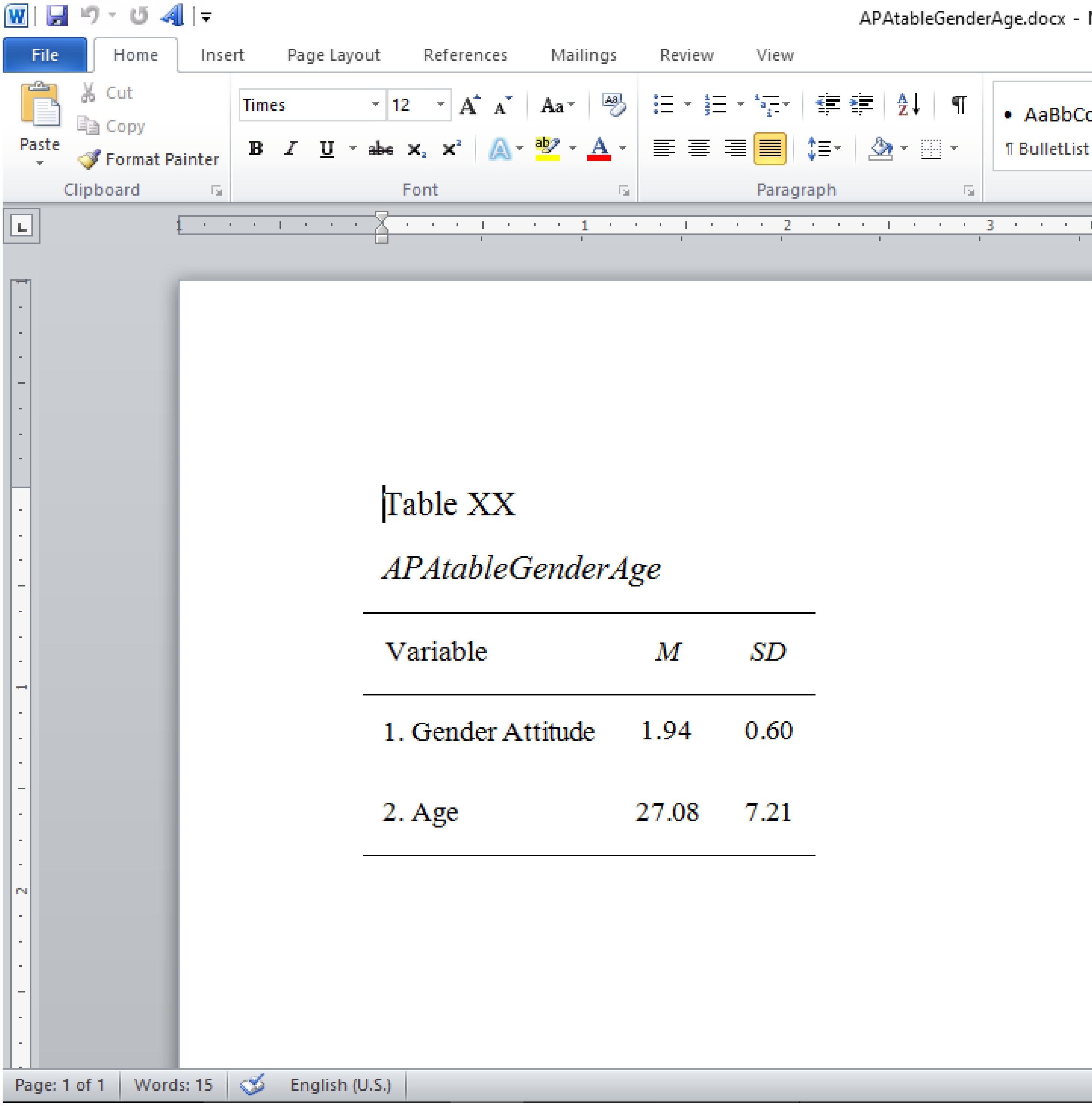


Table XX

APAtableGenderAge

Variable	*M*	*SD*
1. Gender Attitude	1.94	0.60
2. Age	27.08	7.21

Figure 7.7: APAtableGenderAge.docx

7.1.7.1 Write-up

The Gender Attitudes score from 5302 participants had a range of 1–4, a median of 2, a mean of 1.94 and SD=0.6. The score distribution has a sample skewness value of 0.38 and a sample kurtosis value of -0.1. [2]

7.2 Basic graphics

One of R's strong suit is its graphing capabilities. There are several plotting families; including R base(R Core Team, 2016b), lattice(Sarkar, 2016), ggplot2(Wickham and Chang, 2016) and plotrix(Lemon et al., 2016). We prefer to use ggplot2. This subsection briefly includes basics. The number of arguments in a *ggplot* function is large, enabling a user to manipulate every detail in a graph [3] .

7.2.1 Histogram

A histogram is a diagram of rectangles. These rectangles are created as function of frequency/relative frequency given any variable.

7.2.1.1 Histogram of one variable

Useful for distributional evaluation.

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

7.2.1.2 Histogram of one variable by one factor

Useful for evaluating group differences.

```
dataWBT$HEF=droplevels(factor(dataWBT$higher_ed,
                    levels = c(0,1),
                    labels = c("non-college", "college")))

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 = "Gender Attitude",fill='Higher Ed.')+
  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 = "Gender Attitude")+ facet_wrap(~ HEF)+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

[2] Descriptive statistics are calculated with *psych* (Revelle, 2016) package and a histogram 7.8 is created by ggplot2 (Wickham and Chang, 2016).

[3] ggplot cheatsheet might be helpful https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

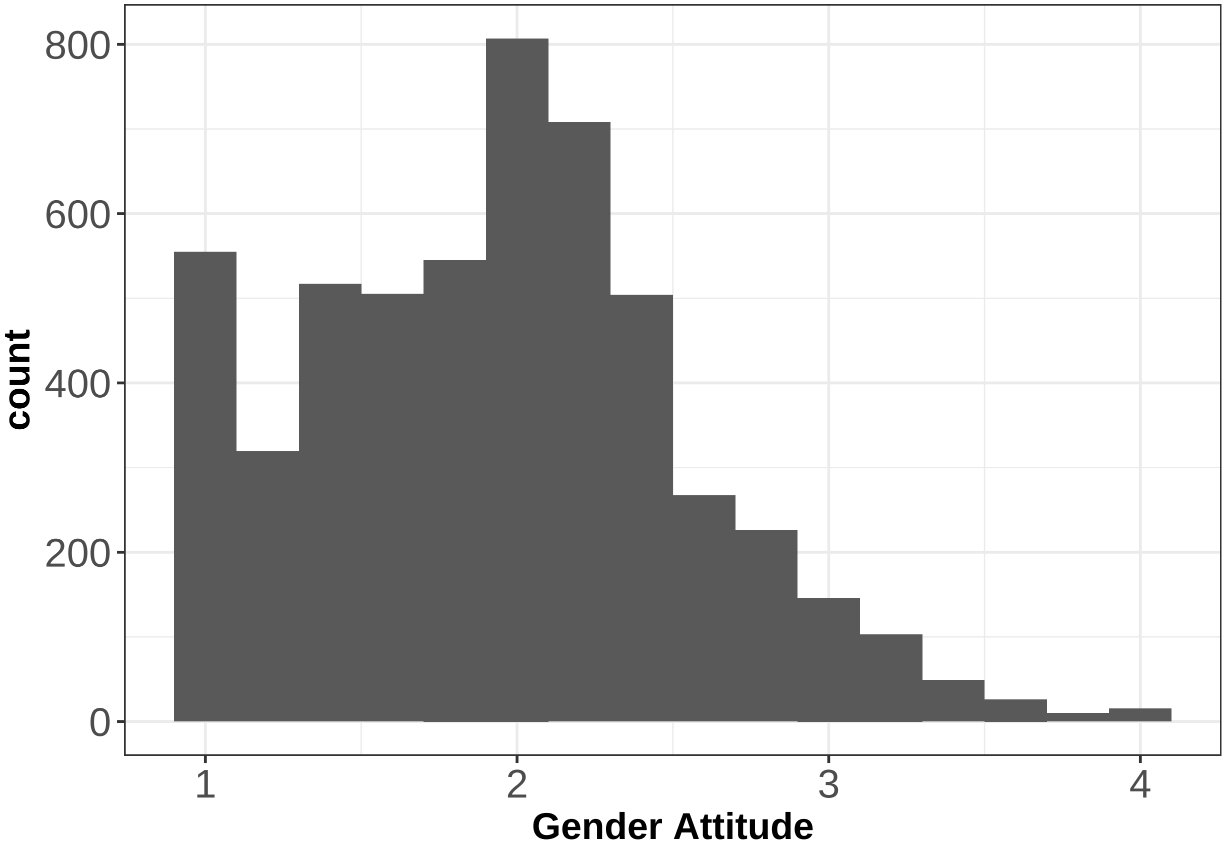


Figure 7.8: Gender Attitudes Score Distribution

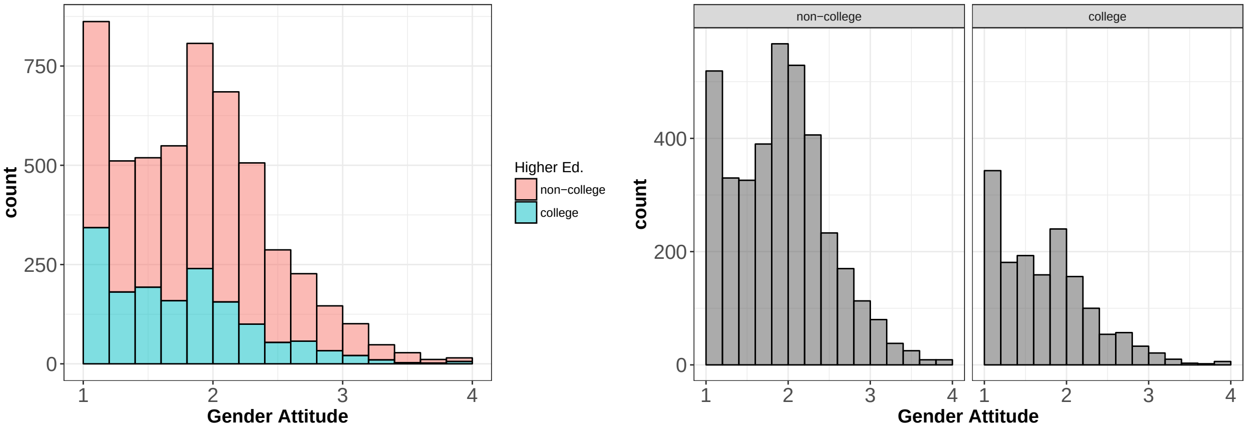
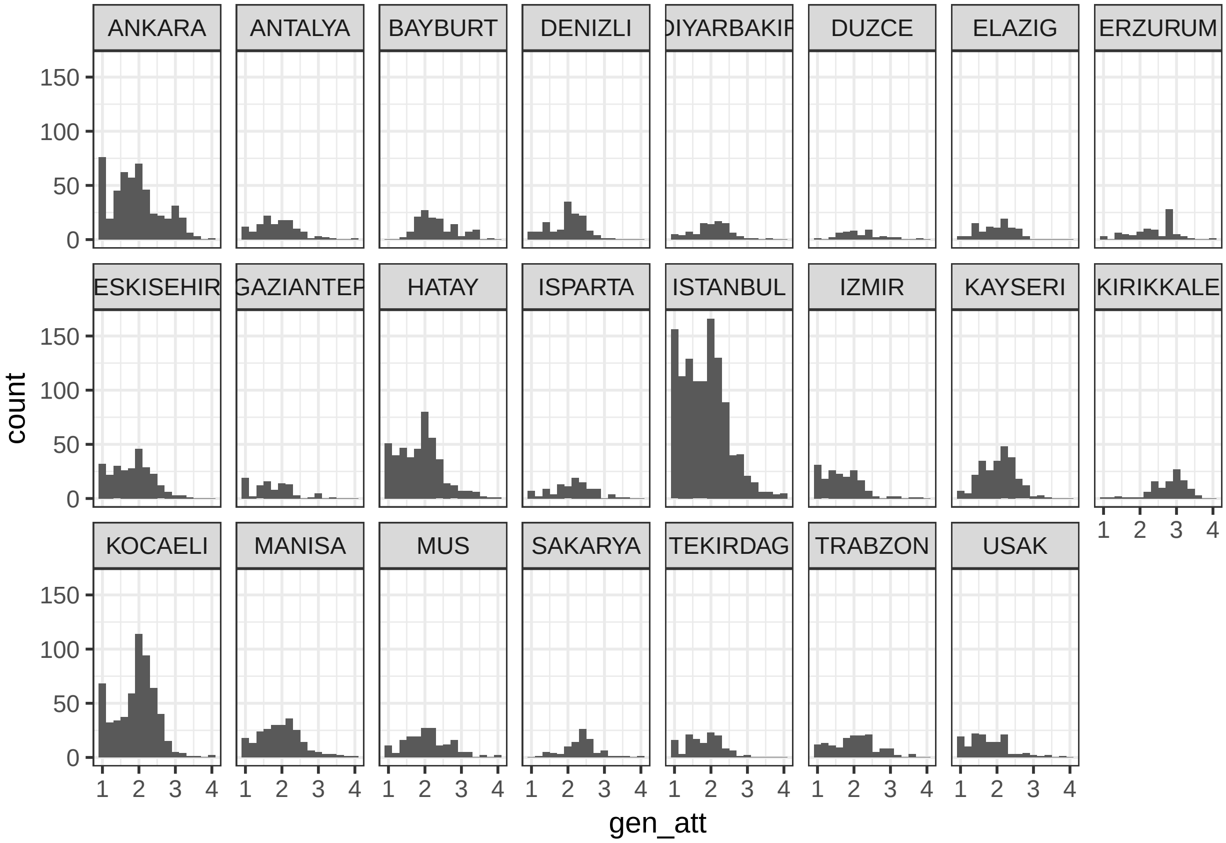


Figure 7.9: Gender Attitudes by Treatment Group

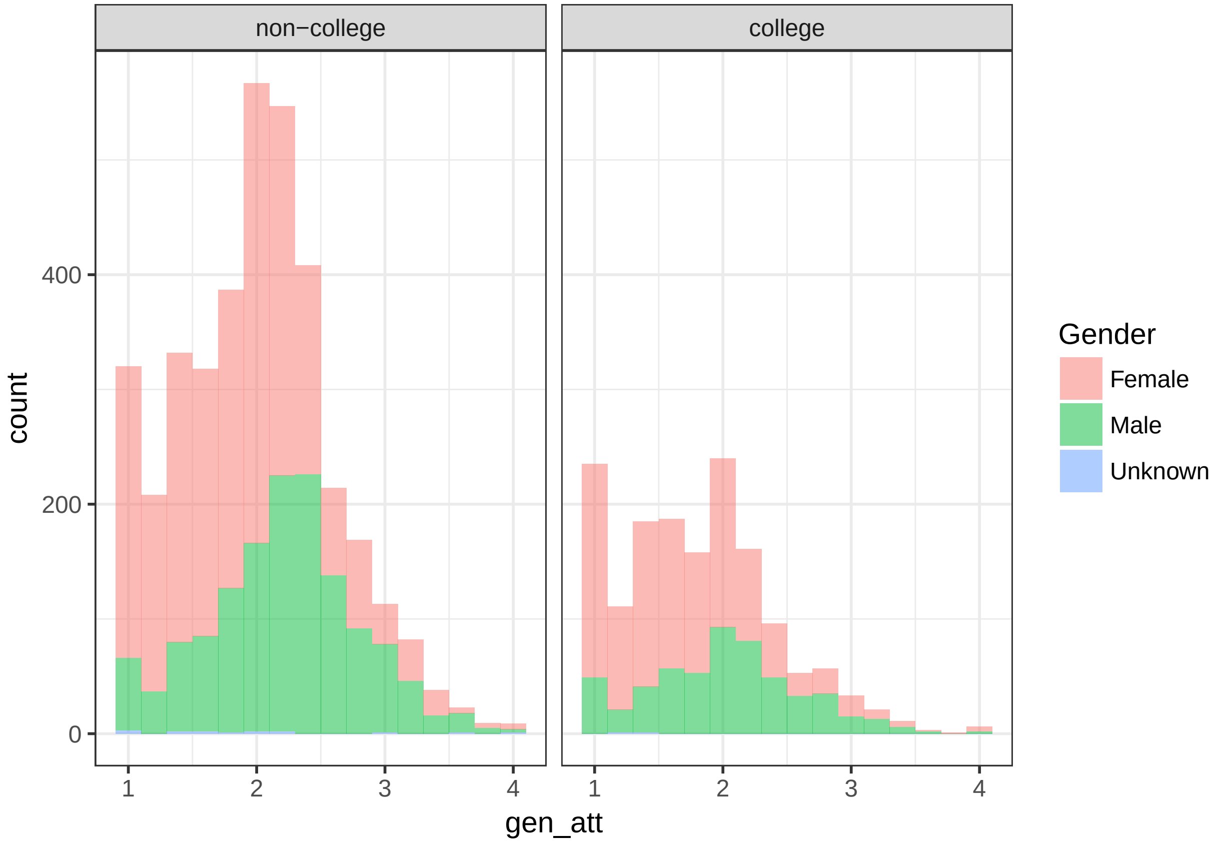
```
library(ggplot2)
ggplot(dataWBT, aes(x = gen_att)) +
  geom_histogram(binwidth = 0.2)+ theme_bw()+
   facet_wrap(~city, ncol = 8)
```



7.2.1.3 Histogram of one variable by two factors

Useful for two way interactions

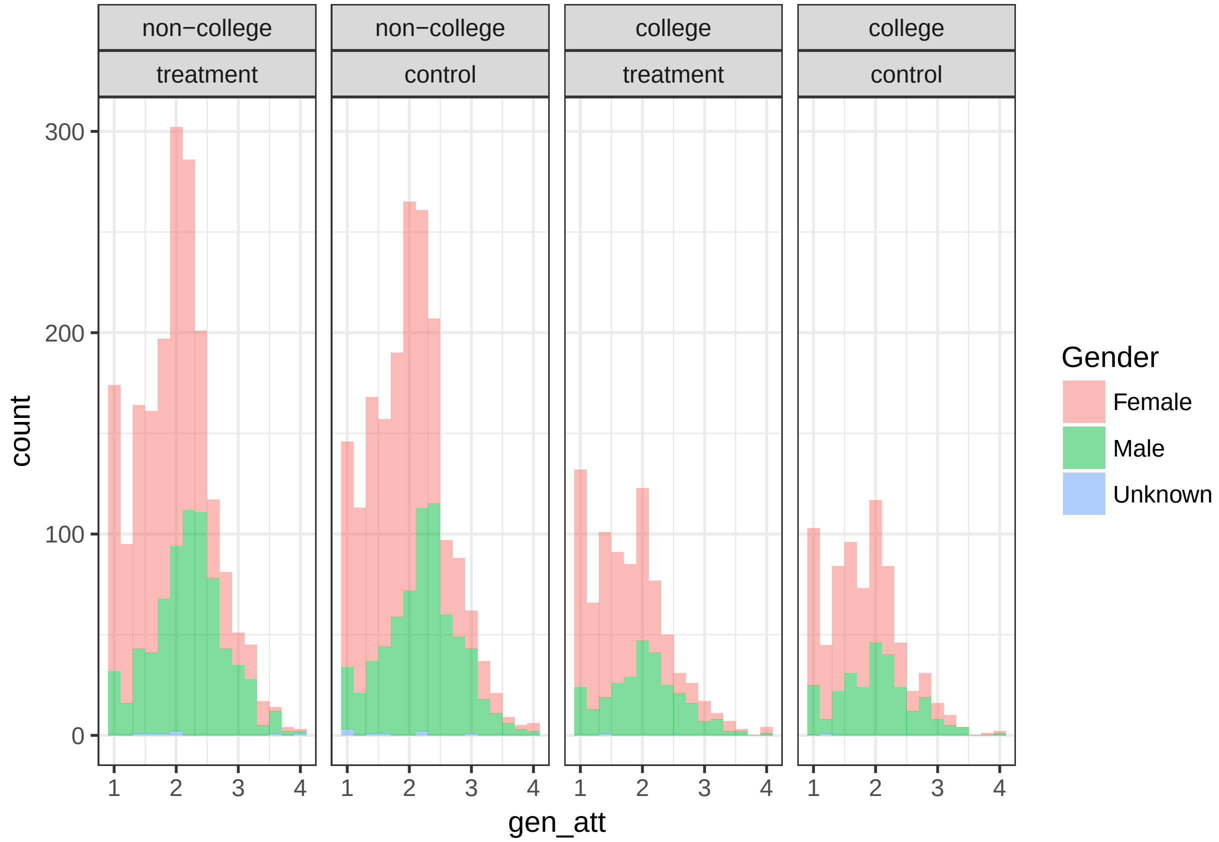
```
dataWBT2=na.omit(dataWBT[,c("gen_att","HEF","gender")])

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 Histogram of one variable by three factors

Useful for three way interactions

```
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 Hypothesis testing introduction

The Cambridge dictionary returns “all the people or animals of a particular type or group who live in one country, area, or place” as the definition of population. In social sciences, generally, a population is “all the people of a particular group”, for example *8 year old students*, or, *8 year old students in a specific country*, or *8 year old dyslectic students*. In any study the researcher can determine the population relevant to the research aims. Any measurable characteristics of the units in a given population can form a variable. In other words, the population of the variable can be definable. In section 5.2.3, possible variable types are identified. The population of the variable includes all of the possible values (outcomes), forms the range and probabilities of occurrence. Densities (for continuous) and mass functions (for discrete) can be used to summarize these probabilities. With a valid distributional assumption for the variable, we can infer from the sample to population.

A random sample from a population might or might not include all of the possible values. But a random sample is expected to be selected so that there is no systematic bias in the selection and therefore to be similar to the population especially when the sample is large. A population parameter is estimated with a model using the information from the sample. Fitting a model consists of evaluating the degree of discrepancy between a model and the observed data. Hypothesis tests (or statistical inferences) based on a fitted model aim to reach a conclusion in terms of the substance of the problem.

7.3.1 The Sampling distribution

A statistic computed from a random sample is a random variable and has a distribution. The most common example of a sampling distribution is the sampling distribution of the mean. The central limit theorem implies that under simple random sampling[4], regardless of the shape of the distribution of the variables, the sampling distribution of the mean can be approximated by a normal distribution ;

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

as the sample size gets larger. In addition if the shape of the distribution is normal, the sampling distribution of the mean is given by (7.6) for all sample sizes.

The standard deviation of a sampling distribution of the mean is called as the standard error of the mean and it is used in statistical inference.

The population parameters μ and σ^2 in (7.6) are unknown, but the expression is useful in understanding how well the sample mean is likely to approximate the population mean. Suppose a researcher plans to draw a simple random sample of size n=10. According to (7.6) the sampling distribution will be approximately normally distributed with mean μ and standard deviation $\sigma\sqrt{10}$. Suppose that unknown to the researcher, $\mu = 100$ and $\sigma = 15$. Then the sampling distribution will have standard deviation $(15\sqrt{10}) = 4.74$ and there will be approximately a 95% chance that the sample mean will be between 90.7 and 109.3, an interval that suggests a sample size of 10 will result in a sample mean which could be quite inaccurate. If the researcher draws a simple random sample of 100, there will be approximately a 95% chance that the sample mean will be between 97.1 and 102.9, an interval that suggests a reasonably accurate sample mean.

At this point, the question is which estimator is unbiased, consistent and efficient to estimate the expected values, hence the population parameters. It can be shown mathematically that Equation (7.1) is an unbiased estimator of μ and Equation (7.2) is an unbiased estimator of σ^2.

7.3.1.1 Unbiased estimation and sampling

To be added

7.3.2 The Confidence Intervals (CI)

Using an assumption about the distribution, information from the sample and an appropriate estimator (to produce a point estimate), confidence intervals can be constructed. A confidence interval might include the population parameter and yields correct decisions, however if it does not include the population parameter, erroneous decisions are made. Creating a CI for a sample mean is straight forward. Assuming sampling from a normal distribution, the distribution is normal [5] and the sample mean is an unbiased estimator. A normal distribution has known properties, the density function implies 95% of the density lies within 1.96 standard deviations from the mean. A visual is given below (Figure 7.10), the probability of a random draw from the blue area is only 5%. Similarly, the probability is 10% for a drawn from the blue or yellow area. The grey area (±1) represents approximately 68% of the density. This information is useful. Using the sample mean and variance, the 95% confidence interval for μ can be created.

7.3.2.1 A confidence interval example

Below R code calculates the sample mean, the standard deviation and the confidence interval for the Gender Attitude scores' mean.

[4] every member of a population has an equal probability of being selected, a selected member does not affect any other member's probablity of being selected

[5] with small sample size, a t distribution. Not assuming sampling from a normal distribution, it is approximately normal with a large sample size

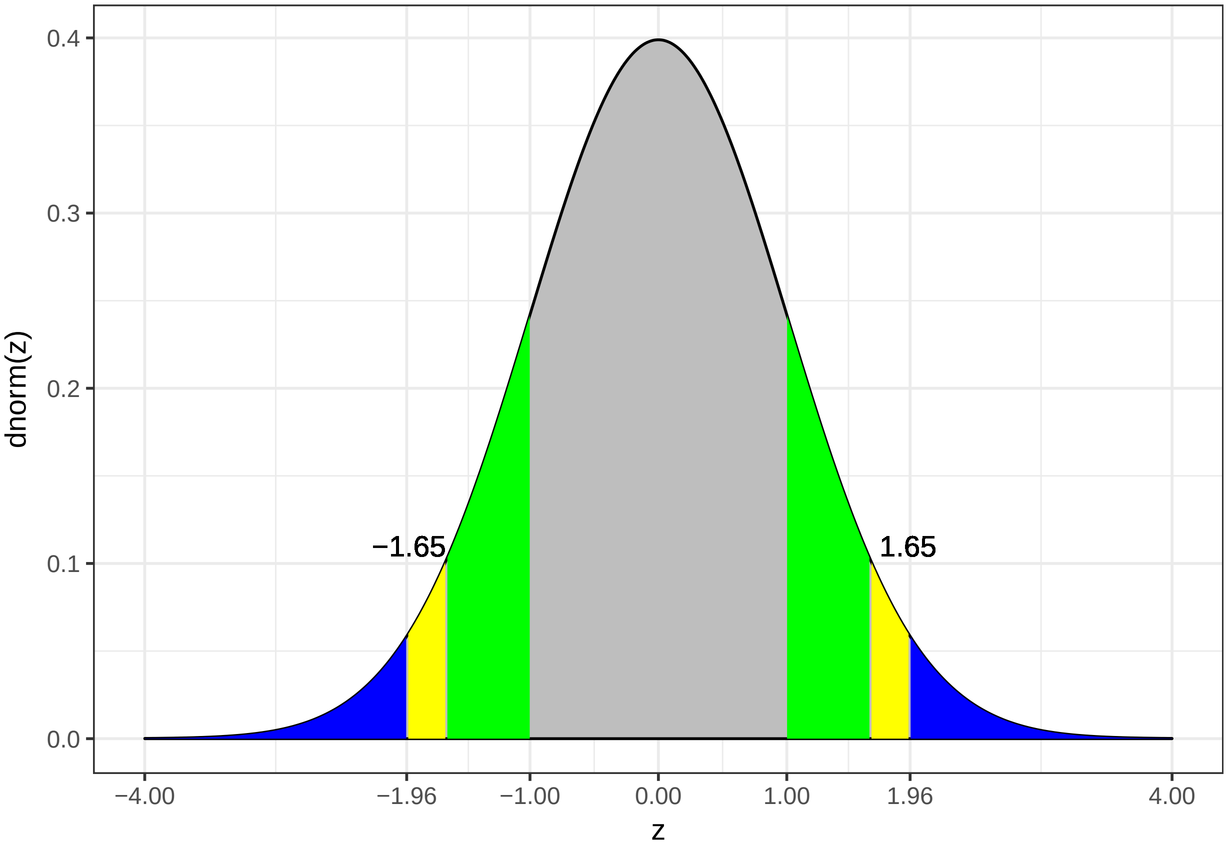


Figure 7.10: The z distribution

```
# the number of available data points, n
GA_n=sum(!is.na(dataWBT$gen_att))

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

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

#95% confidence interval
lower=GA_m - 1.96 * (GA_s/sqrt(GA_n))
lower
## [1] 1.92
upper=GA_m + 1.96 * (GA_s/sqrt(GA_n))
upper
## [1] 1.96

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


#the value 1.96 can be called by qnorm(0.975)
```

7.3.2.2 Write up

The Gender Attitudes score from 5302 participants had a mean score of 1.94 and 95% CI was 1.92–1.96 (SD=0.60).

7.3.3 The null hypothesis

The purpose of a hypothesis test is to determine which of two hypotheses about the population are supported by the sample data. A hypothesis test includes mainly 5 steps;

1) State the null hypothesis (for example $\mu = 0$)
2) Select an alternative hypothesis. (for example $\mu \neq 0$)
3) Select a test statistic
4) Make a decision by comparing the calculated value of the test statistic to the critical value. If the calculated test statistic is more extreme than the critical value then we reject the null hypothesis.The critical value depends on the alternative hypothesis.
5) State a conclusion. That is state what the decision means in term of the substance of the problem.

The null (H_0) and alternative (H_1) hypotheses are established to answer the research question. Statistical evidence is used to decide whether to reject or fail to reject the null hypothesis. Rejecting or retaining a null hypothesis is a decision. The following table shows some important concepts in how statisticians think about hypothesis testing.

State of Nature	Decision	Result
H_0	Fail to reject H_0	Correct
H_0	Reject H_0	Incorrect *(Type I error,α)*

State of Nature	**Decision**	**Result**
H_1	Reject H_0	Correct
H_1	Fail to reject H_0	Incorrect *(Type II error, β)*

Type I error—the act of rejecting H_0 when it is true, a false positive error. Alpha, α, is the probability of rejecting H_0 when it is true. An important goal in hypothesis testing is to ensure that α is sufficiently small. Usually this goal is met by requiring α to be .05, which says that the researcher is willing to tolerate a .05 probability that they will conclude there is a difference (Reject H_0) when H_0 is true.

Type II error - the act of failing to reject a hypothesis that is false. Beta, β, is the probability of retaining H_0 when it is false.

7.3.4 The z score and the z test

A general formula for z is

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

This z-variable, also known as z-score, has a mean of 0 and standard deviation of 1. If X is normally distributed, z will also be normally distributed.

```
# Create z-scores for the Gender Attitudes

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

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

# Scale function can be used for more than 1 variable
# center=T substracts mean from each score.
# scale=T divide the difference by standard deviation
# try scale(dataWBT$gen_att, center=3, scale=2)to substract 3 from each score and divide by 2.
```

The z test for a sample mean is straight forward;

$$z = \frac{\bar{X} - \mu_{hypothesis}}{Standarderrorofthemean} = \frac{\bar{X} - \mu_{hypothoses}}{\sigma_X / \sqrt{n}}$$

This z statistic can be interpreted using a z distribution (Figure 7.10);

a) If the alternative hypothesis states that the observed mean is expected to be lower than the hypothesized mean, the z statistic is compared to z_{alpha} or $-z_{(1-alpha)}$. The null hypothesis is rejected if the z-statistic is less than or equal to z_{alpha}.

b) If the alternative hypothesis states that the observed mean is expected to be different than the hypothesized mean, the absolute value of the z statistic,$|z|$ is compared to $z_{1-(alpha/2)}$. The absolute value of the z-statistic should be larger (or equal) than $z_{1-(alpha/2)}$ to reject the null.

c) If the alternative hypothesis states that the observed mean is expected to be greater than the hypothesized mean, the z statistic is compared to $z_{1-alpha}$. The z-statistic should be larger (or equal) than $z_{1-(alpha)}$ to reject the null.

Here it should be emphasized that a directional alternative hypothesis (cases a and c) has a substantially different criteria compared to a non-directional (case b) hypothesis. The researcher should provide justification for the alternative hypothesis that is used.

7.3.4.1 z test illustration-1 (non-directional)

Stating the null as $H_0 : \mu_{GenderAttitudes} = 2$ and alternative as $H_1 : \mu_{GenderAttitudes} \neq 2$ and using $\alpha = 0.05$;

```r
# the number of available data points, n
GA_n=sum(!is.na(dataWBT$gen_att))

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

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

# set the null
mu_hyp=2

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

#the critical value for alpha=0.05 and nondirectional test
qnorm(1-(0.05/2))
## [1] 1.96
```

The Gender Attitudes score from 5302 participants had a mean of 1.94 and SD=0.6. A one-sample z test revealed that the observed mean is 7.17 standard deviation below the hypothesized mean of 2. Using a rejection criteria of 1.96 ($z_{1-(0.05/2)}$) the difference between the observed mean and the hypothesized mean was concluded to be statistically significant.

7.3.4.2 z test illustration-2 (directional)

In this illustration, the Gender Attitudes scores' population mean is assumed to be 1.9 with a standard deviation of 0.75. When the population standard deviation is known it should be used. Stating the null as $H_0 : \mu_{GenderAttitudes} = 1.9$ and alternative as $H_1 : \mu_{GenderAttitudes} > 1.9$ and using $\alpha = 0.01$;

```r
# set the null
mu_hyp=1.9

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

#the critical value for alpha=0.01 and directional test
qnorm(1-(0.01))
## [1] 2.33
```

Using a critical value of 2.33 ($z_{0.99}$), results indicated that the Gender Attitudes scores' mean was significantly greater than the hypothesized value of 1.9 ($z = 3.94$).

7.3.5 The one-sample t test

Interpreting a z statistic based on a z distribution is not valid for small sample sizes. If the sample size is small, a t distribution with a n-1 degrees of freedom is valid assuming the population has a normal distribution. The procedure is the same as the z-statistic, but the critical values change.

7.3.5.1 t test illustration-1 (non-directional)

In the dataWBT, city *DUZCE* has only 52 participants and 47 available Gender Attitudes scores. For illustrative purposes this city is chosen.

Stating the null as $H_0 : \mu_{GenderAttitudes} = 1.94$ and alternative as $H_1 : \mu_{GenderAttitudes} \neq 1.94$ and using $\alpha = 0.05$;

```
dataWBT_DUZCE=dataWBT[dataWBT$city=="DUZCE",]
#descriptive statistics
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

#critical value
qt(.975,df=46)
## [1] 2.01
```

The Gender Attitudes scores from 47 participants in DUZCE had a range of 1–3.8, a median of 2, a mean of 2.18 and SD=0.55. The score distribution had a sample skewness value of 0.56 and a sample kurtosis value of 0.28.[6] A one sample t-test revealed a significant difference, t(46)=2.94 between the city's observed mean and the hypothesized mean of 1.94 using a critical value of 2.01 ($t_{.975,46}$).

7.3.5.2 t test illustration-2

In the previous example a directional test was conducted in which the alternative hypothesis specified that the population mean would not be equal to 1.94. What will happen if the null hypothesis is $H_1 : \mu_{GenderAttitudes} \leq 1.94$?

For the city *DUZCE*, stating the null as $H_0 : \mu_{GenderAttitudes} = 1.94$ and alternative as $H_1 : \mu_{GenderAttitudes} \leq 1.94$ and using $\alpha = 0.05$;

```
#t test
t.test(dataWBT_DUZCE$gen_att,
       alternative="less",
       mu=1.94,
       conf.level = 0.95)
```

[6] Descriptive statistics were calculated with *psych* (Revelle, 2016) package.

```
##
##  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

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

A one sample t-test, t(46)=2.94, revealed that the evidence does not support a conclusion that the population mean is smaller than 1.94, using a critical value of -1.68 ($t_{.05,46}$).

7.3.6 The p value

The t-test illustrations (the t.test function) reported a p-value. Calculation of a p-value is based on the assumption that the null hypothesis is true and an assumption about the distribution of the test statistic . The p-value aims to inform if the calculated statistic is ordinary or not for a given distribution. Historically, a p-value smaller than the pre-determined alpha value led researchers to conclude whether a finding is statistically significant.

7.3.7 The p value illustration

Assuming a z-distribution is valid, and the calculated z-statistic is 1.80, following visual is drawn.

The blue area corresponds to 3.6% of the density, in other words p=0.0359;

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

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

This p-value is valid for a directional test but not for a non-directional test. When the uncertainty exists for the direction, the following visual depicts the situation;

The blue area, now, corresponds to 7.2% of the density, in other words p=0.0719;

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

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

7.3.8 Statistical power

The power of a statistical test is the probability that it will correctly reject the null hypothesis,and is equal to $1 - \beta$. This probability can be computed a-priori or post-hoc, whereas a post-hoc analysis is less useful. A-priori power analyses is helpful to design a study and to decide the desired sample size. A-priori power analyses are required for the related grant proposals.

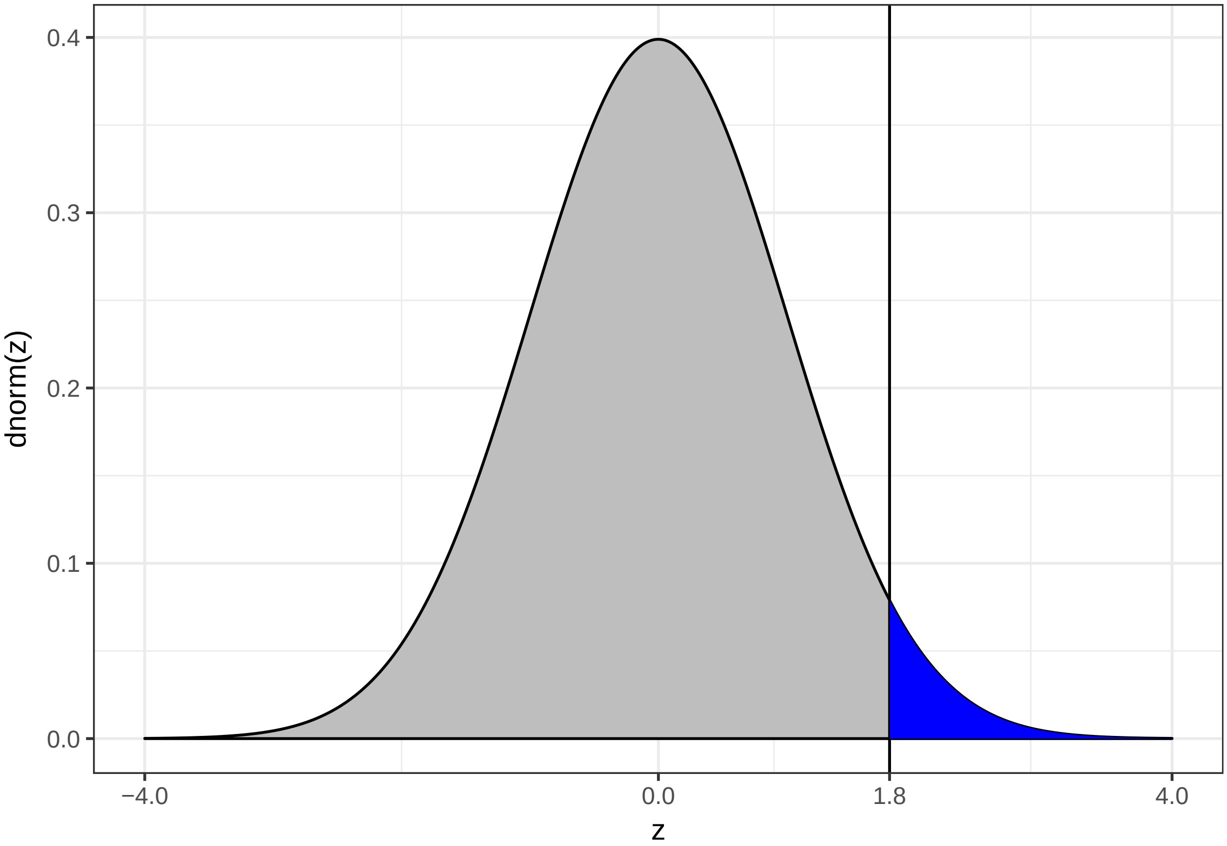


Figure 7.11: The z distribution and z=1.8

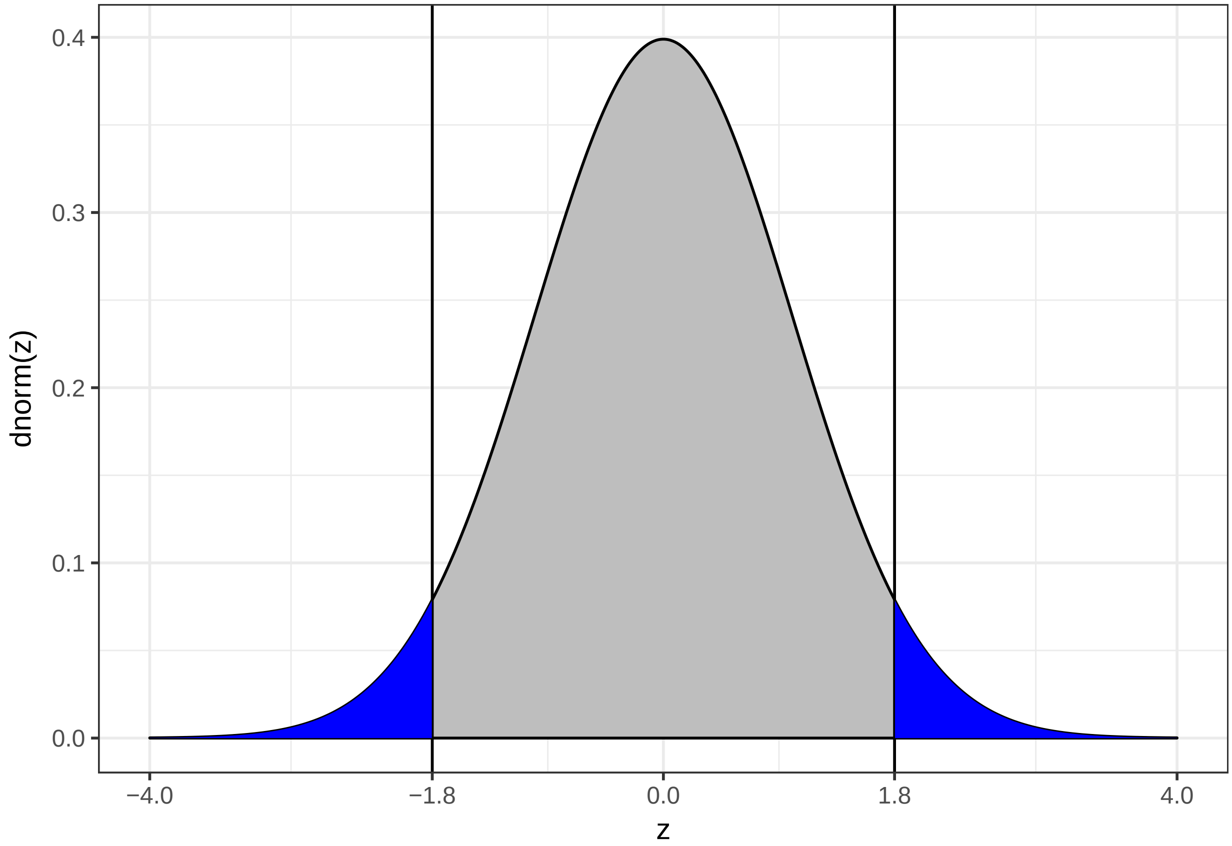


Figure 7.12: The z distribution and abs(z)=1.8

The plot produce by the following R program can be used to explain statistical power[7].

```r
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")+
  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()
```

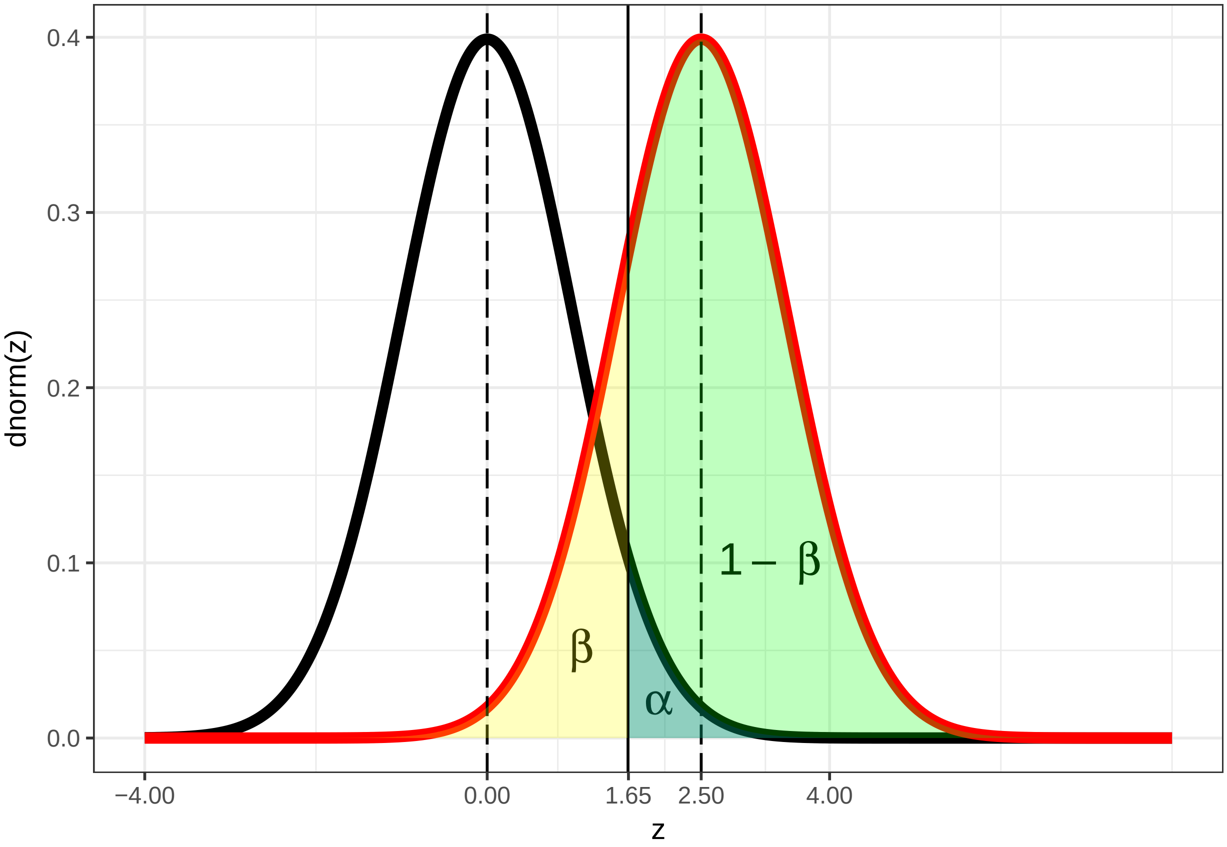


Figure 7.13: Power illustration with z distribution

The z distribution assuming a true null hypothesis ($H_0 : \mu = 0$) is depicted with black borders; the mean for the sampling distribution is 0 under this assumption and is shown with a dashed line. The z distribution assuming (a) the null hypothesis is false and (b) the actual population mean and standard deviation are such that $((\mu - \mu_{hypothesis})(\sigma\sqrt{n}) = 2.5)$ is depicted with red borders. This visual is valid for a directional test with $\alpha = 0.05$, hence, with a critical value of $z_{0.95} = 1.65$. The blue area represents α, the yellow area

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

represent β and the green area represents power. In this particular case the power is .804.

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

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

Figure 7.13 visually shows that[8] a power calculation includes 2 distributions, an alpha value and a test statistic. With these knowns the statistical power can be calculated. It should be noted that a test statistic has its own elements, generally a numerator and a denominator. For a z test, the numerator is the difference between the hypothesized mean and the null, whereas the denominator is the standard error of the mean $(\sigma/\sqrt(n))$. If the power is set to a constant (i.e. .80) , equation can be solved for any desired unknown. Generally, the equation is solved for n, the sample size.

The statistical power is revisited in the following chapters. Each design has its own standard error of the parameter estimate and the test statistic has its own distributional features. For a one sample t-test, the *power.t.test* function is useful;

```
#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
```

This illustration shows that for pre-determined knowns of a mean difference of 0.1, a standard deviation of 0.6, an alpha level of 0.05, a directional test and a desired power of 0.9, the sample size should be 310. In other words, the probability of rejecting the null ($H_0 : \mu = 0$) is .9 with a sample size of 310, a mean difference of 0.1, SD=0.6, alpha=0.05 and a directional test.

7.3.9 In case the z and the t distribution is not valid

Generalization from knowns to unknowns requires assumptions. A test statistic is robust to a violation of an assumption if, for a given sample size, the sampling distribution of the test statistic remains substantially the same under violation of the assumption(Verzani (2014)). It should be noted that a test statistic may be robust to violations of one assumption but not to violation of another assumption. In addition a test statistic that is robust to violation of one assumption, may not be robust to violation of that assumption when a second assumption is also violated. Even when a test statistic is robust to violation of assumptions, there may be a better test statistic to use when those assumptions are violated.

The z statistic for the one sample mean is expected to be robust against the violations of normality when the sample size is larger than 30 (Field et al. (2012), page 198). However, it should be noted that the rate at which the sampling distribution converges to normality depends on the distribution of the data. As a separate note, if the population is assumed to be normal and the sample sizes small, a t distribution is valid.

There are several approaches to produce robust statistics for a one-sample mean test, comprehensively illustrated by Wilcox (2012). Below R code calculates 95% confidence intervals using the second variation of the bootstrap-t method (Wilcox (2012), page 117)

```
#the second variation of the bootstrap-t method
# select DUZCE and perfom listwise deletion using na.omit
```

[8] accurate only for post hoc power

```
dataWBT_DUZCE=na.omit(dataWBT[dataWBT$city=="DUZCE",c("id","gen_att")])

# test whether the Gender Attitudes' mean is equal to 1.94
# assuming normality and using a t-test
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


#Calculate 95% CI using bootstrap (normality is not assumed)
set.seed(04012017)
B=5000       # number of bootstraps
alpha=0.05   # alpha

#x is the variable
# xBAR is the observed mean
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 confidence interval
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 Write up

The Gender Attitudes scores from 47 participants in DUZCE had a range of 1 to 3.8, a median of 2, a mean of 2.18 and SD=0.55. The score distribution had a sample skewness value of 0.56 and a sample kurtosis value of 0.28. Using a critical value of 2.01 ($t_{.975,46}$), a one sample t-test revealed a significant difference, t(46)=2.94 between the city's observed mean and the hypothesized mean of 1.94 . When the normality is assumed, the 95% confidence intervals using a t-distribution were [2.01,2.34]. When this assumption is not made, the 95% confidence intervals using the bootstrap-t method with 5000 replications (Wilcox, 2012) were [2.01,2.34].

7.3.10 Shiny application to visualize sampling distribution

To be added.

Chapter 8

Comparing Two Means, the t-test

Section 7.3.1 introduced the basics of a sampling distribution using the sample mean. When the interest is to compare two means the t-test is useful and the sampling distribution of the mean difference between two groups drives the analyses.

The mean of the sampling distribution of $\bar{Y_1} - \bar{Y_2}$ ($\mu_{\bar{Y_1}-\bar{Y_2}}$) is always equal to $\mu_1 - \mu_2$, but the standard deviation of the sampling distribution ($\sigma_{\bar{Y_1}-\bar{Y_2}}$) depends on the design used to collect the data.

Example: Consider an example in which the tensile strength of wounds closed by Suture and Tape is compared. The design for conducting this study will have one factor, Method of Wound Closure, with two levels, Tape and Suture. The following are two designs for conducting the study:

Within-subjects design. Incisions are made on both sides of the spine for each of 10 rats. Tape was used to close one of the wounds; the other was sutured. For each rat the wound closed by tape was determined randomly. This design is called within-subjects because the measurements under tape and suture are made on the same rat; rats are the subjects in the study.

Between-subjects design. Beginning with 20 rats, 10 are randomly assigned to have a wound closed by tape and the other 10 rats have a wound closed by suture. For each rat an incision is made on one side of the spine. The side is determined randomly for each rat.(Half of the rats assigned to each closure method have the incison on the left side of the spine and half on the right side. We ignore side of the spine as a factor in this example.) This design is called between-subjects because the measurements under tape and suture are made on different rats. An additional requirement for classifying the design as between-subjects is that no attempt was made to match the rats prior to random assignment. For example if the 20 rats were from 10 litters with different parents, the rats might have been matched on litter prior to random assignment.

One can imagine a population mean and a population standard deviation under each closure method.
For example the population mean under tape closure is the mean for an indefinitely large group of rats all of which have a wound closed by tape.

In the following comparison it is assumed that the population mean for tape closing will be the same in the within-subjects and the between-subjects design and that the population standard deviation will be the same in the within-subjects and the between-subjects design.

The corresponding assumptions for the population mean and standard deviation for the suture closing are made.

The following are the symbols for these population parameters.

Parameter for Population	Tape	Suture
Mean	μ_T	μ_S
Standard deviation	σ_T	σ_S

Parameter for Population	Tape	Suture
Sample size	n_T	n_S

Note. More generally, μ_1 and μ_2 for population means for the two treatments and σ_1 and σ_2 for population standard deviations for the two treatments.

Parameter for Sampling Distribution	Between-Subjects	Within-Subjects
Mean ($\mu_{\bar{Y_T}-\bar{Y_S}}$)	$\mu_T - \mu_S$	$\mu_T - \mu_S$
Standard deviation ($\sigma_{\bar{Y_T}-\bar{Y_S}}$)	$\sqrt{\frac{\sigma_T^2+\sigma_S^2}{n}}$	$\sqrt{\frac{\sigma_T^2+\sigma_S^2-2\sigma_T\sigma_S\rho_{TS}}{n}}$

1. ρ_{TS} is the correlation between the tensile strength scores in the tape and suture treatments in the within-subjects design.
2. The difference in the standard errors is due to ρ_{TS}. If this correlation is zero the designs result in the same standard error.

An important goal in designing a study is to make the standard error as small as possible. When the standard error is small the statistic in which we are interested will tend to be close in numeric value to the parameter we are estimating.

In data analysis we must select a formula for a standard error (or for the error variance). Selecting the wrong formula is a critical error in data analysis.

In practice the standard error is selected by classifying the design as between-subjects or within subjects. This means that incorrectly classifying the design is a critical error in data analysis.

8.1 Between-Subjects t-test (The Independent Groups t-test)

The gender attitudes scores for college graduates vs non-collage graduates in the city of USAK are compared. The density plot for each group's gender attitudes scores are shown below.

```r
# load csv from an online repository
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

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


# We explained the functions 'factor' and 'droplevels' in section 5.2.4
# here we create a factor, Higher Education Factor (HEF).
# it is labeled as 'non-college' when the higher_ed variable equals 0,
# 'college' when equals to 1.
# if you dont use droplevels function, you might have an empty level
dataWBT_USAK$HEF=droplevels(factor(dataWBT_USAK$higher_ed,
                    levels = c(0,1),
                    labels = c("non-college", "college")))

require(ggplot2)
```

```r
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 = "Gender Attitude by HEF in USAK")+ facet_wrap(~ HEF)+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

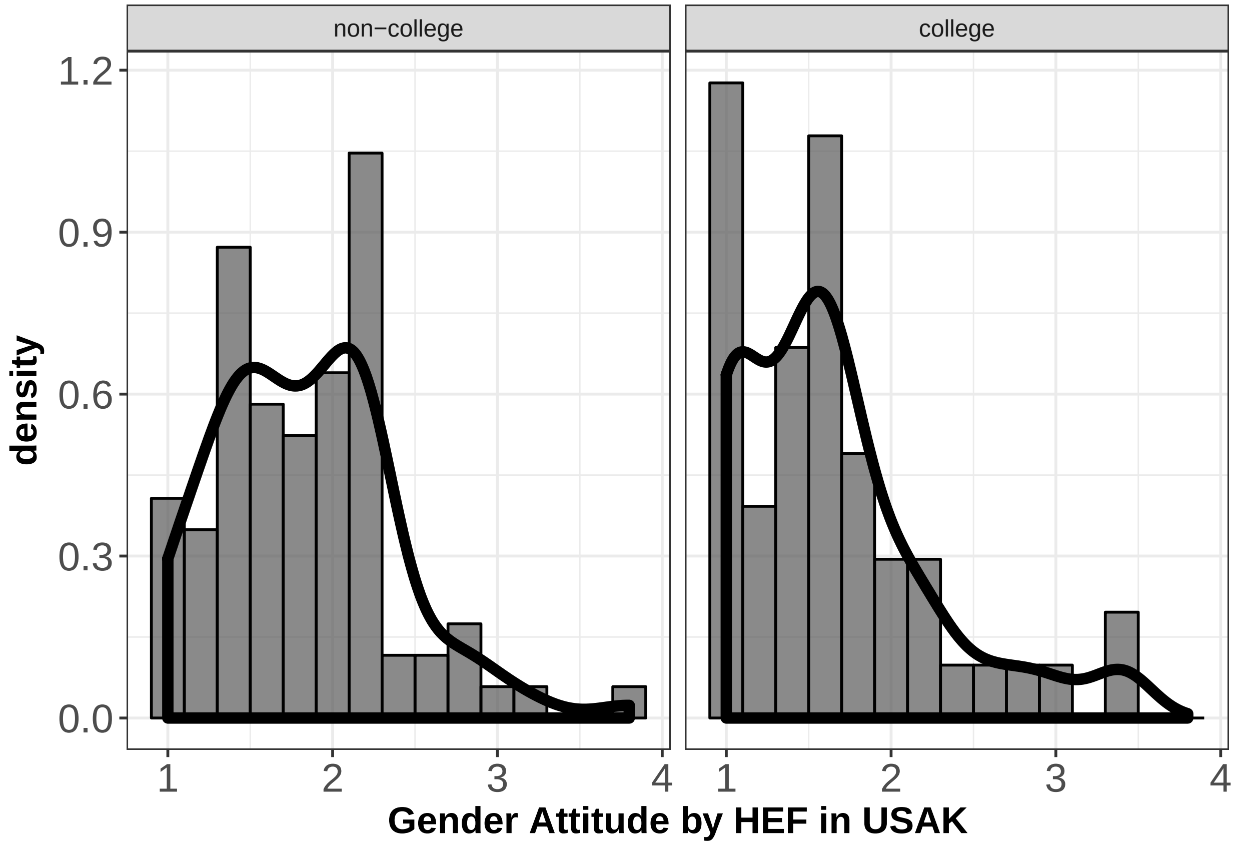


Figure 8.1: Gender Attitudes by Treatment Group

8.1.1 R codes for the independent groups t-test

The following are the steps for conducting the independent groups t-test and R code for implementing the steps

1. Create descriptive statistics
2. Calculate the test statistic

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

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

3. Find the critical value $\pm t_{\alpha/2,n_1+n_2-2}$to test

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

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

```
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
#write.csv(descIDT,file="independent_t_test_desc.csv")

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

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

# critical value for alpha=0.05
qt(.975,df=135)
## [1] 1.98
```

Since 1.963 is smaller than the critical value of $t_{.975,135} = 1.978$, H_0 is retained.

For $H_1 : \mu_1 - \mu_2 > 0$, the critical value is $t_{.95,135} = 1.66$ which would yield the rejection of H_0 given 1.93 is greater than 1.66.

For $H_1 : \mu_1 - \mu_2 < 0$, the critical value is $t_{.05,135} = -1.66$ which would yield the retaining of H_0 given 1.93 is not lower than -1.66.

A more convenient R code would be;

```
# The dataWBT does not have HEF factor,
# you should define it as it is given a few lines above.

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

# greater
```

```
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


# less
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 Write up for non-directional test:

An independent groups t-test showed that in the city of USAK, the gender attitudes scores for the college graduates (n=51, mean=1.64, SD=0.61, skew=1.19, kurtosis=1.09) were not statistically different than the non-college graduates (n=86, mean=1.83, SD=0.54, skew=0.72, kurtosis=0.90), t(135)=1.96, p=0.052. The 95% confidence interval was [-0.002,0.395].[1]:

8.1.1.2 Write up for directional test:

A directional independent groups t-test showed that in the city of USAK, the gender attitudes scores for the college graduates (n=51, mean=1.64, SD=0.61, skew=1.19, kurtosis=1.09) were significantly lower than the non-college graduates (n=86, mean=1.83, SD=0.54, skew=0.72, kurtosis=0.90), t(135)=1.96, p=0.026. The 95% confidence interval was [0.030,∞].

8.1.2 Assumptions of the independent groups t-test

Three assumptions should be met to claim statistical validity for a conventional between-subjects t-test.

[1] The descriptive statistics were calculated with the *psych* package (Revelle, 2016) and the t-test is conducted with the stats package (R Core Team, 2016b).

1. Independence . The scores in each group should be independently distributed. The validity of this assumption is questionable when (a) scores for participants within a group are collected over time or (b) the participants within a group work together in a manner such that a participant's response could have been influenced by another participant in the study. (See 9.2.1.4 for additional discussion)

2. Normality. The scores with each group are drawn from a normal distribution. However Myers et al. (2013) states that when the two groups are equal in size and the total sample size is 40 or larger departures from normality can be tolerated unless the scores are drawn from extremely skewed distributions. As noted earlier, the authors of the current book are hesitant to conduct tests for normality. However the use of robust procedures is advised when there is doubt for the normality.

3. Equal variance. This assumption is also called the homogeneity of variance assumption and means it is assumed that samples in the two groups are drawn from two populations with equal variances. Myers et al. (2013) states that when the sample sizes are equal and larger than 5, even with very large variance ratios ($s_1^2/s_2^2 = 100$) the conventional t-test leads to acceptable Type-I error rates. However this not the case with unequal sample sizes. Field et al. (2012) states that tests for the variance homogeneity, i.e. Levene, might not perform well with small and unequal sample sizes. The problems with tests on variance is that they are not powerful enough to detect inequality of variance even when it is large enough to cause problems with the t test and most are less robust to non-normality than the t test is. The *t.test* function , by default, does not assume equal variances and uses a Welch's t-test.

Even though we briefly summarized the assumptions of the independent groups t-test above, they were only introductory. For example we did not discuss violating equal variance and normality simultaneously. The discussion of what is "acceptable" is another limitation for our brief summary, for example when n1 = n2 = 10 we estimated the Type I error rate for $\alpha = .01$ and a non-directional test to be .018 based on a 100000 replications. Most people would see .018 as liberal with $\alpha = .01$

There is an enormous literature on the effects of violating the assumptions of the independent samples t test on both Type I error rate and power and a great deal is known about when the independent samples t test works well and when it does not. However, because that literature is so large it is difficult to summarize it in a way that will allow data analysts to decide in every situation if the independent samples t test should be used. Perhaps a reasonable summary is that if independence appears to be violated an appropriate alternative to the independent sample t test should be used. If independence does not appear to be violated, then when the sample sizes are equal and at least 20 in each group and the scores are approximately normally distributed the independent samples t test can be used. In other situations alternatives to the independent samples t test should be used.

8.1.3 Using Welch's t test

Welch' t-test can be conveniently implemented in R and is a reasonable choice for comparing means for independent groups when the normality is not severely violated, the groups have different sample sizes and each groups' sample size is reasonable large, (e.g. > 20) , and the homogeneity of variance assumption is not made.

```
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 Write up for non-directional Welch's t-test:

An independent groups Welch's t-test showed that in the city of USAK, the gender attitudes scores for the college graduates (n=51, mean=1.64, SD=0.61, skew=1.19, kurtosis=1.09) were not statistically different than the non-college graduates (n=86, mean=1.83, SD=0.54, skew=0.72, kurtosis=0.90), t(95.89)=1.90, p=0.06. The 95% confidence interval was [-0.008,0.402].

When the departures from the normality is severe, especially when the groups demonstrate substantially different distributions, a percentile bootstrap procedure is effective (Wilcox (2012),page 171).

```
#Calculate 95% CI using bootstrap (normality is not assumed)
set.seed(04012017)
B=5000       # number of bootstraps
alpha=0.05   # alpha

# define groups
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)

## non-directional
# D star lower
output[as.integer(B*alpha/2)+1]
## [1] -0.0134

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

##Directional x2>x1
# D star lower
output[as.integer(B*alpha)+1]
## [1] 0.022

#wrong direction x2<x1
# D star upper
output[as.integer(B*(1-alpha))]
## [1] 0.358
```

8.1.3.2 Write up for percentile bootstrap method:

In the city of USAK, the gender attitudes scores for the college graduates (n=51, mean=1.64, SD=0.61, skew=1.19, kurtosis=1.09) were not statistically different than the non-college graduates (n=86, mean=1.83, SD=0.54, skew=0.72, kurtosis=0.90) given that the 95% confidence interval was [-0.013,0.390].[2]

For a directional test: When the direction is appropriately stated in the alternative hypothesis, the lower limit of the 95% CI is 0.022 and yields the rejection of the null hypothesis of $H_0 : \mu_{non-college} = \mu_{college}$ in favor of $H_1 : \mu_{non-college} - \mu_{college} > 0$.

For a directional test: When the direction is NOT appropriately stated in the alternative hypothesis, the upper limit of the 95% CI is 0.358 and yields the retaining of the null hypothesis of $H_0 : \mu_{non-college} = \mu_{college}$ against the $H_1 : \mu_{non-college} - \mu_{college} < 0$.

8.1.4 Effect size for the independent groups t-test

A t statistic tells whether the mean difference is large in a statistical sense but not in a substantive sense. To judge whether a mean difference is large in a substantive sense one can use an effect size. Cohen's effect size is difference between the mean divided by the pooled standard deviation and can be computed using;

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

Effect sizes are often judged in terms of criteria suggested by Cohen (1962).

Effect Size	Description
.2	Small
.5	Medium
.8	Large

```
##  the normality and the equal variances assumptions are made
## given the robust procedures provided roughly the same results
n1=51
n2=86
tval=1.96

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

#or by the package effsize
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
```

[2] The descriptive statistics were calculated with the *psych* package (Revelle, 2016) and the non-directional percentile bootstrap method with 5000 replications was conducted with the base package (R Core Team, 2016b).

```
## 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
## -0.00843  0.70078
# experiment noncentral=T.
```

The effsize package (Torchiano, 2016) reported an effect size of 0.35 with a 95% CI of [-0.008, 0.701]

8.1.5 Extra: Practical significance vs statistical significance

There are a number of points to keep in mind about practical significance (a term similar to practical significance is clinical significance.) versus statistical significance.

What do these terms mean? In treatment studies, statistically significant means large enough to be unlikely to have occurred by sampling error if the population means are equal whereas practically significant means large enough to be judged as practically important. Note then that significant has a different meaning in the two terms.

In treatment studies, practical significance can be measured by the mean difference or, when the scale of measurement is not well understood, by the effect size.

The claim is sometimes made that and effect can be practically significant but not statistically significant. This would mean that the effect is judged to be large but is not statistically significant. The problem with this claim is that an effect that is large but not statistically significant can only occur in a small study. Therefore the effect will be imprecisely estimated, which undermines the credibility of the claim that the effect is practically significant.

Another claim sometimes made is that an effect can be statistically significant, but not practically significant. This claim can be correct. For example, suppose there were 400 participants in an experiment, resulting in 200 participants in each group. The researcher found a small ES of 0.20 which is significantly different than zero ($t = 2$, $p < .05$). If we regard an effect size of .2 as not practically significant then we have an effect that is statistically, but not practically significant.

8.1.6 Missing data techniques for the independent groups t-test

To be added

8.1.7 Supportive graphs for the independent groups t-test

To be added

8.1.8 Power calculations for the independent groups t-test

Section 7.3.8 provided the basics of statistical power.

```
#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
```

This illustration shows that for the pre-determined knowns of a mean difference of 0.35, a standard deviation of 0.6, an alpha level of 0.05, a non-directional test and a desired power of 0.95, the sample size should be 78 in each group. In other words, the probability of rejecting the null ($H_0 : \mu_1 - \mu_2 = 0$) is .95 with a sample size of 156, a mean difference of 0.35, SD=0.6, alpha=0.05 and a non-directional independent t-test.

8.2 The dependent groups t-test (Within-subjects t-test)

To examine whether surgical tape or suture is a better method for closing wounds, for each of 20 rats incisions were made on both sides of the spine. One of the wounds was closed by using tape; the other was sutured. The side closed by tape was determined at random. After 10 days the tensile strength of the wounds was measured. The following are the data.

```
wounds=data.frame(ratid=1:20,
                  tape=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),
                  suture=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))

# Create plot data
library(tidyr)
plotdata=gather(wounds, method, strength, tape:suture, factor_key=TRUE)

require(ggplot2)
ggplot(plotdata, aes(x = strength)) +
  geom_histogram(aes(y = ..density..),col="black",alpha=0.7) +
  geom_density(size=2) +
  theme_bw()+labs(x = "strength")+ facet_wrap(~ method)+
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=14,face="bold"))
```

8.2.1 R codes for the dependent groups t-test

The following are the steps for conducting the dependent groups t-test and R code for implementing the steps

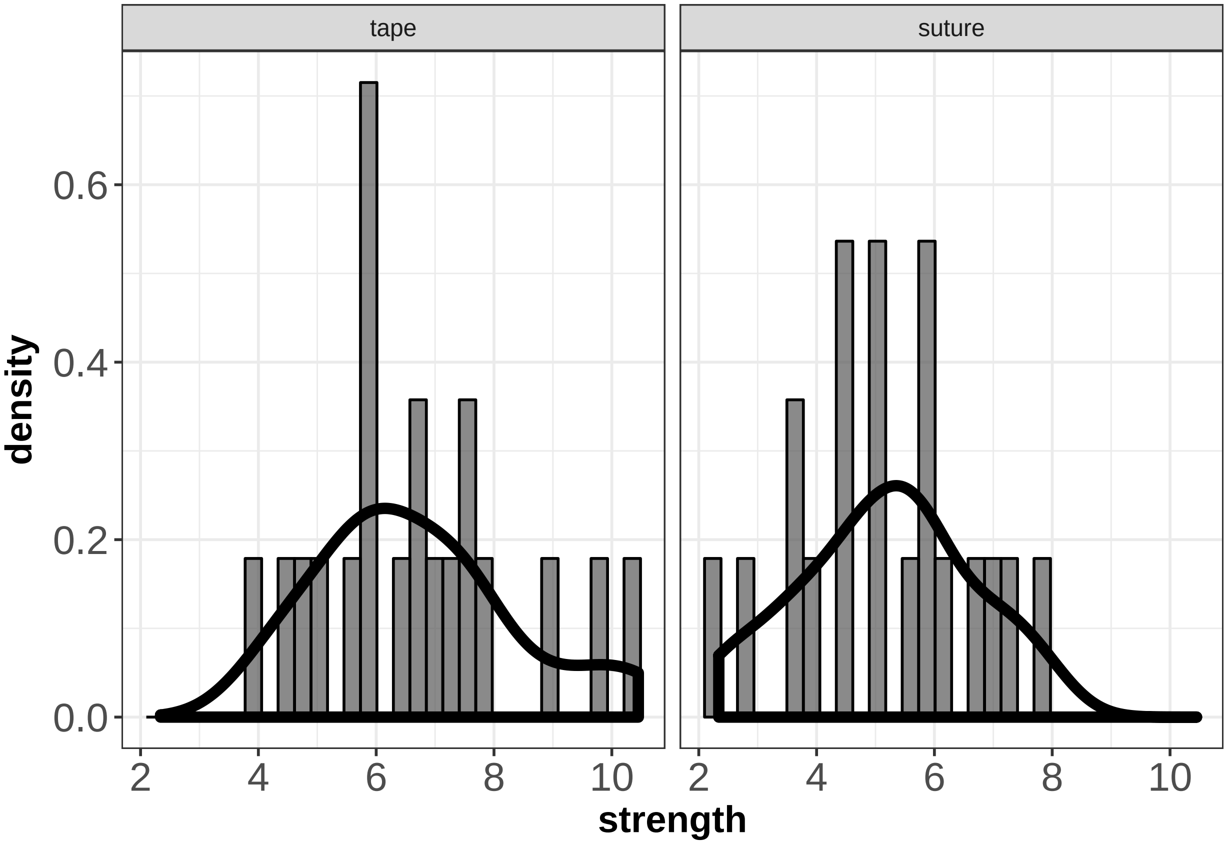


Figure 8.2: Wounds example

1. Create descriptive statistics
2. Calculate the test statistic

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

3. Find the critical value $\pm t_{\alpha/2,n-1}$to test

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

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

```
library(psych)
descDT=with(wounds,describe(cbind(tape,suture)))
descDT
##        vars  n mean   sd median trimmed  mad  min   max range  skew
## tape      1 20 6.67 1.71   6.53    6.54 1.45 3.97 10.45  6.48  0.55
## suture    2 20 5.17 1.49   5.17    5.19 1.30 2.34  7.87  5.53 -0.08
##        kurtosis   se
## tape      -0.45 0.38
## suture    -0.87 0.33

corDT=with(wounds,cor(tape,suture,use="complete.obs"))
corDT
## [1] 0.354


# estimated standard error
ese=sqrt(((1.71^2+1.49^2)-(2*1.71*1.49*corDT))/(20))

# t-statistic
tstatistic=(6.67-5.17)/ese

# critical value for alpha=0.05
qt(.975,df=19)
## [1] 2.09
```

Given 3.67 is grater than the critical value of $t_{.975,19} = 2.09$, H_0 is rejected

A more convenient R code would be;

```
library(psych)
with(wounds, t.test(tape,suture,paired=T,
                                    alternative="two.sided",
                                    conf.level=0.95))
##
##  Paired t-test
##
## data:  tape and suture
## 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 Write up for non-directional dependent groups t-test:

A dependent groups t-test showed that the tensile strength after surgical tape (mean=6.67, SD=1.71, skew=0.55, kurtosis=-0.45) was statistically different than the tensile strength after the suture (mean=5.17, SD=1.49, skew=-0.08, kurtosis=-0.87), t(19)=3.67, p=0.002 ,r=0.35. The 95% confidence interval was [0.64,2.35].

8.2.2 Assumption for the dependent groups t-test

The score difference ($Y_{1i} - Y_{2i}$) should be normally distributed and the difference scores should be independent.However,the dependent t test is expected to be robust to normality with large sample sizes.

8.2.3 Robust estimation for the dependent groups t-test

When the departures from the normality is severe, a percentile bootstrap procedure can be employed (Wilcox (2012),page 201).

```
#Calculate 95% CI using bootstrap (normality is not assumed)
set.seed(04012017)
B=5000       # number of bootstraps
alpha=0.05   # alpha

wounds=data.frame(ratid=1:20,
                  tape=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),
                  suture=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){
  #sample rows
  bs_rows=sample(wounds$ratid,replace=T,size=nrow(wounds))
  bs_sample=wounds[bs_rows,]
  mean1=mean(bs_sample$tape)
  mean2=mean(bs_sample$suture)
  output[i]=mean1-mean2
  }
output=sort(output)

## Uni-directional
# d star lower
output[as.integer(B*alpha/2)+1]
## [1] 0.686

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

##Directional x2>x1
# d star lower
output[as.integer(B*alpha)+1]
## [1] 0.837
```

```
#wrong direction x2<x1
# d star upper
output[as.integer(B*(1-alpha))]
## [1] 2.14
```

8.2.3.1 Write up for an non-directional percentile bootstrap method:

The tensile strength after surgical tape (mean=6.67, SD=1.71, skew=0.55, kurtosis=-0.45) was statistically different than the tensile strength after the suture (mean=5.17, SD=1.49, skew=-0.08, kurtosis=-0.87) given that the 95% confidence interval was [0.667,2.2555].[3]:

8.2.4 Effect size for the dependent groups t-test

A simple effect size formulae for a dependent t test is (Equation 7 in Lakens (2013))[4];

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

```
##  the normality and the equal variances assumptions are made
## given the robust procedures provided roughly the same results
n=20
tval=3.6678

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

library(effsize)
cohen.d(wounds$tape,wounds$suture,
        paired=T, conf.level=0.95,noncentral=F)
##
## Cohen's d
##
## d estimate: 0.82 (large)
## 95 percent confidence interval:
##   inf   sup
## 0.135 1.505
```

The effsize package (Torchiano, 2016) reported an effect size of 0.820 and the 95% CI was [0.135, 1.505]

8.2.5 Missing data techniques for the dependent groups t-test

To be added

8.2.6 Supportive graphs for the dependent groups t-test

To be added

[3] The descriptive statistics were calculated with the *psych* package (Revelle, 2016) and the non-directional percentile bootstrap method with 5000 replications was conducted with the base package (R Core Team, 2016b).

[4] it goes to infinity as r goes to 1 even when the means are very similar. Equation 10 in Lakens (2013) is more appropriate which is $\frac{meandifference}{(SD_1+SD_2)/2}$

8.2.7 Power calculations for the dependent groups t-test

Section 7.3.8 provided the basics of statistical power.

```
#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
##
##               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
```

This illustration shows that for the pre-determined knowns of a mean difference of 0.35, a standard deviation of 0.6, an alpha level of 0.05, a non-directional test and a desired power of 0.95, the sample size (number of pairs) should be 41. In other words, the probability of rejecting the null ($H_0 : \mu_1 - \mu_2 = 0$) is .95 with a sample size of 41, a mean difference of 0.35, SD=0.6, alpha=0.05 and a non-directional paired t-test.

8.3 Common Designs

We first present examples of designs commonly used in studies in the social and behavioral sciences to compare two means. The steps used in such studies are

1. obtain scores under each of the two treatments
2. compute the mean for each treatment, and
3. compare the means using a statistical hypothesis test.

An important distinction in selecting a statistical test is whether the scores in the two treatments are correlated or independent. In this handout We classify the designs by whether the scores in the two treatments are correlated or independent. Then We turn to a presentation of terminology for describing designs. This terminology facilitates discussion of designs and determining the correct data analysis procedure to use with a design.

8.3.1 Designs in which Scores in the Two Treatments are Correlated

We want to be able to determine whether the scores used to compute one mean are likely to be correlated with the scores used to compute the second mean. While this goal would seem to require analyzing the data, the surface characteristics of the design used to collect the data can be used to determine whether or not the scores are likely to be correlated.

8.3.1.1 Repeated measures designs

These are designs in which multiple measurements of the same variables are made on the same subjects.

1. **Subjects as own control design:** To examine whether activation of a concept in semantic memory increases accessibility of related concepts, 100 college students were asked to read pairs of words. The first member of each pair was either a weapon word (such as “dagger” or “bullet”) or a non-weapon

word. The second member was always an aggressive word (such as “destroy” or “wound”). On each of 192 trials, a computer presented a priming stimulus word (either a weapon or non-weapon word) for 1.25 seconds, a blank screen for 0.5 seconds, and then the target aggressive word. The experimenter instructed the participants to read the first word to themselves and then to read the second word out loud as quickly as they could. The computer recorded how long it took to read the second word. Average reaction time was computed for each participant under each type of prime word. The data could be recorded in a table like the following

	Prime Word	
Subject	Weapon	Non-weapon
1		
2		
...		
100		

Based on the idea that some participants read more quickly than others, we would expect the reaction times under the two types of prime words to be correlated.

2. **Longitudinal designs:** Mathematics achievement is measured twice for 48 6th grade students: at the beginning of the school year and at the end of the school year. The purpose is to test whether or not the means change over time. The data could be recorded in a table like the following

	Time	
Subject	Beginning	End
1		
2		
...		
48		

Because the same students are measured on each occasion we expect the scores to be correlated over time.

8.3.1.2 Blocking designs

These are designs in which participants are placed in pairs; the members of each pair are expected to perform similarly.

1. **Randomized Block Design:** A study was conducted to examine the effects of metacognitive instruction on reading. Thirty second-grade students were administered a reading test and placed in pairs based on the results.

Pair	Ranks on Reading Pretest
1	1,2
2	3,4
...	...
15	29,30

As shown, the students with the two highest scores were in the first pair, the students with the second highest scores were in the second pair, and so forth. From within each pair one student was randomly assigned to the metacognitive training and one to the control treatment.

Following completion of training the students were tested again on reading. The purpose was to determine whether or not type of training affected mean reading. The data can be recorded in a table like the following

Training		
Pair	Metacognitive	Control
1		
2		
...		
...		
15		

Clearly the scores on the reading pretest will be correlated for pairs of students. However, the scores that are to be analyzed are the scores on the reading posttest. Will these be correlated? Because the students within the first pair have the two highest reading pretest scores, we would expect the student assigned from this pair to the metacognitive treatment to have among the highest scores on the reading posttest; similarly for the student assigned to the control treatment. The students within the last pair have the two lowest reading pretest scores. Therefore we would expect the student assigned from this pair to the metacognitive treatment to have among the lowest scores on the reading posttest; similarly for the student assigned to the control treatment.

The term block is a more general term than pair. It refers to a group of subjects who are homogeneous on some variable. When there are just two treatments a randomized block design (RBD) can be diagrammed as follows:

Treatments		
Block	1	2
1		
2		
...		
n		

Each block is a pair of subjects. One member of the block is exposed to treatment 1 and the other is exposed to treatment 2.

2. **Nonrandomized block design:** A study is conducted to investigate state anxiety levels of physically abused children in a stressful situation. A control group consists of non-abused children matched (matched is a synonym for blocked when each block consists of a pair of subjects) on trait anxiety with the abused children. There were 20 abused children in the study. The data could be recorded in a table like the following:

Type of Child		
Pair	Abused	Control
1		
2		
...		
20		

We expect the state anxiety scores to be correlated because of the matching on trait anxiety.

3. **Familial Designs:** Twenty-five pairs of mothers and adult daughters are surveyed about their political views. The purpose is to test for mean differences between mothers and daughters. The data could be

recorded in a table like the following:

Pair	Type of Person	
	Mother	Daughter
1		
2		
...		
25		

We expect the political views of mothers and daughters to be at least somewhat correlated.

4. **Dyad Designs:** Fifty pairs of African-American and European-American students are formed. The pairs complete a task involving cooperation. Following completion of the task, subject rate the cooperativeness of their partner. The data could be recorded in a table like the following

Ethnic Background		
Pair	African American	European American
1		
2		
...		
25		

We expect the cooperativeness scores for members of a pair to be related.

8.3.2 Designs in which Scores in the Two Treatments are Independent

1. **Completely Randomized Design:** It has been proposed that pain can be treated with magnetic fields. Fifty patients experiencing arthritic pain were recruited. Half of the patients were randomly assigned to be treated with an active magnetic device and half were assigned to be treated with an inactive device. All patients rated their pain after application of the device. The purpose is to determine whether or not type of device affects mean pain ratings. The data can be recorded in a table like the following:

Device	
Magnetic	Inactive
.	
.	
.	

Note that there is no way to pair the scores and that therefore the scores cannot be correlated.

2. **Nonrandomized Design:** Fifty 8th grade boys and 50 8th grade girls take a test on addition of two-digit addition. The test is computer generated and measures the amount of time taken to answer each question. The purpose is to determine whether or not there are gender differences in mean time to respond. Again there is no way to pair the scores and that therefore the scores cannot be correlated.

Chapter 9

Analysis of Variance (ANOVA)

9.1 Terminology

Designs are usually described using a standard terminology. The following is an introduction to this terminology.

Factor a collection of treatments. For example, in the Magnetic vs. Inactive device study, device is a factor. In the priming study, type of prime word is a factor.

Level an instance of a factor. In the Magnetic vs. Inactive device study 8.3.2, magnetic device is a level of the type of instruction factor, as is inactive device. In the priming study, weapon word is a level of the type of prime word factor, as is non-weapon word.

Crossed factors two factors are crossed if each level of one factor occurs in combination with every level of the second factor. For example, consider the diagram, on the following page, of a repeated measures design in which the treatment factor has two levels.

	Levels of Treatment Factor	
	1	2
Subjects		
1		
2		
.		
n		

Subjects can be considered a factor and is crossed with the treatment factor since each subject occurs in combination with each treatment.

Nesting one factor is nested in a second factor if each level of the first factor occurs in combination with only one level of the second factor. For example, consider the following diagram of an independent samples design in which the treatment factor has two levels.

Levels of Treatment Factor	
1	2
S_1	S_{n+1}
S_2	S_{n+2}
S_3	S_{n+3}
...	

Levels of Treatment Factor	
S_n	S_{2n}

Subjects are nested in treatments because each subjects appears in only one treatment.

Within-subjects factor a factor that is crossed with subjects. The name derives from the fact that the levels of the factor vary within a subject as can be seen in the diagram for the repeated measures design. The following designs have a within-subjects factor: subjects as own control and longitudinal, both of which are examples of repeated measures designs.

Within-blocks factor a factor that is crossed with blocks. The name derives from the fact that the levels of the factor vary within a block as can be seen in the following diagram.

Block	Levels of factor	
	1	2
1		
2		
...		
n		

The following designs have a within-blocks factor: randomized block, nonrandomized block, familial, and dyads.

Many people do not distinguish between within-subjects and within-blocks factors, because they lead to the same method of analysis. Typically, we will not distinguish between the two types of factors and will label both as a within-subjects factor.

Between-subjects factor a factor that has subjects nested in its levels; the subjects in the levels are not crossed with blocks. The qualifier following the semi-colon is necessary to distinguish a between-subjects factor from a within-blocks factor because in both factors a subject is assigned to only one level of a factor. This can be seen from the diagram for the independent samples design:

Levels of Treatment Factor	
1	2
S_1	S_{n+1}
S_2	S_{n+2}
S_3	S_{n+3}
...	
S_n	S_{2n}

9.2 Between Subjects ANOVA

The name between-subjects derives from the fact that the levels of the factor vary between subjects.

9.2.1 One-way Between Subjects ANOVA

The structural model for a one-factor between subjects ANOVA is $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$, in which Y_{ij} is the score for the participant i in group j, μ is the grand mean of the scores, α_j is the effect of the level j, and ϵ_{ij}

is the error term (nuisance). It can be shown that $\mu_j = \mu + \alpha_j$, where μ_j is the the mean for the jth level of the factor.

Generally, the interest is on α_j because it represents $\mu_j - \mu$. This interest leads to hypothesis testing: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_j$

The alternative hypothesis states that at least one population mean is different. It is possible to test the null by partitioning the variance, for a one factor model using the notation by Myers et al. (2013)

SV	df	SS	MS	F
Total	$N-1$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{..})^2$		
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	$N-J$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2$	$SS_{S/A}/df_{S/A}$	

SV	EMS
Total	
A	$\sigma^2_{S/A} + \frac{1}{J-1}\sum_j n_j(\mu_j-\mu)^2$
S/A	$\sigma^2_{S/A}$

where SV=Source of Variance, df=degrees of freedom, SS=Sum of squares, MS= Mean Square, EMS= Expected Mean Square, A is the between subjects factor with J levels, S/A is the subjects within A, N is the total sample size, j=1,...,J factor level indicator, i=1,...,n_j individual indicator, Y_{ij} is the individual score, $\bar{Y}_{..}$ is the grand mean, $\bar{Y}_{.j}$ is the group j's mean.

The ratio of $MS_A/MS_{S/A}$, when the null is true and assumptions are met, follows an F distribution with J-1 and N-J degrees of freedom; hence, if $MS_A/MS_{S/A}$ is larger than the $F_{\alpha,J-1,N-J}$ the null is rejected.

9.2.1.1 Effect size for one-way between-subjects ANOVA

To simplify the illustration, let us assume each treatment level has the same number of participants, $n_1 = n_2 = \cdots = n_j = n$. Hence, the expected mean square for A is $\sigma^2 + n\theta^2_A$ in which

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

.

The estimate of θ^2_A, the $\hat{\theta}^2_A$ is equal to $\frac{MS_A - MS_{S/A}}{n}$, and the estimate of $\sigma^2_{S/A}$, the $\hat{\sigma}_{S/A}$ is equal to $MS_{S/A}$

As stated in Section 8.1.4 to judge whether a mean difference is large in a substantive sense one can use an effect size. For a one-way between subjects ANOVA, reporting at least one type of effect size is a general practice. Among them, omega-hat-squared ($\hat{\omega}^2$), eta-hat-squared ($\hat{\eta}^2$) and f are well known.

9.2.1.1.1 Omega-squared for one-way between-subjects ANOVA

Omega-hat-squared is the proportion of total variance that is due to the factor. $\hat{\omega}^2 = \frac{(J-1)\hat{\theta}^2/J}{((J-1)\hat{\theta}^2/J)+\hat{\sigma}^2_{S/A})}$

An omega-squared is considered small if it is 0.01, medium if 0.06, large if 0.14 Myers et al. (2013).

9.2.1.1.2 Eta-squared for one-way between-subjects ANOVA

$\hat{\eta}^2 = \frac{SS_A}{SS_{Total}}$ also attempts to estimate the proportion of total variance that is due to the factor.

$\hat{\eta}^2$ is larger than $\hat{\omega}^2$ because $\hat{\eta}^2$ is a positively biased statistics, that is, it tends to be too large, especially when n is small.

$\hat{\eta}^2$ is probably the most widely used effect size for ANOVA and also reported in a regression fashion as R^2.

9.2.1.1.3 Effect size f for one-way between-subjects ANOVA

Cohen's $f = \frac{\hat{\theta}_A}{\hat{\sigma}_{S/A}}$. An f value is considered small if it is 0.10, medium if 0.25, large if 0.40.

9.2.1.1.4 A general note on the Effect Size Measures

For illustrative purposes, we briefly summarized effect size measures for equal sample size in each group. In practice it is generally not common to have equal sample sizes. It is also not common to have a single factor design. In addition, factors in a design are either measured or manipulated, which which affect the effect size computation. The *ezANOVA* function (Lawrence (2016)) reports generalized eta-squares based on Bakeman (2005). The work by Bakeman (2005) encourages researchers to use generalized eta-squared defined by Olejnik and Algina (2003). Hence, a convenient choice for a researcher is to use the *ezANOVA* function, while paying attention to the *observed* argument to declare the measured factors. On the other hand, if it is not desired to be dependent on an R package, the researcher can examine and apply the formulae by Olejnik and Algina (2003).

9.2.1.2 Testing specific contrasts of means

Either in addition to or in place of the ANOVA, specific contrasts (comparisons) of means may be tested. A contrast is a weighted sum of means in which the weights sum to zero. There are two classes of contrasts: pairwise contrasts and complex contrasts. To illustrate these classes consider a one-way design in which the factor has three levels, a control treatment and two active treatments. Let the population means for these levels be μ_1, μ_2, and μ_3, respectively. In a pairwise contrast two means are compared and the weights are 1 for one mean, -1 for another and zero for all others. A pairwise contrast of the means for the active treatments is $(0)\mu_1 + (1)\mu_2 + (-1)\mu_3$. The complex contrast $(-1)\mu_1 + (-.5)\mu_2 + (-.5)\mu_3$ is a comparison of the mean for the control group to the average of the means for the two active treatments. Under the assumptions of a one-way between-subjects ANOVA, the null hypothesis that a contrast is equal to zero can be tested using

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

9.2.1.3 Testing all possible pairwise comparisons

There are several procedures for testing all possible pairwise contrasts. An important issue in such testing is the error rate to control. Controlling an error rate means keeping it at or below some conventional level (e.g., .05). Two of the most common error rates are the per comparison error rate and the familywise error rate. The per comparison error rate is the probability of making a Type I error when one of the contrasts is tested. To control the per comparison error rate the critical value for a pairwise comparison is $\pm t_{(1-\alpha 2),N-J}$. When this critical value is used, the per comparison error rate is α. The family wise error rate is the probability of falsely rejecting one of more of the contrasts. If all pairwise contrasts are equal to zero, the family wise error rate is between α and $[J(J-1)2]\alpha$. The upper limit can be quite high even when the number of levels of the factor is small. For example if there are J=3 levels, the upper limit is 3α. There are several procedures for controlling the familywise error rate.

9.2.1.3.1 Trend analyses following one-way between-subjects ANOVA

To be added.

9.2.1.4 Assumptions of the one-way between-subjects ANOVA

The assumptions of the one-way between-subjects ANOVA are the same as the assumptions of the independent samples t test.

1. Independence. The scores in each group should be independently distributed and the scores in different groups should also be independent. The validity of this assumption in regard to independence within groups is questionable when (a) scores for participants within a group are collected over time or (b) the participants within a group work together in a manner such that a participant's response could have been influenced by another participant in the study. The validity of this assumption in regard to independence between groups is questionable when the factor is a within-subjects factor rather than a between-subjects factor. Violating the independence assumption is a critical violation that usually can be addresses by adopting an analysis appropriate for the lack of independence. For example, if there are different participants in each group, but within each group there are subgroups of participants who work together then according to (b) above independence is likely to have been violated. This violation can be addressed by using multilevel analysis. If there are different participants in each group, but the participants in the groups have been matched, using a randomized block ANOVA can address the violation of independence.

2. Normality. The scores with each group are drawn from a normal distribution. Statistical power is likely to be compromised if the distributions of scores have long tails. When the sample sizes are equal violating normality is not likely to affect the type I error rate, unless the non-normality is severe and the sample sizes are small.

3. Equal variance. This assumption is also called the homogeneity of variance assumption and means it is assumed that samples in the J groups are drawn from J populations with equal variances. Violation of the equal variance assumption is likely to affect the Type I error rate except when the sample sizes are equal and fairly large.

Even though we briefly summarized the assumptions of the one-way between subjects ANOVA above, they were only introductory. If independence does not appear to be violated, then when the sample sizes are equal and at least 20 in each group and the scores are approximately normally distributed the one-way between subjects ANOVA can be used. In other situations alternatives should be used. When the robust analyses (e. g. Wilcox (2012)) and conventional analyses yield the same decisions about **all hypothesis tests**, results of the conventional analyses can be reported due to their greater familiarity to most readers.

9.2.1.5 R codes for a one-way between-subjects ANOVA

For illustrative purposes, the city of KOCAELI is subsetted from the DataWBT (Section 2.3). The gender attitudes scores is the dependent variable and the highest degree completed is the between subjects factor. This factor had seven levels; no-degree, primary school, middle school, high school,vocational high school, 2 year college and bachelors. However, there is only one participant in the *no-degree* group. We combined the no-degree and primary school groups. The gender attitude score for this participant is 1.6. [1]

For illustrative purposes, the city of KOCAELI is subsetted from the DataWBT (Section 2.3). The gender attitudes scores is the dependent variable and the highest degree completed is the between subjects factor.

[1] Removing this participant from the ANOVA would have had no substantial effect on the results.

This factor had seven levels; no-degree, primary school, middle school, high school,vocational high school, 2 year college and bachelors. However, there is only one participant in the *no-degree* group. We relocated this individual into primary school level. The gender attitude score for this participant is 1.6. Removing this participant from the ANOVA has no substantial effect on the results.

Step 1: Set up data and report descriptive

```r
# load csv from an online repository
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#remove URL
rm(urlfile)

#select the city of KOCAELI
# listwise deletion for gen_att and education variables
dataWBT_KOCAELI=na.omit(dataWBT[dataWBT$city=="KOCAELI",
                                c("id","gen_att","education")])


#There is only 1 participant in the level "None", merge it into Primary school
# the gender attitude score for this participant is 1.6
library(car)
dataWBT_KOCAELI$eduNEW <- recode(dataWBT_KOCAELI$education,
                                 "'None'='Primary School (5 years)'")

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)'")
#table(dataWBT_KOCAELI$eduNEW)

##optional re-order levels (cosmetic)
#levels(dataWBT_KOCAELI$eduNEW)
dataWBT_KOCAELI$eduNEW = factor(dataWBT_KOCAELI$eduNEW,
                         levels(dataWBT_KOCAELI$eduNEW)[c(4,3,1,6,2,5)])


#which(dataWBT_KOCAELI$education=="None")

#drop empty levels
dataWBT_KOCAELI$eduNEW=droplevels(dataWBT_KOCAELI$eduNEW)


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

```
#select relevant descriptives
# 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
#write.csv(desc1BW,file="onewayB_ANOVA_desc.csv")
```

Step 2: Check assumptions

```
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 = "Gender Attitude by Degree in Kocaeli")+ facet_wrap(~ eduNEW)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="bold"))
```

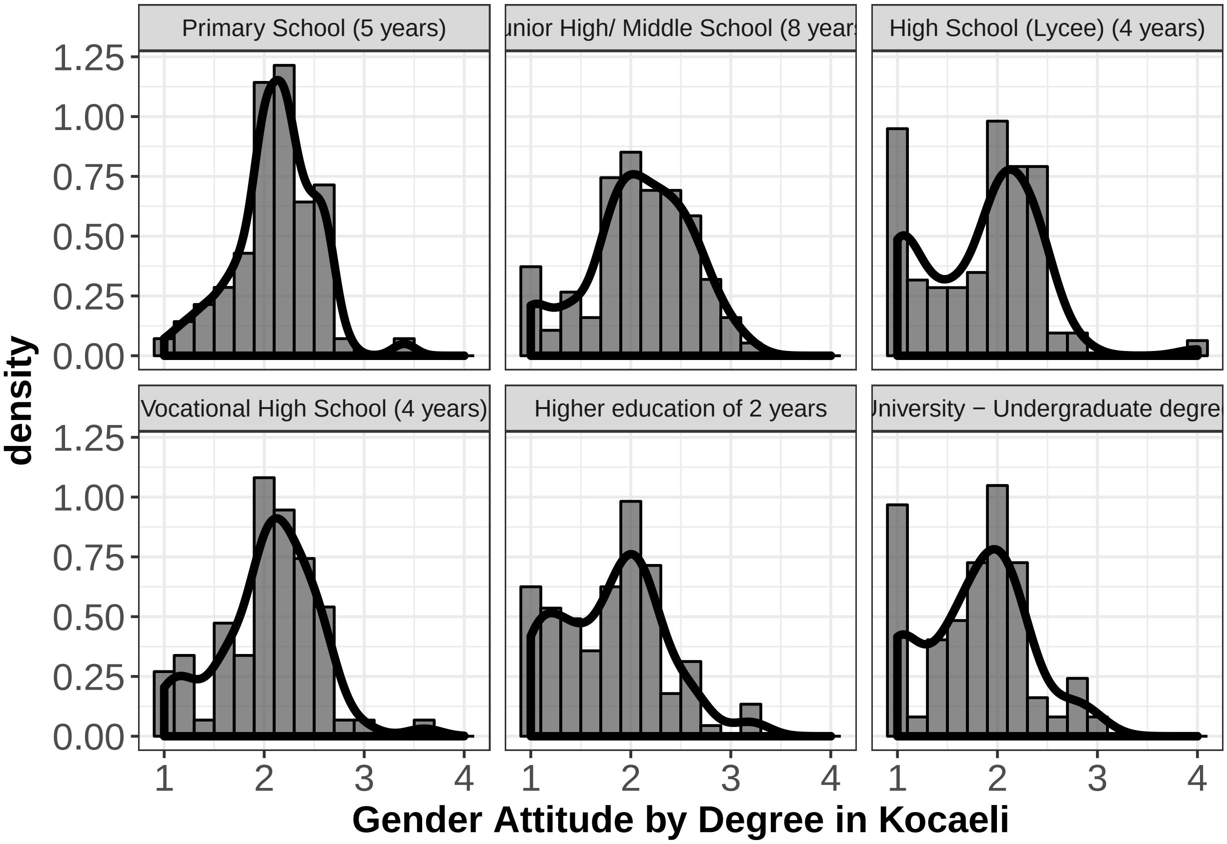


Figure 9.1: Gender Attitudes by Degree

Departures from the normality do not seem to be severe.

```
require(ggplot2)
ggplot(dataWBT_KOCAELI, aes(eduNEW,gen_att)) +
  geom_boxplot() +
  labs(x = "Education",y="Gender Attitude by degree in Kocaeli")+coord_flip()
```

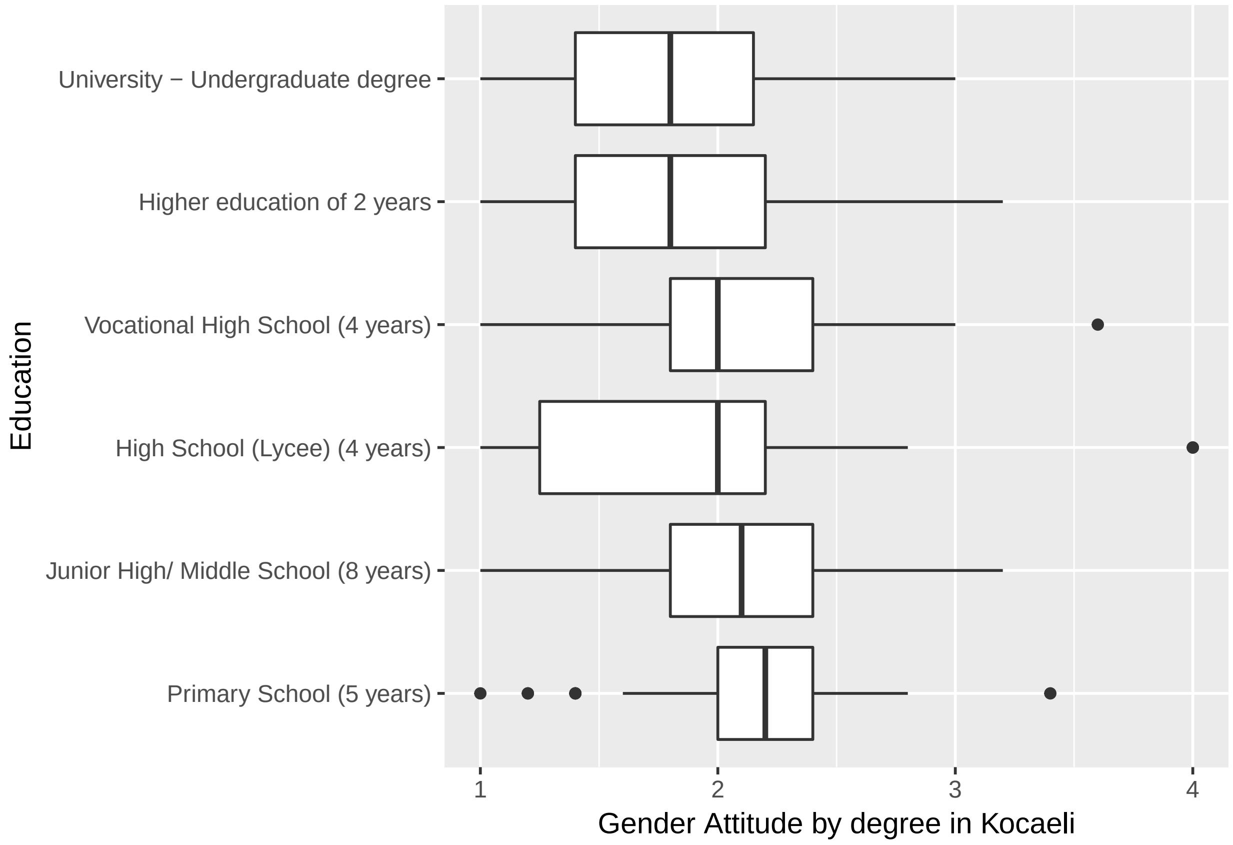


Figure 9.2: Gender Attitudes by Degree

Homogeneity of variance is questionable but not severely violated.

Step 3: Run ANOVA

For illustrative purposes, let us ignore the violations first. The *ezANOVA* function (Lawrence (2016)) reports the F test, the Levene Test and an effect size. Type of the effect size depends on the model. For further details, please carefully study the Table 1 in Bakeman (2005), an open access article, or Olejnik and Algina (2003). The Levene test rejects the null hypothesis of equal variances across factor levels.

```
library(ez)
#the ezANOVA function throws a warning if id is not a factor

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

# set the number of decimals (cosmetic)
options(digits = 3)


#alternative 1 the ezANOVA function
```

```
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     *



# critical F value
qf(.95,5,564)
## [1] 2.23
```

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

ezANOVA can calculate three different types of sums of squares
for main effects and interactions.
For a one-way between-subjects design the F test is the same
for all three types and this warning can be ignored.
```

The same results can be obtained with the *lm* (linear model) function in R Core Team (2016b).

```
# alternative 2 the lm function
alternative2=lm(gen_att~eduNEW,data=dataWBT_KOCAELI)

#Table 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
```

The *aov* function in R Core Team (2016b) is the third alternative.

```
#alternative 3 the aov function
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
```

The *pairwise.t.test* function in the *stats* package (R Core Team (2016b)) is convenient. Provide the preferred procedure by using *p.adjust.method* argument, for example *p.adjust.method* =*"Holm"* to use the adjustment given by Holm (1979). Five other procedures are available with this function, please see *?p.adjust*.

```
# pairwise comparisons
# Table 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 Robust estimation and hypothesis testing for a one-way between-subjects design

Several approaches to conducting a robust one-way between subjects ANOVA, have been presented by Wilcox (2012) One of the convenient robust procedure , a heteroscedastic one-way ANOVA for trimmed means, has been compressed into the *t1way* function, available via WRS-2 (Mair and Wilcox (2016)). Please use *?t1way* for the current details, this promising package is being improved frequently .

```
library(WRS2)

#t1way
# 20% trimmed
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% trimmed
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% trimmed
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 pairwise comparisons

#level order
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)"
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 Example writeup for one-way between-subjects ANOVA

For our illustrative example, results of hypothesis tests conducted using robust procedures did not disagree with the results of the ANOVA and pairwise comparisons of means. This was expected given the assumptions were not severely violated. When the robust analyses and conventional analyses yield the same decisions about *all hypothesis tests*, results of the conventional analyses can be reported due to their greater familiarity to most readers. A possible write up for our illustrative example would be:

An ANOVA was performed to investigate whether the gender attitudes scores differ across education level. The means,standard deviations, skewness and kurtosis values of the gender scores, grouped by the highest-degree obtained, are presented in Table 1. The analysis of variance indicated a significant difference in the gender attitudes scores , F(5,564) = 7.27, p < .001, $\hat{\eta}^2_G = .06$. Table 2 is the ANOVA table for this analysis. Pairwise comparisons were planned a' priori. The familywise error rate was selected for control and the Holm procedure (Holm (1979)) was used. The results of the pairwise comparisons are presented in the Table 3. Nine out of fifteen comparisons yielded statistically significant results; (primary school vs lycee, primary school vs 2-year-collage, primary school vs undergraduate,… (provide details). Robust statistical procedures yielded the same conclusions.

9.2.1.8 Missing data techniques for one-way between-subjects ANOVA

To be added

9.2.1.9 Power calculations for one-way between-subjects ANOVA

To be added

9.2.2 Two-Factor Between Subjects ANOVA

This topic concerns designs in which there are two between-subjects factors: factor A with J levels and factor B with K levels for a total of JK combinations of levels; each combination is called a cell. The factors are between-subjects if (a) different subjects appear in each cell and (b) subjects are not matched in any way. In the simplest version of this design, each factor has two levels. For example, consider if a researcher is interested in the effect of the gender and college education on the gender attitudes scores. The following is a depiction of a study designed to investigate these two factors.

	non-college	college	
Female	μ_{11}	μ_{12}	$\mu_{1\cdot}$
Male	μ_{21}	μ_{22}	$\mu_{2\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	

Also shown are the parameters about which hypotheses will be tested: the population cell means $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$, row means ($\mu_{1\cdot}$, $\mu_{2\cdot}$) and column means ($\mu_{\cdot 1}$, $\mu_{\cdot 2}$). The general term for a row or column mean is a marginal mean.

Symbolically the hypothesis of no interaction can be written as $H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$. The interation is also a comparison of the two simple effects of gender $(\mu_{21}\mu_{11}and\mu_{22}\mu_{12})$ leading to the null hypothesis $H_0 : \mu_{21}\mu_{11} = \mu_{22} - \mu_{12}$. If one of the null hypotheses is true the other must also be true and if one is false the other must also be false.

Interaction The first null hypothesis of interest is the hypothesis of no interaction between the two factors. Before defining an interaction, we first define a simple main effect. A simple main effect refers to differences among the cell means in a particular row or in a particular column. In the current example, there are two types of simple main effects: simple main effects of gender and simple main effects of college education.

For each level of graduation from college type there are two simple main effects gender: the simple main effect of gender at college graduates (μ_{12} versus μ_{22}) and the simple main effect of gender at non-college graduates (μ_{11} versus μ_{21}). The interaction is a comparison of these two effects.

For each type there are two simple main effects. There is a simple main effect of gender at college graduates (μ_{12} versus μ_{22}) and a simple main effect of gender at non-college graduates (μ_{11} versus μ_{21}).

There is a simple main effect of education for Female (μ_{11} versus μ_{12}) and a simple main effect of education for Male (μ_{21} versus μ_{22}).

The main effects Effects defined in terms of marginal (row and column) means are called main effects. Symbolically, the main effect of gender is $\mu_{1\cdot} - \mu_{2\cdot}$, and the hypothesis of no main effect due to gender is $H_0 : \mu_{1\cdot} - \mu_{2\cdot} = 0$. Similarly, the hypothesis of no main effect due to college education is $H_0 : \mu_{\cdot 1} - \mu_{\cdot 2} = 0$.

When there is an interaction:

1. Inspection of the main effect for a factor is misleading when the directions of the simple main effects of the factor are not the same at all levels of the second factor.

2. It is a matter of opinion as to whether it is misleading to inspect the main effect for a factor the directions of the simple main effects of the factor are the same at all levels of the second factor.

When the main effect is misleading about the effect of a factor, the cell means are the proper basis for studying the effects of the factor.

The structural model for a two-factor between subjects ANOVA is $Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \epsilon_{ij}$, in which Y_{ijk} is the score for the participant i in first factor level j , and the second factor level k; μ is the grand mean of the scores, α_j is the effect of the level j of the first factor, β_k is the effect of the level k of the second factor, $\alpha\beta_{jk}$ is the interaction effect and ϵ_{ij} is the error term (nuisance).

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	$N - JK$	

SV	df	F
Total	$N - 1$	

9.2.2.1 Type I, II and III sum of squares

As we pointed out in the section on one-way between-subjects designs, the F test of the main effect is the same for all three types of sums of squares. This is not true in designs with two or more between-subjects factors. In designs with three or more between-subjects of effects F tests for interaction other than the highest order interaction can vary across the types of sums of squares. Selecting among the three types can be an important decision and we refer the reader to Carlson and Timm (1974) for a discussion of the issues in selecting among the three types of sums of squares in experimental studies and to Appelbaum and Cramer (1976) for a discussion of the issues in survey studies.

9.2.2.2 R codes for a two-way between-subjects ANOVA

For illustrative purposes, the city of Kayseri is subsetted from the DataWBT (Section 2.3). The gender attitudes scores is the dependent variable, gender and higher education indicator are the between subjects factors.

Step 1: Set up data and report descriptive

```
# load csv from an online repository
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#remove URL
rm(urlfile)

#select the city of Izmir
# listwise deletion for gen_att and education variables
dataWBT_Kayseri=na.omit(dataWBT[dataWBT$city=="KAYSERI",c("id","gen_att","higher_ed","gender")])

# Higher education is coded as 0 and 1, change it to non-college, college
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)

#drop empty levels
dataWBT_Kayseri$gender=droplevels(dataWBT_Kayseri$gender)

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


# set the number of decimals (cosmetic)
```

```
options(digits = 3)

#get descriptives
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)) } ))
# Table 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_desc.csv")
```

Step 2: Inspect assumptions

```
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 = "Gender Attitudes by HEF and Gender in Kayseri")+ facet_wrap(~ HEF+gender)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="bold"))
```

Departures from the normality do not seem to be severe.

```
require(ggplot2)
ggplot(dataWBT_Kayseri, aes(x=gender, y=gen_att))+
  geom_boxplot()+
  facet_grid(.~HEF)+
  labs(x = "Gender",y="Gender Attitude by Gender and HEF in Kayseri")
```

Variances look similar.

Step 3: Run ANOVA

The *ezANOVA* function (Lawrence (2016)) reports the F test, the Levene Test and an effect size. Type of the effect size depends on the model and indirectly depends on the type of sum of squares used. The *type* argument (1,2 or 3) transmits the choice.

```
library(ez)
#the ezANOVA function throws a warning if id is not a factor

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

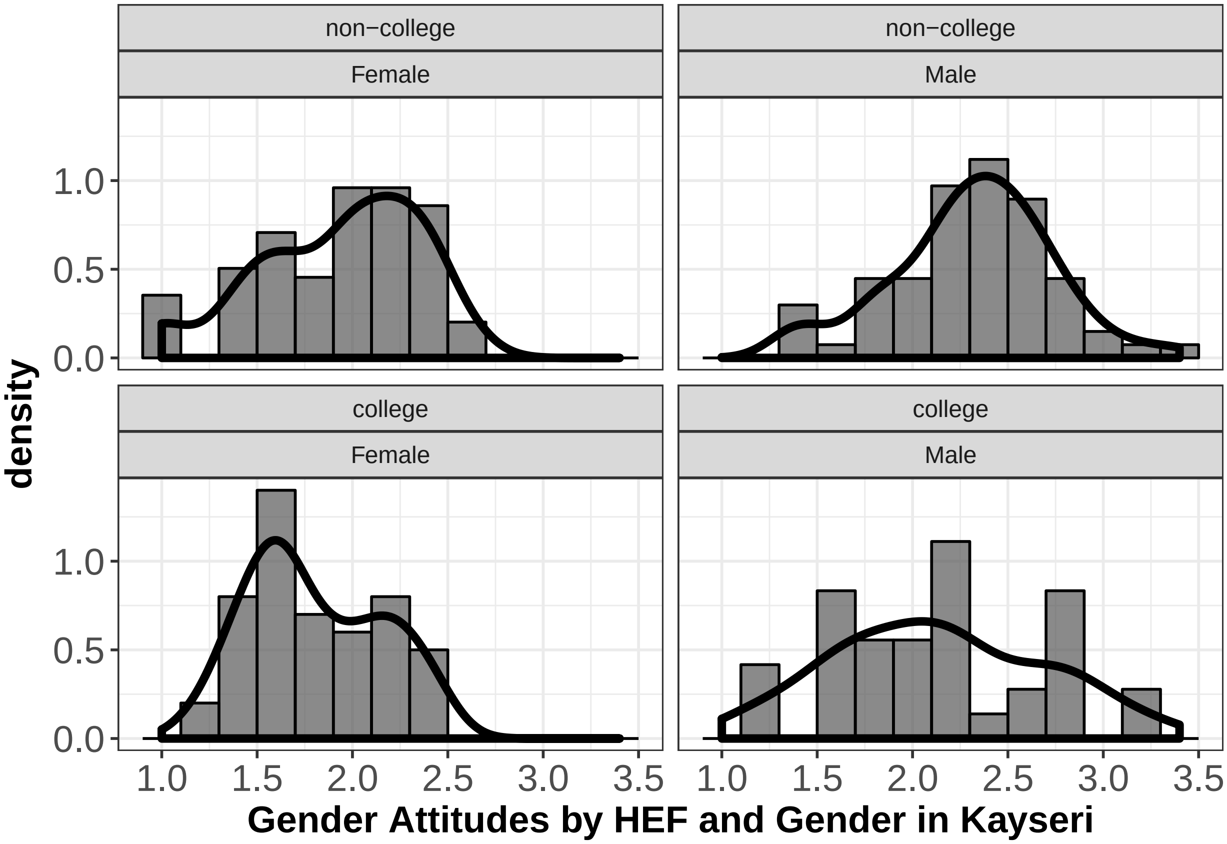


Figure 9.3: Gender Attitudes by HEF and Gender

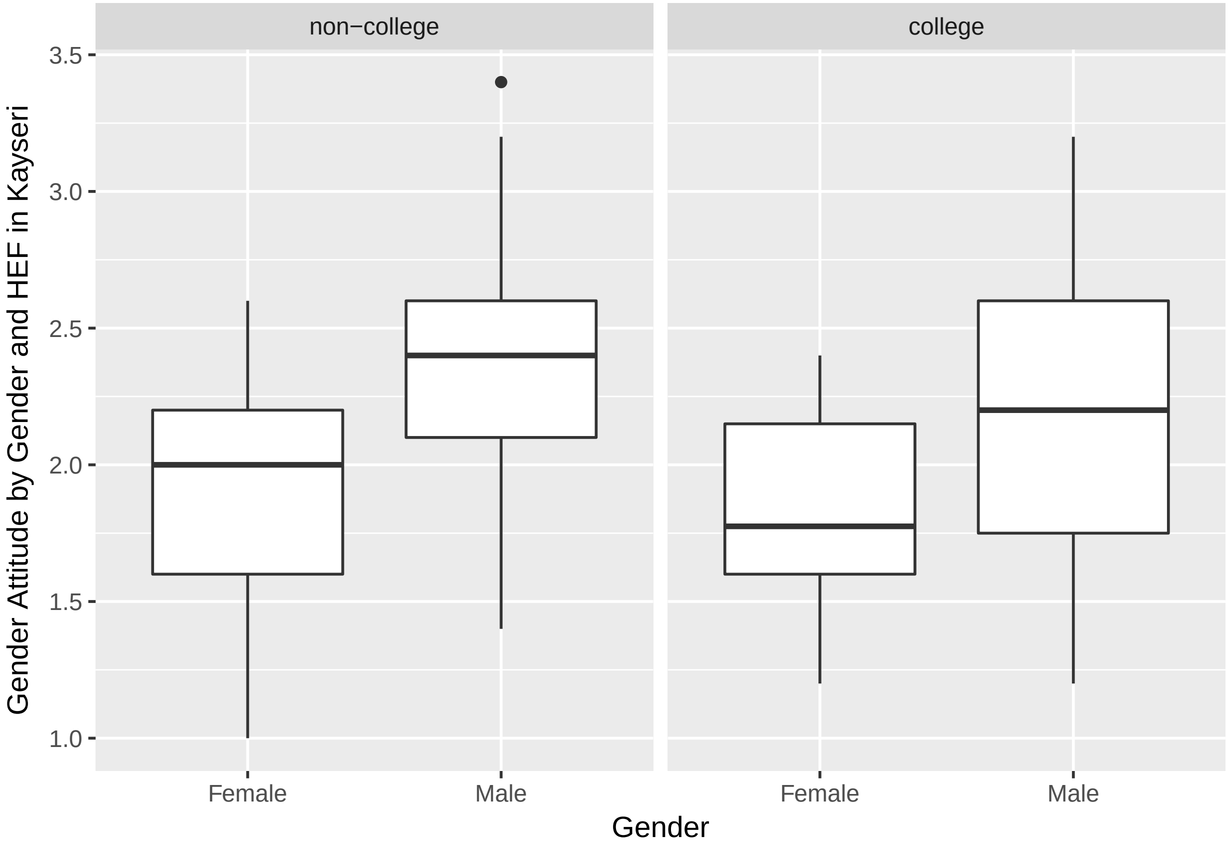


Figure 9.4: Gender Attitudes by Degree

```
#alternative 1 the ezANOVA function
alternative1 = ezANOVA(
    data = dataWBT_Kayseri,
    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


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

# critical F value
qf(.95,1,248)
## [1] 3.88
```

9.2.2.3 Robust estimation and hypothesis testing for a two-way between-subjects design

Several approaches to conducting a robust two-way between subjects ANOVA, have been presented by Wilcox (2012) One of the convenient robust procedure , a heteroscedastic two-way ANOVA for trimmed means, has been compressed into the *t2way* function, available via WRS-2 (Mair and Wilcox (2016)). Please use *?t2way* for the current details, this promising package is being improved frequently .

```
library(WRS2)

#t2way
# 20% trimmed
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% trimmed
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% trimmed
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
```

9.2.2.4 Example writeup two-way between-subjects ANOVA

For our illustrative example, robust procedures did not disagree with our initial analyses. This was expected given the assumptions were not severely violated. When the robust analyses yield very similar results, we prefer to report initial results to ease communication. A possible write up for our illustrative example would be:

Descriptive statistics for the gender attitudes scores as a function of gender and higher education in the city of Kayseri are presented in Table 4. A 2x2 ANOVA was reported. F tests were conducted at $\alpha = .05$. There was a significant difference for gender $F(1,248) = 45.39, p < .001$. There was also a significant difference for the college effect $F(1,248) = 6.24, p = .013$. However, there was no significant interaction between the gender and higher education status, $F(1,248) = 0.25, p = .617$. The *ezANOVA* (Lawrence (2016)) function reported generalized eta hat squared ($\hat{\eta}_G^2$) of 0.15 for the gender effect and 0.02 for the college-effect. Table 5 is the ANOVA table for these analyses.

9.2.2.5 Follow ups for two-way between-subjects ANOVA

To be added.

9.2.2.5.1 Pairwise comparisons following two-way between-subjects ANOVA

To be added.

9.2.2.5.2 Contrats comparisons following two-way between-subjects ANOVA

To be added.

9.2.2.6 Missing data techniques for two-way between-subjects ANOVA

To be added

9.2.2.7 Power calculations for two-way between-subjects ANOVA

To be added

9.3 Within Subjects ANOVA

This procedure can be used when there are score on the same participants under several treatments or at several time points and is then called repeated measures ANOVA. It can also be used in blocking designs and is then called randomized block ANOVA. Compared to between-subjects designs, this procedure is expected to eliminate influence of individual differences, in other words, to reduced variability, and thus, to reduce error. This development results in more power than independent-samples ANOVA with the same sample size. Of course there are issues other than power that must be considered in selecting a design. For example, if the goal is to compare reading achievement under three instructional methods, using a design in which each participant is exposed to the three methods will be problematic because the effect of exposure to one treatment will continue to influence reading ability during the other treatments.

9.3.1 One-way Within-Subjects ANOVA

The structural model following Myers et al. (2013) notation for a non-additive model;

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

where i represents the individual, i=1,...,n; j represents the levels of the within-subjects factor (i.e, the repeated measurement or the treatment factor), j=1,...,P. Y is the score; μ is the grand mean; η_i represents the difference between individual's average score over the levels and the grand mean; α_j represents the difference between the average score under level j of the within-subjects factor and the grand mean; $(\eta\alpha)_i j$ represents the interaction, and ϵ_{ij} represent the error component. Because $(\eta\alpha)_i j$ and ϵ_{ij} have the same subscripts, they are confounded.

Generally, the interest is on α_j. This interest leads to hypothesis testing:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_P$$

The alternative hypothesis states that at least one population mean is different. The ANOVA table for a one-way with-subjects ANOVA is;

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

Note on additivity The simplest explanation of additivity is the situation that would justify $(\eta\alpha)_i j = 0$ in Equation (9.1). This unrealistic restriction implies that the effect of levels of the within-subject factor waves is the same for all individuals.

Shown below in Table 9.10 are data for an experiment in which each person participates in four treatments defined by the amount of alcohol consumed. The dependent variable is a reaction time measure.

Treatment means, treatment standard deviations, and subject means are shown below. A subject mean is

Table 9.10: Original Alcohol Data

id	noALC	twoOZ	fouroz	sixoz
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 9.11: Correlation Coefficients for Reaction Time Data

	noALC	twoOZ	fouroz	sixoz
noALC	1.00	0.93	0.88	0.88
twoOZ	0.93	1.00	0.89	0.94
fouroz	0.88	0.89	1.00	0.95
sixoz	0.88	0.94	0.95	1.00

the average of the four scores for that subject.

```
# set the number of decimals (cosmetic)
options(digits = 2)

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

#treatment mean
apply(owadata[,-1],2, mean)
##  noALC  twoOZ fouroz  sixoz
##    2.8    3.5    6.2    9.0

#treatment sd
apply(owadata[,-1],2,sd)
##  noALC  twoOZ fouroz  sixoz
##   1.49   0.93   1.04   1.31
```

Because each participant has a score for each treatment, amount of alcohol is a within-subjects factor and it is possible to calculate a correlation for each pair of treatments. These correlations, which are presented in Table 9.2, indicate that reaction time is highly correlated for each pair of treatments. Recall that corresponding to each correlation there is a covariance;

$$Cov_{pp`} = S_pS_{p`}r_{pp`}$$

where p and p' are two levels of the alcohol consumption factor. For example the correlation between the scores in the first two levels of the alcohol consumption factor is $r_{02} = 0.93$. And the corresponding covariance is $Cov_{02} = 1.5 * 0.9 * 0.93 = 1.26$

The alcohol consumption factor is a within-subjects factor. Consequently the F statistic for comparing the four treatment means should be appropriate for a design with a within-subjects factor. Let

P = the number of levels of the within-subjects factor, in the example P=4 ;

$\bar{C}$ = the average covariance; in the example $\bar{C} = 1.26$.

The appropriate F statistic is

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

where MS_A and $MS_{S/A}$ are calculated as they are for a between-subjects factor and the W emphasizes the F statistic is for a within-subjects factor. The critical value is $F_{\alpha,P-1,(P-1)(n-1)}$. The denominator mean square, MS_{SA} , is read mean square Subjects x A where A is the generic label for the treatment factor. Recall that the F statistic for a between-subjects factor is $F_B = MS_A/MS_{S/A}$. Comparison of F_W and F_B shows that F_W incorporates the correlations between the pairs of treatments and F_B does not. (Recall that, in like fashion, the dependent samples t statistic incorporates the correlation whereas the independent samples t statistic does not.) As a result, when applied to a design with a within-subjects factor $F_W \geq F_B$.Therefore, incorrectly using will usually result in a loss of power.

9.3.1.1 Assumptions of one-way within-subjects ANOVA

Sphericity is an assumption about the pattern of variances and covariances.If the data are spherical, the difference between each pair of repeated measures has the same variance for all pairs.

Example covariance matrix;

	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

Sphericity holds;

$Y_p - Y_{p'}$	$\sigma_p^2 + \sigma_{p'}^2 - 2\sigma_{pp'}$
$Y_1 - Y_2$	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's epsilon —measures how severely sphericity is violated.

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

Estimates of ϵ are Greenhouse-Geisser ($\hat{\epsilon}$) and Huynh-Feldt ($\tilde{\epsilon}$)

$\tilde{\epsilon}$ can be larger than 1.0; if it is $\tilde{\epsilon}$ is set equal to 1.0.

Critical value assuming sphericity is $F_{alpha,(P-1),(n-1)(P-1)}$.

Approximately correct critical value when sphericity is violated $F_{alpha,\epsilon(P-1),\epsilon(n-1)(P-1)}$.

normality of errors in Equation (9.1), ϵ_{ij} is assumed to be normally and independently distributed with a mean value of zero.

normality of η_i in Equation (9.1), η_i is assumed to be normally and independently distributed with a mean value of zero.

Together the assumptions listed immediately above imply that the repeated measures are drawn from a multivariate normal distribution.

9.3.1.1.1 The relationship between additivity and sphericity

Although assumptions can be stated in terms of η_i and ϵ_{ij}, a simpler approach is that the repeated measures are drawn from a multivariate normal distribution with covariance matrix that meets the sphericity assumption. If the data meet the sphericity assumption, the difference between each pair of repeated measures has the same variance for all pairs.

Assuming that the data are drawn from a multivariate normal distribution and within each level of the within-subjects factor the scores are independent, having equal variance and covariances (compound symmetry) is a sufficient condition for the F test on the within-subjects factor to be valid.

If additivity holds and the equal variance assumption holds then compound symmetry holds. But compund symmetry is a stricter assumption than the sphericity. Considering that sphericity is a necessary and sufficient requirement for the F test on the within-subjects factor to be valid (assuming that data are drawn from a multivariate normal distribution and scores are independent within each level of the within-subjects factor the) checking for spericity is more important than checking for additivity. In addition because there are adjusted degrees of freedom procedures that adjust the F test on the within-subjects factor for violation of sphericity, it is not necessary to test for sphericity.

9.3.1.2 R codes for a one-way within-subjects ANOVA

For illustrative purposes, a subsample from an original cluster randomized trial study (for details see Daunic et al. (2012)) was taken. The subsample included 1 control-classroom and 17 students. The variable of interest is the problem solving knowledge. Each wave was approximately one year apart. Higher scores mean higher knowledge.

Step 1: Set up data

```
#enter data
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))
```

Report descriptive

```
# set the number of decimals (cosmetic)
options(digits = 3)

##the long format will be needed
#head(PSdata)
library(tidyr)
data_long = gather(PSdata, wave, PrbSol, wave1:wave3, factor_key=TRUE)

#get descriptives
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)) } ))
# Table 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")
```

The covariance matrix might be helpful.

```
# check covariance  Table 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
```

Step 2: Check assumptions

```
ggplot(data_long, aes(x=wave, y=PrbSol))+
  geom_boxplot()+
  labs(x = "Wave",y="Problem Solving Knowledge scores")
```

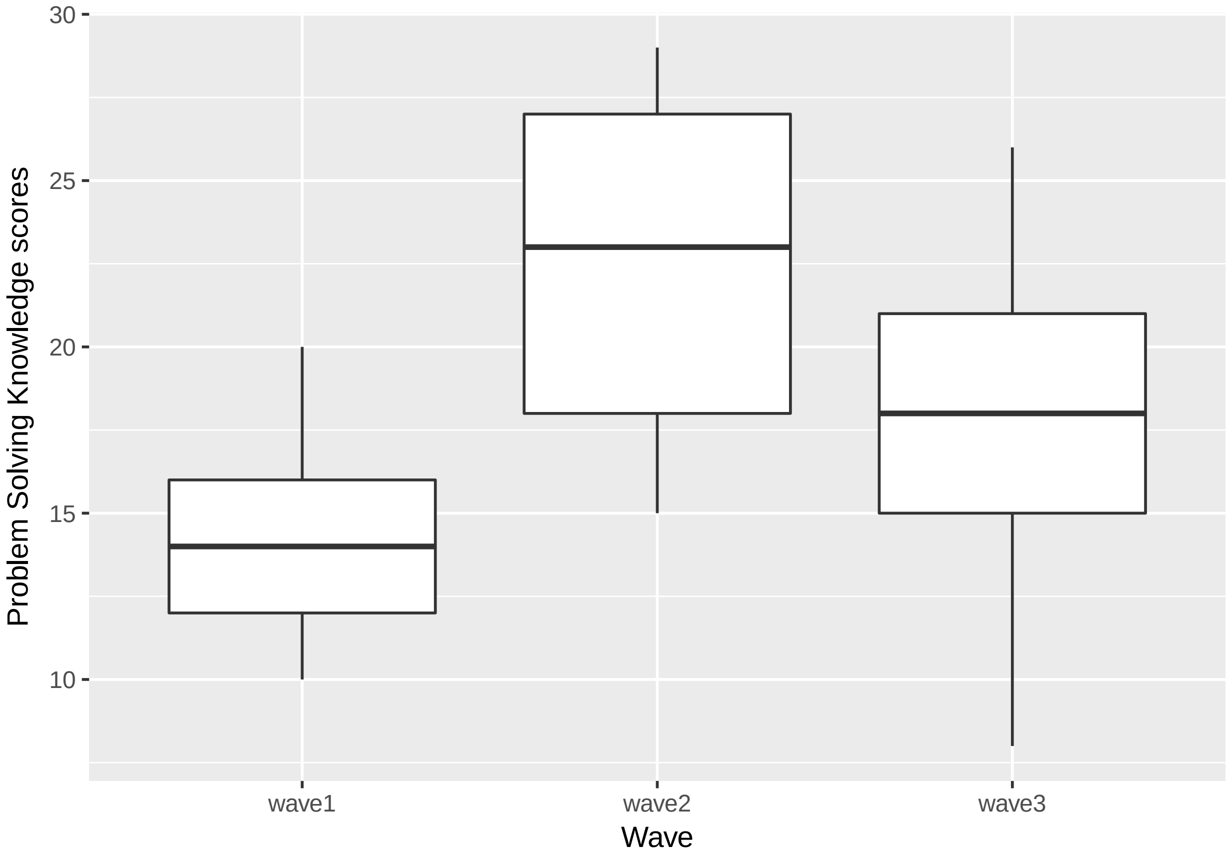


Figure 9.5: Problem Solving Knowledge score by wave

We will test for the sphericity assumption using Mauchly's test integrated in the *ezANOVA* function, but this graph implies that it might be violated.

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

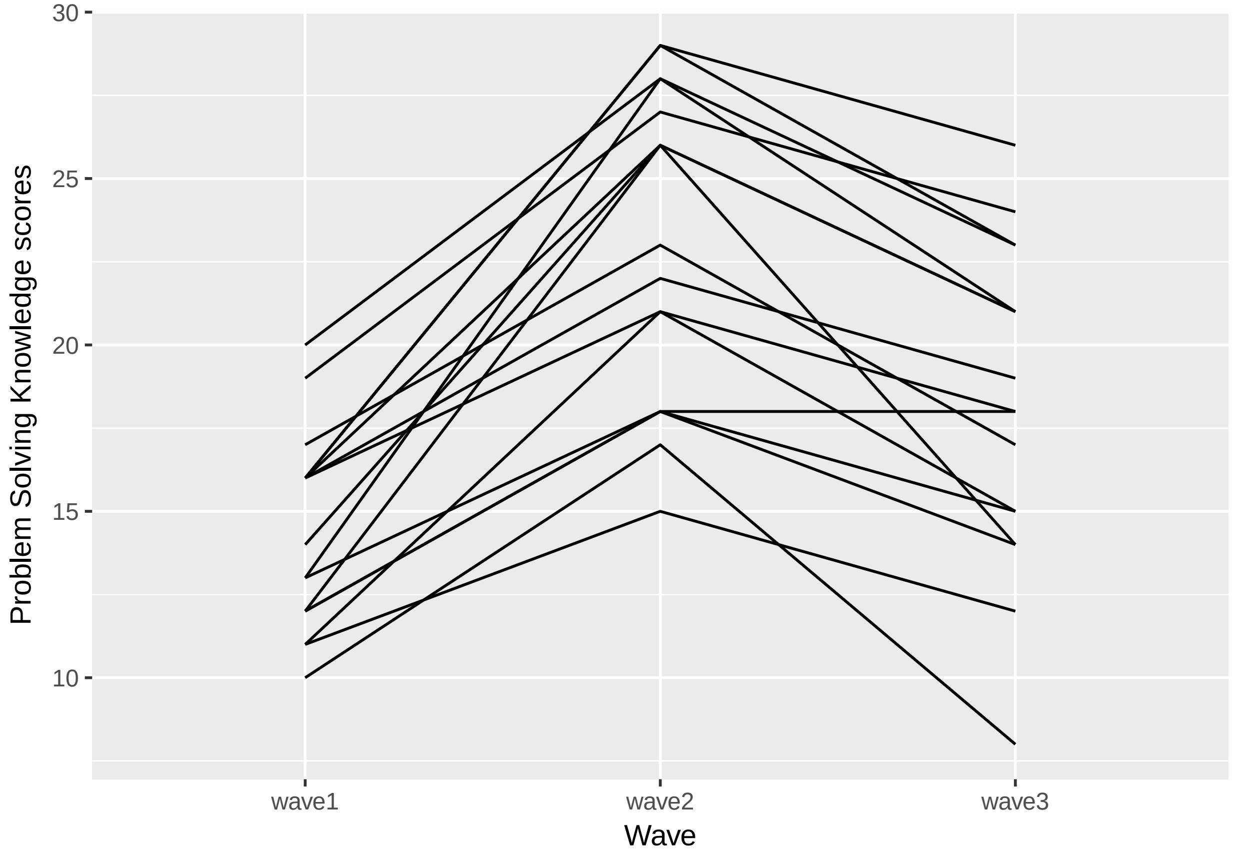


Figure 9.6: Problem Solving Knowledge score by wave, line graph

This graph, which plots the problem solving scores by wave, suggests that the $\eta\beta_{ij}$ interaction terms are not likely to all be zero; therefore assuming sphericity while testing the hypothesis of equal wave means is not likely to be justified. The tukey.add.test function in *asbio* (Aho (2016)) investigates H_0 *: main effect and blocks are additive.*

```
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

# if additivity exists a randomized block approach might be appropriate
#additive=with(data_long,lm(PrbSol~id+wave))
#anova(additive)
```

The Tukey additive test rejects the null hypothesis of additivity agrees with the line graph. In other words, a non-additive model is more appropriate for the problem solving knowledge data.

Step 3: Run ANOVA (including checks for the sphericity and normality of residuals assumptions)

The *ezANOVA* function (Lawrence (2016)) reports the F test, the Levene Test and an effect size. Type of the effect size depends on the model and indirectly depends on the type of sum of squares used, the *type* argument (1,2 or 3) transmits the choice.

```
library(ez)
#alternative 1 the ezANOVA function

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)
```

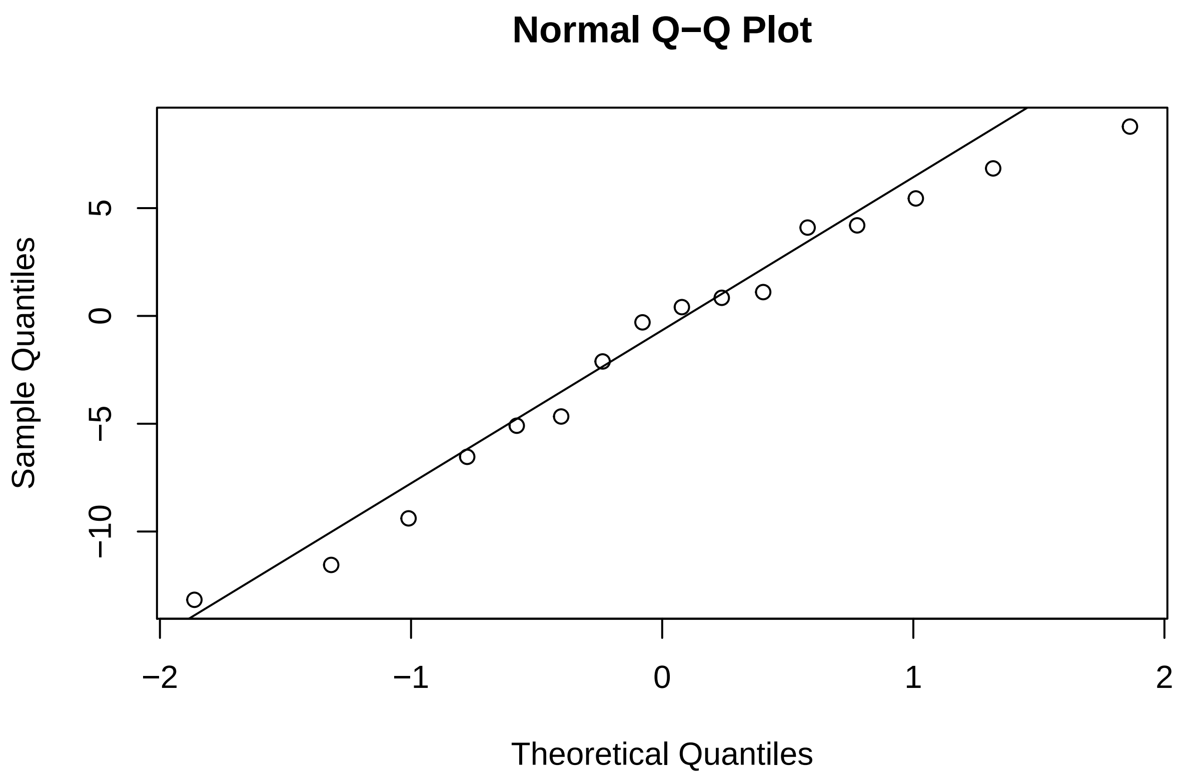


Figure 9.7: Inhibit model residuals

The distribution of the residuals is not severely non-normal.

The *aov* function is the second alternative.

```
# alternative 2 the aov function
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
```

9.3.1.3 Robust estimation and hypothesis testing for a one-way within-subjects design

One of the convenient robust procedures, a heteroscedastic one-way repeated measures ANOVA for trimmed means, has been compressed into the *rmanova* function, available via WRS-2 (Mair and Wilcox (2016)).

```
library(WRS2)

#rmanova
# 20% trimmed
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
```

9.3.1.4 Example writeup one-way within-subjects ANOVA

Descriptive statistics for the problem solving knowledge scores at each wave are presented in Table 6. The covariance matrix is presented in Table 7. A one-way within ANOVA was reported. F test was conducted at $\alpha = .05$. The assumptions of one-way within subjects ANOVA are satisfied. There was a significant difference between waves $F(2, 32) = 61.2, p < .001$. The *ezANOVA* function reported a generalized eta hat squared ($\hat{\eta}^2_G$) of 0.43.

9.3.1.5 Follow ups for one-way within-subjects ANOVA

To be added.

9.3.1.6 Missing data techniques for one-way within-subjects ANOVA

To be added.

9.3.1.7 Power calculations for one-way within-subjects ANOVA

To be added.

9.4 Mixed Design

To be added.

Chapter 10

Correlation

In our chapter 7, we introduced descriptive statistics; mean, variance, median, kurtosis, etc. These descriptive statistics aimed to ease the communication for a single variable. In other words, instead of transferring the entire raw data set to a colleague (or to a machine), providing these descriptives is generally satisfying and easier. However when the interest is in the association between variables, other measures are needed.

The sum of cross products, $S_{XY} = \sum(X - \bar{X})(Y - \bar{Y})$, can provide some information about the association. For example Figure 10.1 depicts an X and a Y variable. The sum of cross products for these two variables is zero.

```
##        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
```

The covariance between two variable is simply $Cov_{XY} = S_{XY}/n - 1$, but its a scale dependent measure, the correlation coefficient on the other hand generally has its bounds.

10.1 Pearson correlation coefficient

Pearson introduced a correlation coefficient in 1896. This coefficient ranges between -1 and +1, can be calculated as $Cov_{XY}/S_X S_Y$. This coefficient measures the linear relationship between two variables. Figure 10.1 depicts a correlation of zero. Even though X and Y in this figure are related to form a 14-sided polygon, the relation is not linear. Hence the correlation is zero. Figure 10.2 depicts several other associations; (A) is a perfect positive linear relationship, (B)is a positive correlation of .7, (C) substantially no linear relation, (D) is a correlation of -.4 and (E) is a correlation of -1.

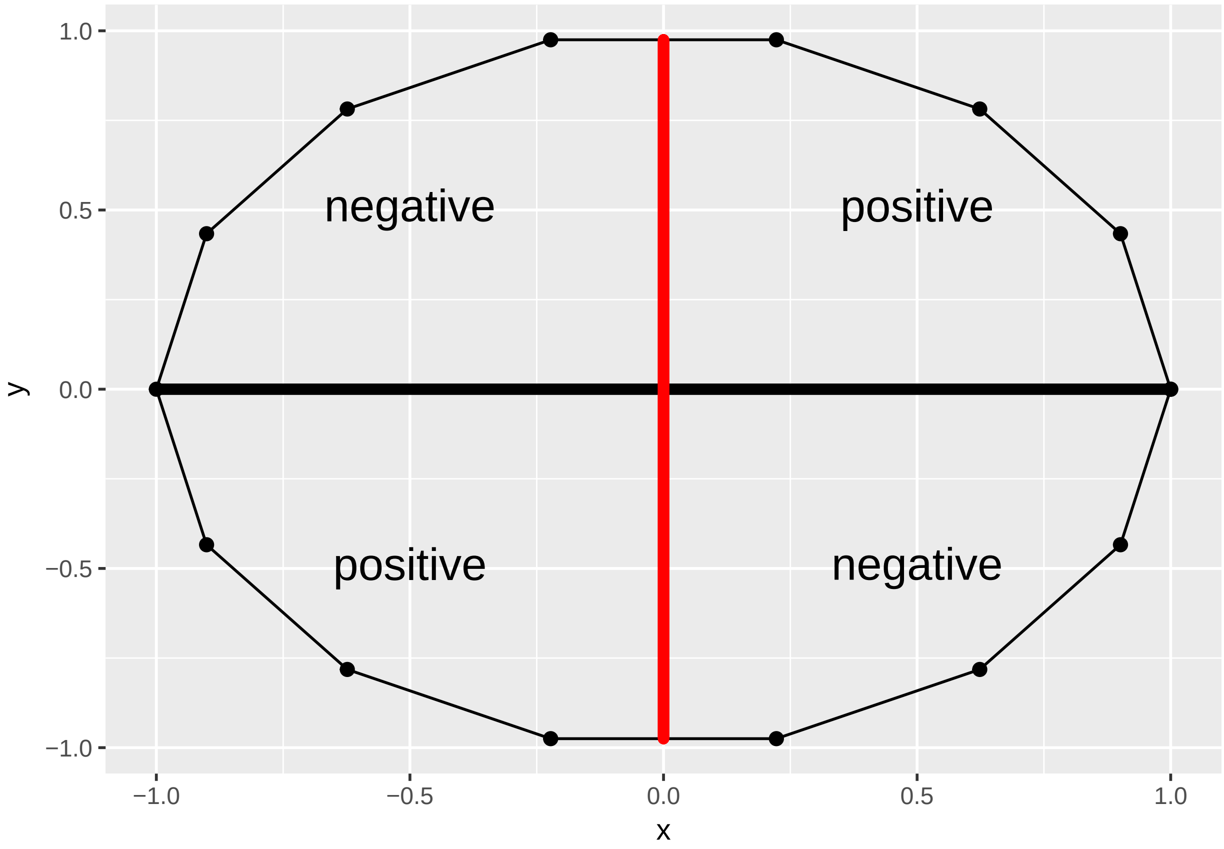


Figure 10.1: Sum of cross products=0

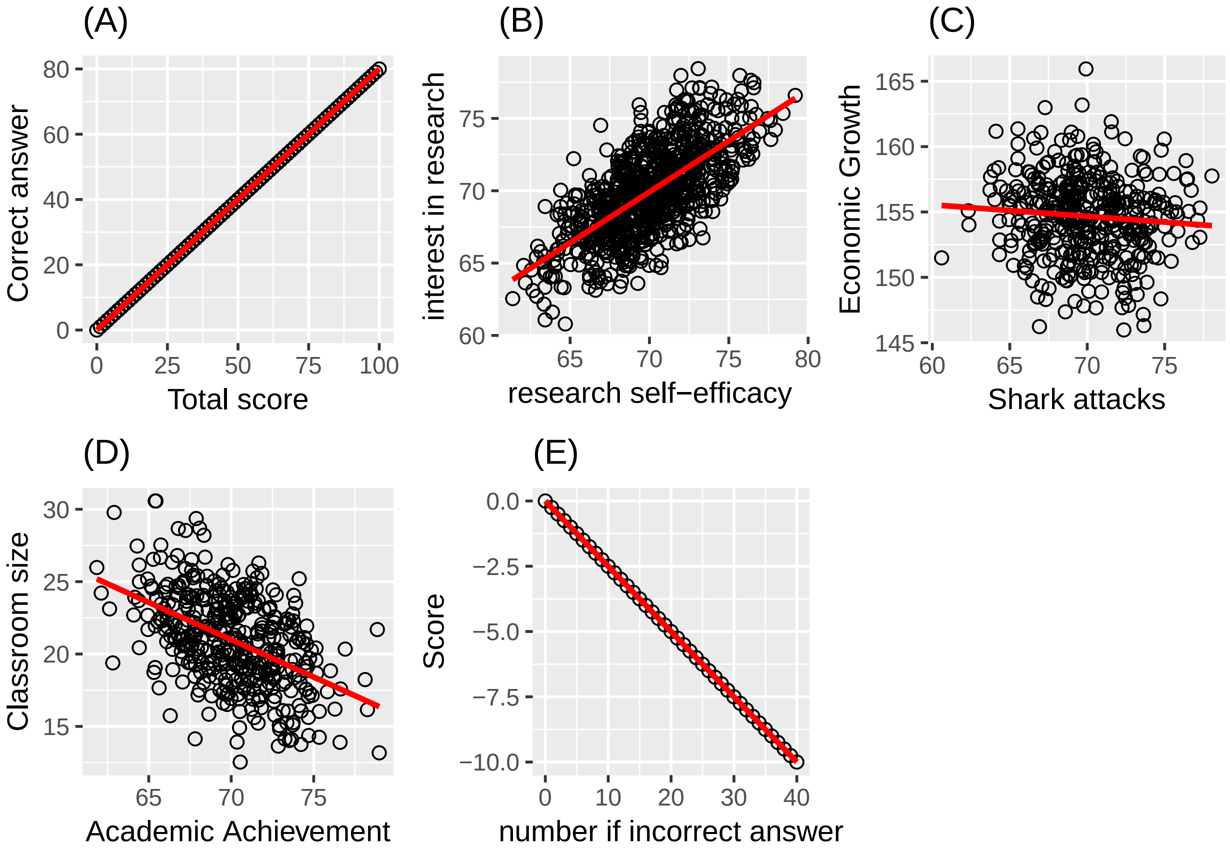


Figure 10.2: Correlation examples

10.1.1 Inference on a Pearson correlation coefficient

Information from the sample (r) can be utilized to make judgement about the population (ρ).

The z transformation , assuming a bivariate normality and a sample size of at least 10 (Myers et al. (2013)), is a helpful procedure to reach a judgement. The transformation equation is;

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

The standard error is;

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

Hence the confidence intervals are $z_r \pm z_{\alpha/2}\sigma_r$. Back transformation is needed to make interpretation about the correlation coefficient; $r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$.

Utilizing a normal distribution, a null hypothesis can be tested;

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

The t distribution can also be utilized to test $H_0 : \rho = 0$.

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

The distribution for this statistic follows a t distribution with a degrees of freedom of $n-2$.

10.1.2 R codes for Pearson Correlation coefficent

For illustrative purposes we selected the city of Bayburt. The Pearson correlation is computed for the association between the Gender Attitudes scores and the annual income per person. The income per person is calculated as “total household income” divided by the “total number of residents in the house”.

```
# load csv from an online repository
urlfile='https://raw.githubusercontent.com/burakaydin/materyaller/gh-pages/ARPASS/dataWBT.csv'
dataWBT=read.csv(urlfile)

#remove URL
rm(urlfile)

#select the city of Elazig
# listwise deletion for gen_att and education variables
dataWBT_Bayburt=dataWBT[dataWBT$city=="BAYBURT",]
#hist(dataWBT_Bayburt$income_per_member)
```

The bivariate distribution can be seen in 10.1.2. This is an interactive graph, please use your mouse to inspect it, created with the *rgl* package (Adler and Murdoch (2017)).

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

Gender Attitudes and Income

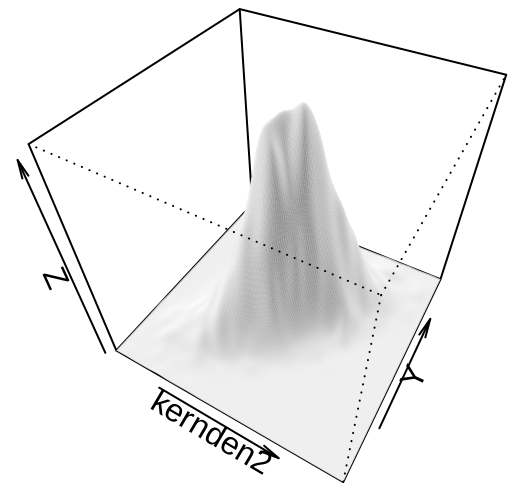


Figure 10.3: Bivariate Normal Distribution

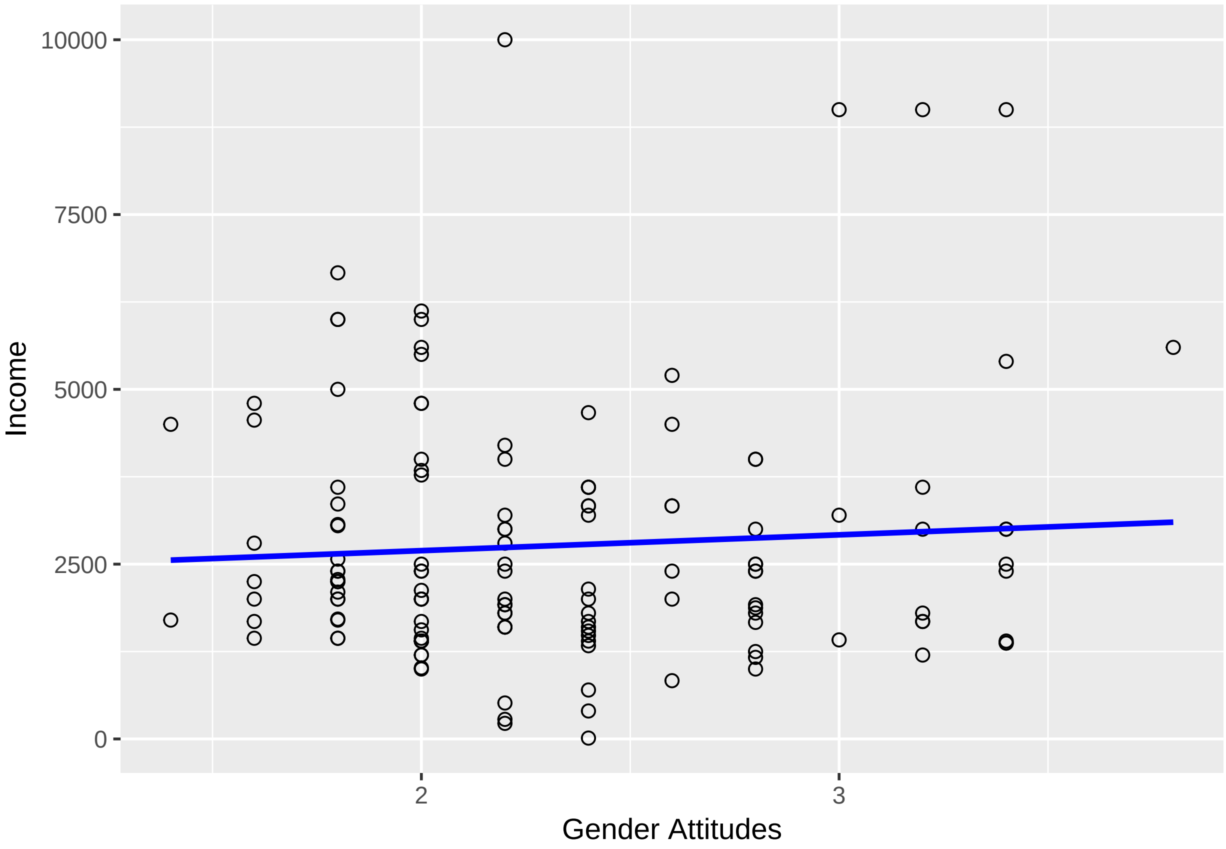


Figure 10.4: Bayburt: Gender attitudes vs income scatterplot

Bivariate normality seems to be violated. For comparison, below graph 10.3.depicts a bivarite normal distribution with r=0.7. Nevertheless, for illustrative purposes we use these data to test the null hypothesis $H_0 : \rho = 0$ against the non-directional alternative hypothesis $H_1 : \rho \neq 0$.The scatter plot is provided in 10.4.

The correlation between these two variables is computed by the *cor* function in the *stats* package (R Core Team (2016b)). The *cor.test* function in the same package performs the t-test and provides a confidence interval based on Fisher's z transformation.

```
#use ?cor to see use="complete.obs" is doing casewise deletion

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
```

These procedures can easily be hard coded.Stating $H_0 : \rho = 0$ and $H_0 : \rho \neq 0$

```
sample_r=0.06641641
r0=0        #the null
sample_n=137       # the number of complete.cases
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))

#the z test statistic
(zr-z0)/sigmar
## [1] 0.77


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


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


(exp(2*ll)-1)/(exp(2*ll)+1)  #transformback
## [1] -0.102
(exp(2*ul)-1)/(exp(2*ul)+1)  #transformback
## [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
```

A percentile bootstrap method might perform satisfactorily as a robust approach (Myers et al. (2013))

```
#Calculate 95% CI using bootstrap (normality is not assumed)
set.seed(31012017)
B=5000       # number of bootstraps
alpha=0.05   # alpha

#gender attitudes and income
originaldata=dataWBT_Bayburt2

#add id
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)

## Non-directional
# lower limit
output[as.integer(B*alpha/2)]
## [1] -0.138

# d star upper
output[B-as.integer(B*alpha/2)+1]
## [1] 0.252

# investigate the WRS package
# library(WRS2)
# pball(originaldata[,-3])
```

There are alternatives to percentile bootstrapping for a correlation coefficient, extensively discussed by Wilcox (2012). The WRS 2 package offers two alternatives, the percentage bend correlation and the Winsorized correlation. Only for illustrative purposes below is an R code;

```
# investigate the WRS package
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)
##
## Robust correlation coefficient: -0.0197
## Test statistic: -0.229
## p-value: 0.82
```

Write up: We tested a null hypothesis stating the gender attitudes scores and income variable are correlated against an non-directional alternative hypothesis. The Pearson correlation coefficient was $r = .066, p = .44$, the confidence interval with a .05 probability of a type I error using the z transformation is -.10 to .23. The null hypothesis is retained. This conclusion is consistent with the bootstrap results, using 5000 iterations, the 95% CI is -.138 to .252.

Sign difference note The Pearson correlation coefficent is .066 but not significantly different than zero. The WRS package functions also agreed to retain the null but the coefficent was negative. The income variable was sligthly skewed due to a small number of relatively large income values. In fact, when the World Bank team analyzed the data using a regression, they top-coded and transformed the income variable (for details Hirshleifer et al. (2016)). Let us top-code and transform the income variable, inspect bivariate normality and calculate the Pearson correlation;

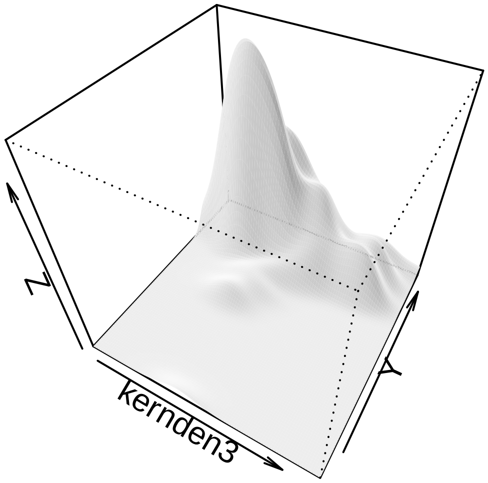


Figure 10.5: Top-coded and transformed income variable

```
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
```

Top-coding and transforming the income variable produced a distribution relatively closer to normal. The sign of the Pearson correlation coefficent is negative.

10.2 Spearman's rho and Kendall's tau

When the data is in the rank format, or there is a need for protection against outliers[1] when working with continuous data the Spearman correlation coefficient is used. If the number of ties in the ranks is not large, procedures provided for the Pearson correlation coefficient can be utilized. Setting the method argument to "spearman", the *cor.test* function first transforms the data into ranks and performs the procedures introduced for the Pearson coefficient.

10.2.1 The R code for Spearman's rho and Kendall's tau

We calculated the Pearson correlation coefficient to assess the association between the gender attitudes scores and the income for the participants in Bayburt. The Spearman correlation coefficient can conveniently be calculated by R;

```
#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="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:
```

[1] Here protection refers to being less sensetive to outliers compared to Pearson coefficient. However Spearman's rho and Kendall's tau might be more sensetive to outliers compared to robust procedures, see Wilcox (2012).

```
##     rho
## -0.0508
```

When there are ties, the *cor.test* function corrects the Spearman coefficient but the exact p value can not be calculated. Instead *exact=FALSE* argument yields a p value based on a t distribution. Field et al. (2012) suggests using Kendall's tau with large number of ties;

```
#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
```

The *exact=FALSE* argument with *method="kendall"* uses normal approximation.

The Spearman correlation between the gender attitudes scores and income was $r_S = -.051, p = .56$, and the Kendall's tau was $\tau = -.037, p = .54$

10.3 Biserial and Point-Biserial Correlation Coefficients with R

The association between a continuous variable and a dichotomously reflected latent continious variable can be examined with a biserial correlation. In psychometrics , for example, biserial correlation is used for calculating the correlation between a total test score (continuous) and a dichotomous item score (assumed to underlie a latent variable).

For illustrative purposes let us use dichotomized item1[2] and the gender attitudes score. The *biserial* function in the *psych* (Revelle (2016)) package can calculate the bi-serial correlation;

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

The point-biserial correlation is calculated for an association between a dichotomous variable and a continuous variable. The *cor.test* function with *method="pearson"* can be used to calculate a point-biserial correlation. The association between the gender and the gender attitudes scores is examined below;

```
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",
```

[2] This item is indeed dichotomized by the Worldbank team in their analyses

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

10.4 Phi Correlation Coefficient with R

When the two variables are dichotomous, a phi (ϕ) correlation coefficient is calculated. For illustrative purposes we calculated the phi coefficent between the gender and the wage variable. This variable equals to "yes" if one of the house members receives wage in the past 12 months. The *phi* function in the *psych* package requires the 2 x 2 matrix of frequencies to calculate the phi coefficient.

```
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

phi(genderWAGE)
## [1] -0.13
```

10.5 Issues in Interpreting Correlation Coefficients

Several issues arise in interpreting correlation coefficients.

Causation A correlation coefficient does not imply causation. For any correlation there are at least four possible interpretations involving cuasation: (a) X causes Y, (b) Y causes X, (c) both X and Y share one or more common causes, and (d) X and Y have different causes, but these causes are correlated.

The magnitude Whether a correlation of .6 is large or not depends on the context. For example suppose the .6 is the correlation between scores on two forms of a standardized achievement tests. This correlation is called an alternate forms reliability coefficient. Alternate forms reliability coefficients for standardized tests are expected to be at least .70 and preferably higher, so the .6 correlation would be regarded as small. Now suppose the correlation is between GRE scores and GPA. The correlation between GRE scores and GPA is typically somewhere between .10 and .30, so a .60 correaltion would be a very large correlation coefficient.

Outliers Correlation coefficients can be misleading when the data set contains outliers.

Reliability If either X or Y contains measurement error, the effect of the measurement error is to attenuate the correlation coefficient. Attenuate means to make the correlation coefficient closer to zero than it would have been if there had been no measurement error.

Its possible to correct for attenuation using

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

where r_{xx} and r_{yy} is the reliability coefficients.

- *When NOT to correct for attenuation:* When a variable is used for practical decision making and we are interested in the validity of those decisions, we should NOT correct for attenuation because the decisions are made on the basis of an observed variable, not a true variable.

- *When to correct for attenuation:* We can correct for attenuation when our motivation is to examine theory.

- *Comparison of Correlation Coefficients:* A comparison of correlation coefficients for two variables with a third variable can be affected by differences in reliability for the first two variables. If we are interested in theoretical relationships between variables and we want to compare the strength of relationship of two constructs (call these A and B and let them be measured by X1 and X2) with a third (call this C and let it be measured by Y), the comparison of the strength of relationship between A and C to the strength of relationship between B and C is compromised if X1 and X2 have different reliability coefficients. To compare strength of relationship we want the reliability of X1 and X2 to be the similar. Of course, it is best if both reliability coefficients are high, but it is critical that that are quite similar.

Unit of analysis A correlation calculated for one unit of analysis (e.g., individuals without regard to school) should not be applied to other units of analysis (i.e., individuals within schools or school means).

Variance in the two variables being correlated The correlation coefficient for two variables can be strongly affected by the amount of variance for the variables being correlated. Other things being equal when the variance of either or both variables is small, the correlation will tend to be small. If the variance for either or both variables is artificially small, misleading small correlation coefficients can occur. Variance can be artificially small due to

- Categorizing Based on Quantitative Variables

- Limited Range Scales

- Restriction of range

- Floor and Ceiling Effects

Chapter 11

Multiple Linear Regression, A Short Introduction

Scientific development requires that knowledge be transferred reliably from one study to another and, as Galileo showed 350 years ago, such transference requires the precision and computational benefits of a formal language. Pearl (2009)

The *formal language* in the quote refers to *mathematical equations.* Galton for example, in late 1800s, used equations to describe the relationship between the weights of mother and daughter pea seeds. Galton's work followed by Pearson's contributions led to initial idea of regression [1].

In the year 2016, the Web of Science reported that 60000+ abstracts of academic articles included the term "regression". The literature is vast, oftentimes the regression is mentioned as the *workhorse.* It is extensively used by frequentist and Bayesian statisticians, and more generally by data scientists in hundreds of different disciplines. The explanation of the popularity of regression analysis is simple, unless they are simulated by a machine, connections between variables, whether observed or latent variables, in a data set requires more complex statistical solutions than are provided by correlation coefficients.

It is not feasible to cover regression in a book chapter. We briefly introduce basics of a relatively simple multiple regression model.

11.1 Matricies and Least Square Estimation

In a multiple regression framework, demonstrating process of model fitting based on matrices and least squares estimation should have at least two benefits; (a) a simple demystification of the procedure, (b) a workable and sensible foundation for readers with a desire to move further in the advanced topics. The following sections use two different data sets. The first data set includes only 12-cases to show calculations and is named the synthetic data. The second data is simulated with a larger sample size for illustrative purposes and is named the simulated data.

Consider a case in which data on three variables are collected and the researcher is interested in the relationship of one of the variables (i.e., the dependent variable) to the other two variables (i.e., the independent variables). Further, these three variables as continuous. The regression model in this scenario is

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

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

where i represents individuals i=1,...,n, Y is the dependent variable, X_1 and X_2 are independent variables, βs are the regression coefficients and ϵ is the random error term(residuals) . This model can be presented in a matrix equations

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

In this more general form, all the independent variables are represented in the X matrix and the regression coefficients are represented by the β matrix. Let us assume the researcher has the following data

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

This synthetic data set has only 12 cases. The researcher can form 2 matrices and use these to calculate $\hat{\beta}$, the estimate of β.

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

Using the least square procedure the β coefficients can easily be estimated;

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

Let's calculate this with R for the synthetic-data;

```
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
```

The regression equation is

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

where $\hat{Y}_i$ is the predicted value for the i^{th} individual. Equation (11.1) was derived to minimize the error sum of squares: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Y'Y - \beta'X'X\beta$.These estimates are Best Linear Unbiased Estimates.

Each independent variable has a mean of zero because they are mean-centered. Therefore, zero represents a score at the center of the distribution for both X_1 and X_2 and is therefore an interpretable score for both X_1 and X_2. When both predictors are zero (at their mean), the ($\hat{Y}_i$) is 2.92. That is, for participants with independent variables scores equal to the mean on both independent variables the expected dependent variable score is 2.92. An increase in X_1 of 1 unit is predicted to correspond to an increase of 0.20 units in Y when the X_2 variable is held constant. Similarly, an increase in X_2 of 1 unit is predicted to correspond to an increase of 1.55 units in Y while controlling for X_1. The term "controlling for" ("ceteris paribus") is necessary to describe the effect of an independent variable in a multiple regression. The coefficients .20 and 1.55 would provide information about the association of the dependent and independent variables, if the researcher had substantial understanding of the unit of measurement for the independent variables, that is, the importance of a "1 unit" change in each variable.

11.1.1 a) "Essentially, all models are wrong, but some are useful."

This aphorism belongs to Box and Draper (1987). The researcher should provide a convincing discussion about the relevance of the variables included in the regression model to the research questions addressed by the research. If there are important omitted variables, the beta coefficients are probably not valid. Hence the researcher is obligated to provide justifications on variable selections to claim usefulness of the results.

Consider the case below ;

```
#omit X2 from the synthetic-data
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
```

For our synthetic data, X_1 and X_2 had a correlation of .68. If the researcher fails to include X_2 in the model, the coefficient for X_1 is estimated to be 1.09. This is a dramatic change from 0.20. Omitting predictors that are related to both the other predictors in the model and the dependent variable will cause the coefficients for the variables that have not been omitted to be misleading. Therefore an important part of the theoretical justification of a regression model is a discussion of variables that may have been omitted.[2]

In addition to omitted variable issue, the validity of the results from a regression model (the usefulness) is also directly related to the sampling process and appropriate reflection of this process in the model. For example, if sampling weights exist they should not be ignored in the analyses. *Sampling and regression* is beyond the scope of this chapter.

11.1.2 b) Strength of relationship between the dependent and independent variables

The sum of squares for Y, which is also known as the total sum of squares, can be decomposed into two parts, *the model sum of squares*, which is also the sum of squares for the predicted values, and *the error*

[2] This might lead to a clue on popularity of controlled randomized trials.

sum of squares. The ratio of the model sum of squares to *the total sum of squares*, is called the sample squared multiple correlation coefficient and symbolized as R^2. The coefficient R^2 measures the strength of association between the dependent variable and the independent variables. Examine the R code below given for the synthetic data;

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

# SS 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
```

Also known as *coefficient of determination*, R^2 is a biased estimator of the population squared multiple correlation coefficient. A more nearly unbiased estimate is the adjusted squared multiple correlation coefficient. One benefit of adjusted R^2 is computational simplicity. Examine the R code below given for the synthetic data

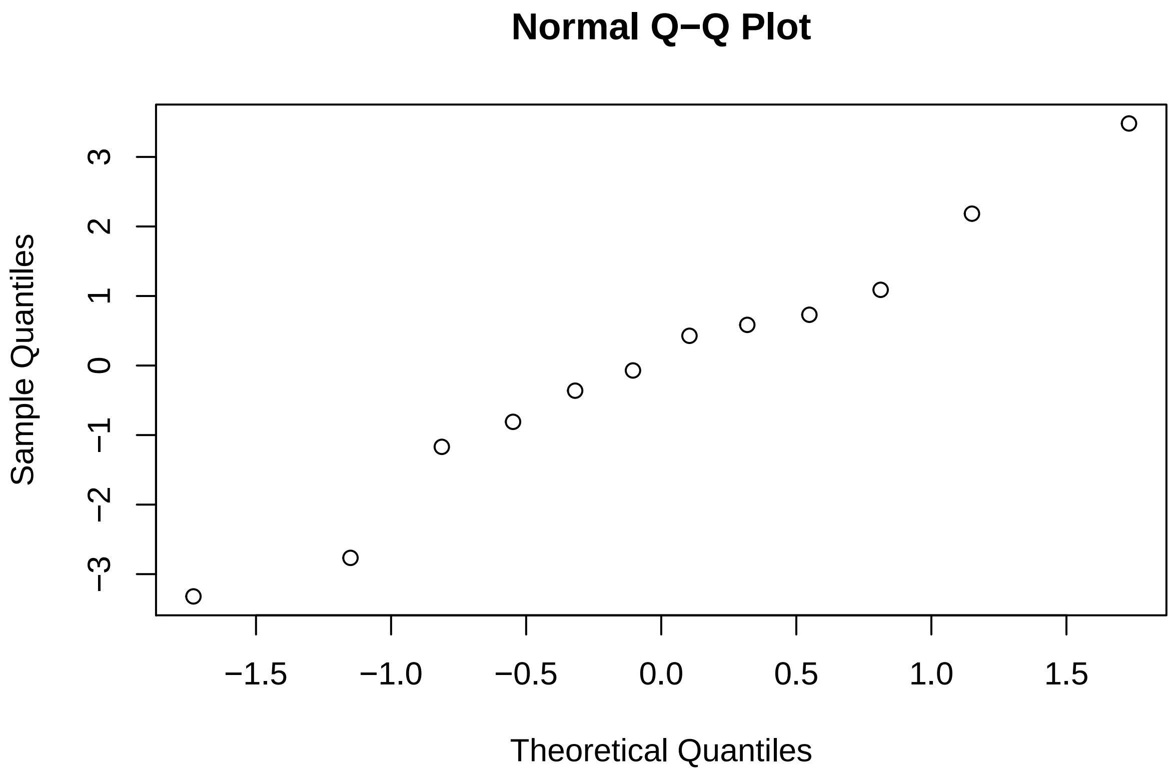
```
Rsquared=ModelSS/TotalSS
#sample size
n=12

#the number of predictors
p=2

# include intercept? 1 for yes, 0 for no

int_inc=1


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

R^2 and R^2_{Adj} are useful coefficients; they provide information on how much of the variance is explained. Note that in this example $R^2 = .879$ and $R^2_{Adj} = .852$ are very similar. However if were $R^2 = .25$ then R^2_{Adj} would be .08. When R^2 is 1, the model successfully explains 100% of the variance in Y and when R^2 is 0 the model does not explain any of the variance in Y. The coefficients R^2 and R^2_{Adj} are also useful for comparing the strength of relationship for different set of predictors to predict a specific outcome. The interpretation of the R^2 is similar to the interpretation of a correlation coefficient. Depending on the context a small R^2 value might be regarded as substantial, or an R^2 value of .7 might be regarded as low.

11.1.3 c) Residuals and influential data points

Residuals provide information for assessing potential problems with the model. Inspecting residuals can provide information about deviations from the assumed linearity of the relationships of the dependent variables to the independent variable. Inspecting the distributional properties of residuals is needed to provide evidence for the validity of statistical inference. For example, because the normality assumption is made when conducting significance tests and calculating confidence interval, residuals should follow a straight line on a Quantile-Quantile (QQ) plot. Examine the R code below given for the synthetic data:

```
#Predicted values
Yhat=X%*%betahat
residuals=Y-Hyat
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)
```



There are three common types of residuals;

- Unstandardized residuals, that is, $Y_i - \hat{Y}_i$. Unstandardized residuals are on the same scale as Y.

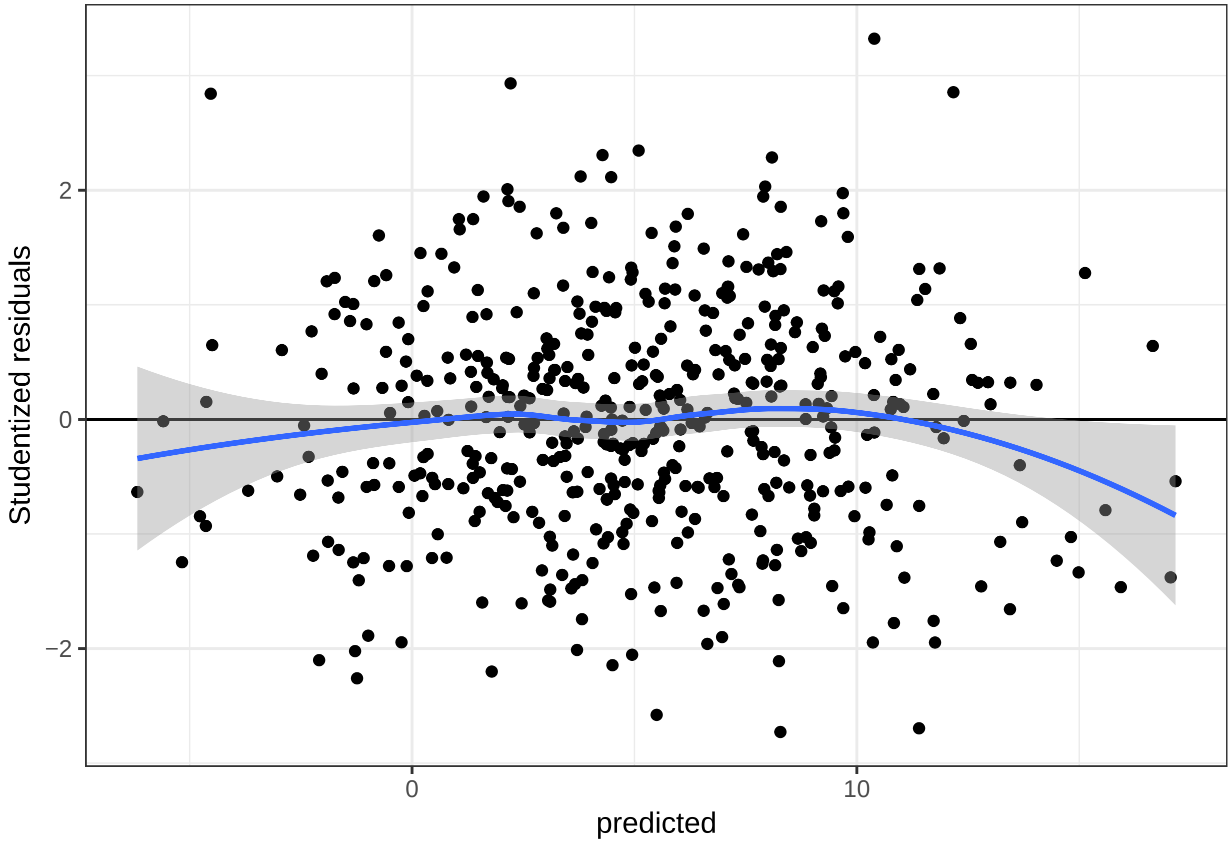
- Standardized residuals: The residuals divided by the overall standard deviation of residuals; Standardized residuals are on a z-score scale (M = 0, SD = 1). When residuals are assumed to be normally distributed, it is common practice to identify outliers as Y values for which the absolute value of the standardized residual is larger than 2. However, it should be noted that this practice can be misleading because outliers can cause the regression coefficients to be poorly estimated and/or can increase the standard deviation of the residuals and both effects can cause poor outlier detection. In addition, if the residual are in fact normally distributed approximately %5 of the participants will have residuals beyone +/- 2.00. See Wilcox (2012) for more information about outlier detection.

- Studentized residual: A studentized residual is ratio of the unstandardized residual to the estimated standard error of the residual.

When investigating residuals , these three types of residuals generally lead to same conclusions. The standardized residual are forced have a z-scale, and thus, -2 and +2 are commonly pronounced cut offs. The studentized residuals are connected to the t distribution; $t_{n-p'-1}$ where n is the sample size p' is the number of coefficients in the model (i.e intercept+two predictors =3). It is argued that when detecting outliers in residuals, investigating the studentized residuals is more convenient (Rawlings et al. (1998)).

Scatter plots of residuals vs. predicted values can provide information about whether the assumed linear relationships between the independent variables and the dependent variable are adequate. Ideally the scatter plot should not show a detectable pattern. Here is a plot of studentized residuals vs fitted values, from a regression model fitted to simulated-data in which the linearity assumption is adequate. The simulated data have a sample size of 500 and two independent variables.

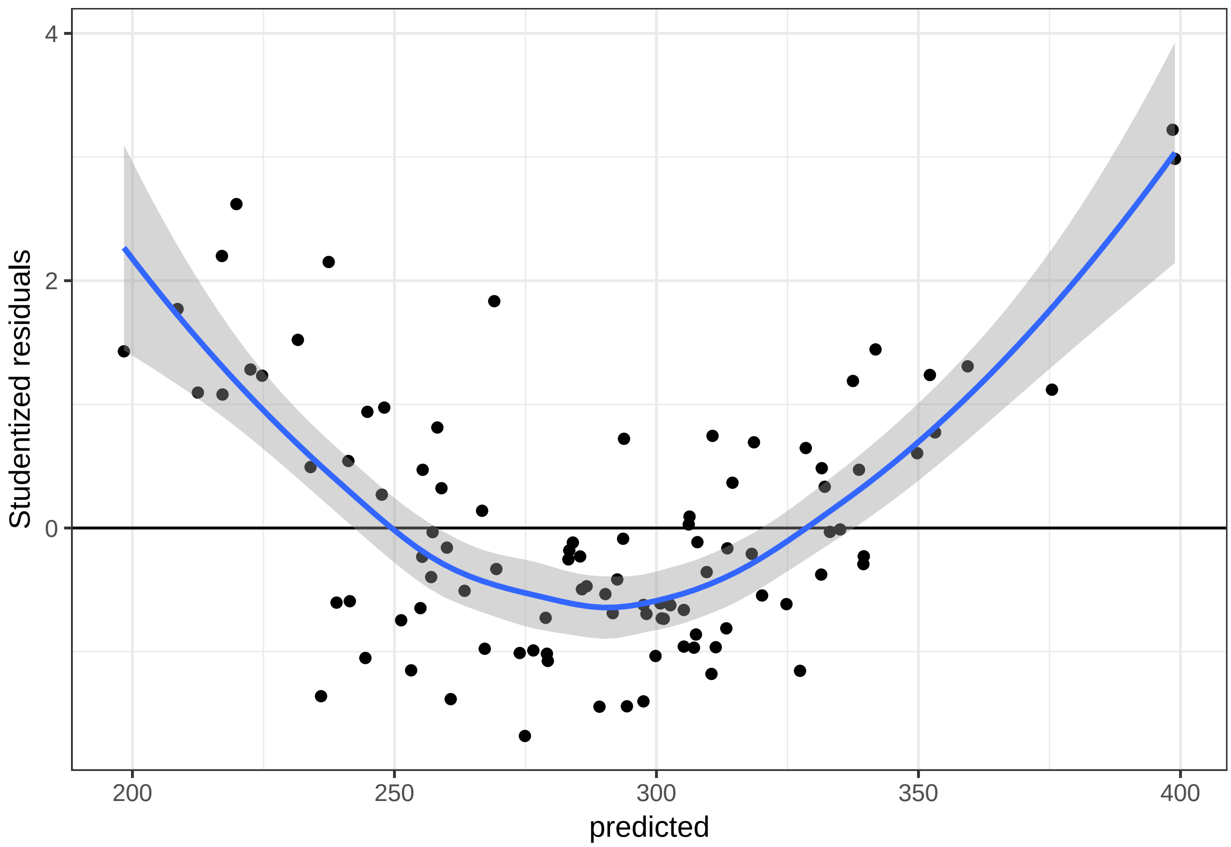
```r
#simulate data
library(mvtnorm)
sigma <- matrix(c(4,2,2,3), ncol=2)
xx <- rmvnorm(n=500, mean=c(0,0), sigma=sigma)
yy=5+xx[,1]*2+xx[,2]*-3+rnorm(500,0,1.5)
model=lm(yy~xx[,1]+xx[,2])
errors=rstudent(model)
predicted=predict(model)

#Standardized Residuals vs Yhat
library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ ylab("Studentized residuals")+
  theme_bw()+stat_smooth()
```



The blue line above, which is determined independently from the regression model, should be compared to the horizontal line at 0. The more similar the two lines, the less likely the linearity assumption is violated.

Here is a plot, studentized residuals vs fitted values, from a mis-specified regression model on a simulated-data. The sample size is 100, there are two independent variables, and the relationship of Y and X2 is quadratic.

```
#simulate data
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)

#Studentized Residuals vs Yhat
library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ylab("Studentized residuals")+
  theme_bw()+stat_smooth()
```



There is a pattern indicating that the model is omitting a quadratic association.However, this graph does not inform about the source of the quadratic association, see non-linearity section below.

Unusual residuals should be inspected. Even when the residuals are substantially normally distributed and there is substantially no-pattern for the residual vs predicted value plot, there might be unusual residuals. Deciding whether a residual is unusual or not (e.g 3,4 or 5 standard deviation above), and more importantly whether to keep the observation in the data set or not requires justifications. Examine the code below to simulate data and examine the studentized residuals:

```
#simulate data
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$SUTresiduals=rstudent(model)
# how many of the residuals are larger than a critical value?
# lets use alpha=.05
sum(abs(tempdata$SUTresiduals)>qt(c(.975), df=100-3-1))
## [1] 8

#which observations?
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
```

Assume we justified the use of $t_{.975,96}$ as the critical value, in which alpha=.05. We should expect approximately $n * .05$ (in our case 100*.05=5) cases larger than the critical value. In this particular case, even though 8 cases were identified, none of them seems unusual.

If the researcher detects an abnormality and further, if the researcher decides to remove the observation from the data, it should be done one observation at a time. The justification of removing a data point should be given clearly. A better alternative, on the other hand, may be to use an estimation method that is robust to outlying data points.

R program is convenient for investigating influential data points. Examine *?influence.measures* below for the simulated data set;

```
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
```

This output reports 5 different measures.

In this example, cases 12, 33, 41, 42, 50, 64 and 100 are reported to be *potentially* influential. As they highlighted by an asterisk, they labeled as potential using the covariance ratio criteria (cov.r). This value reports the impact of an observation on the sampling variances of the regression coefficients. Values larger than $1 + (3p'/n)$ and lower than $1 - (3p'/n)$ are labeled as influential, in our case, n=100 and p'=3, hence the cut offs are 1.09 and .91.

The Dfb (DFBETAS) for each predictor reports how much the coefficient for the predictor changes when the case is removed. It is the difference between the two coefficients divided by an estimate of the standard error of the new coefficient and therefore is on the scale of a t statistic. R places an asterisk if the value is larger than $2/\sqrt(n)$. For this specific illustration the cut off value is $2/\sqrt(100) = .2$.

The dffit reports the change in the predicted value for the i^{th} case when the i^{th} case is removed from the data. The criterion for identifying potentially influential data points is $2 * \sqrt{\frac{p'}{n}}$.

Cook's distance (cook.d) measures the influence of a particular case on all of the estimated coefficients and values larger than $F_{.5,p',n-p'}$ are highlighted. Cook's distance also measures influence of omitting a particular case of the predicted values for all of the remaining cases.

Leverage Values (Hat Diag) measure the distance of an observation compared to other independent variables. Values larger than $2p/n$ are considered to identify potentially influential data points.

It is researcher's responsibility to examine any potentially influential data points.

11.1.4 d) *Equal variance assumption*

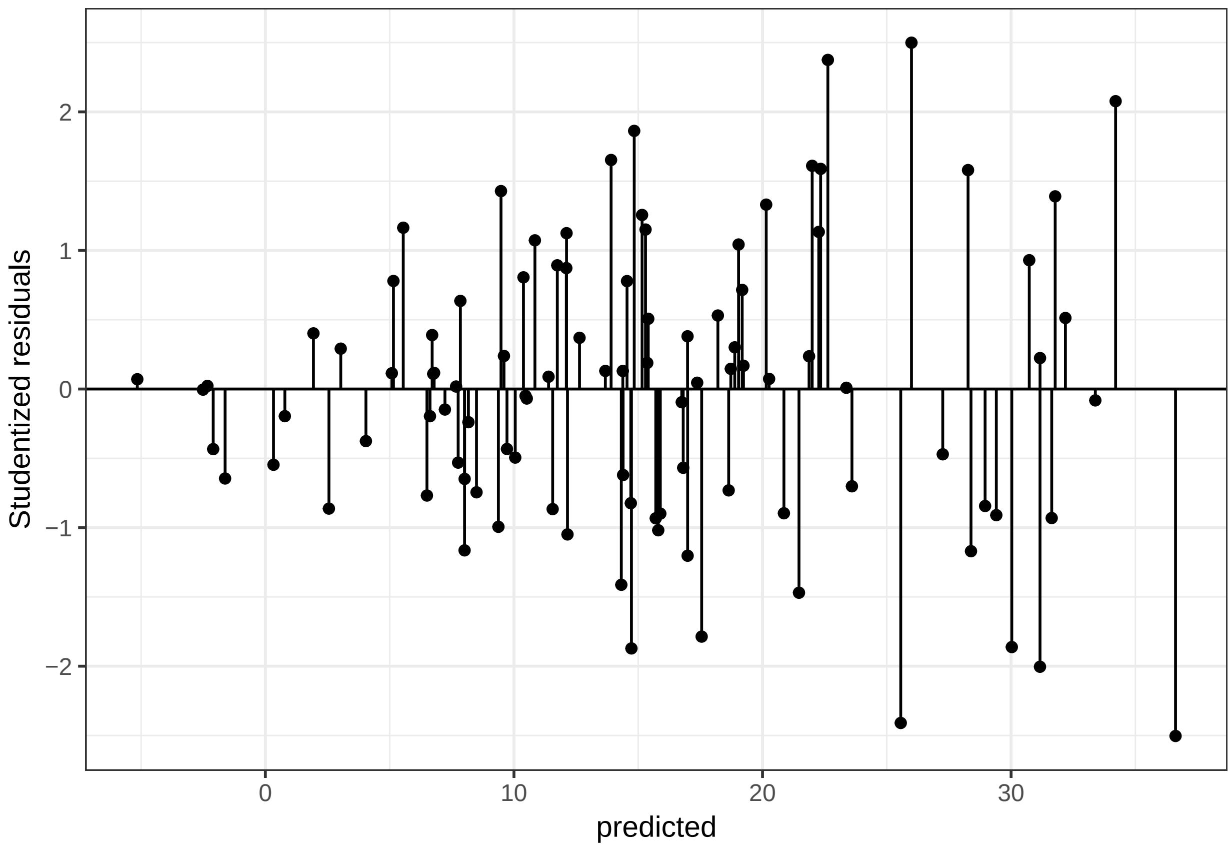
The standard errors of the of the coefficients are calculated as the square roots of the diagonal elements of $\hat{\sigma}^2(XX)^{-1}$, where $\hat{\sigma}^2$ is the variance of the residuals. Examine the code below given for the synthetic data set:

When using the OLS with an assumption of normally distributed Y variable, the distribution of β can be obtained. Examine the code below given for the synthetic-data set;

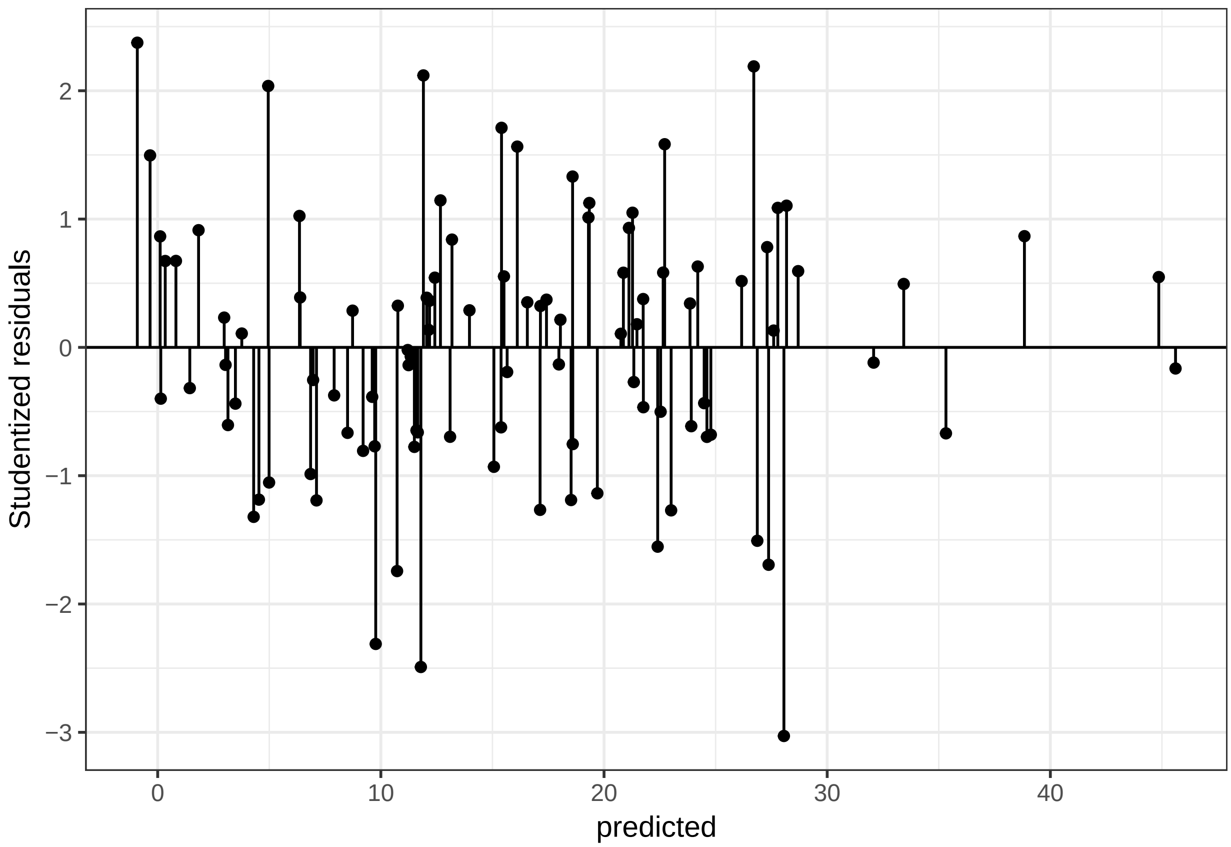
```
#Residuals
s2 <- (t(residuals) %*% residuals)/(nrow(Y)-nrow(betahat))
Var_betahat <- s2[1,1]*solve(t(X)%*%X)
```

The equation $\sigma^2(X'X)^{-1}$ is valid under the assumption of homogeneity, that is, observations on the Y variable have a common variance controlling for the independent variables. In other words, every observation of Y has the same amount of information (Rawlings et al. (1998)). With this assumption, regression coefficients are selected to minimize $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. In this expression equal weights are given to the residuals for every case.. If homogeneity is questionable the estimator can be modified to allow for unequal weights or replaced. Alternatively the Y variable can be transformed or the estimator of the standard error can be modified (see package 'sandwich' Lumley and Zeileis (2015)). Otherwise, the standard error of $\hat{\beta}$ could be underestimated or overestimated. Underestimation results in Type I error rates that are larger than the alpha level used in hypothesis tests and confidence intervals and over estimation results in reduced statistical power. It is common practice to plot residuals against the predicted values to study heterogeneity. Examine the code below to simulate data with unequal variance and examine the studentized residuals:

```
#simulate data
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 function
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)

#Studentized Residuals vs Yhat
library(ggplot2)
plotdata=data.frame(errors,predicted)
ggplot(plotdata, aes(x = predicted, y = errors)) +
  geom_point() + geom_hline(yintercept=0)+ylab("Studentized residuals")+
  geom_segment(mapping=aes(xend = predicted, yend = 0)) +
  theme_bw()
```



The variance with smaller $\hat{Y}$ values are smaller. Below is a graph for a regression model on a simulated data with equal variance.



11.1.5 e) Hypothesis testing

The F test is used within a multiple regression framework to test $H_0 : \beta_1 = ... = \beta_p = 0$, a hypothesis stating that the p regression coefficients are all equal to zero in the population. The alternative hypothesis states that at least one coefficient is not zero. The null hypothesis can be tested using the statistic $MS_{regression}/MS_{residual}$. This statistic follows an F distribution with p and $n - p'$ degrees of freedom. As mentioned earlier, p is the number of predictors and p' is the number of coefficients (p'=p if there is no intercept). Examine the code below given for the synthetic data, setting Type I error rate = .05;

```
# Model SS and Total SS calculated before
dfREG=2  #(p=2, predictors X1 and X2)
dfRES=9  #(n-p', 12-3)
MSreg=ModelSS/dfREG
MSres=(TotalSS-ModelSS)/dfRES

MSreg/MSres
##      [,1]
## [1,] 32.8

#critical F
qf(.95,dfREG,dfRES)
## [1] 4.26
1-pf(MSreg/MSres,dfREG,dfRES)
##          [,1]
## [1,] 7.39e-05
```

The t-test is used for investigating $H_0 : \beta_X = \beta_{hyp}$ vs $H_1 : \beta_X \neq \beta_{hyp}$. Most commonly $\beta_{hyp} = 0$

The statistic $(b_X - \beta_{hyp})/SE(b_X)$ follows a t-distribution with N-p' degrees of freedom.Examine the code below given for the synthetic-data;

```
# test if the coefficient for X2 is different than 0
Bhyp=0  #hypothesized value

# estimated coefficient for X2 (see betahat calculated before)
bx2=betahat[3]

# estimated SE for X2 (see var_betahat calculated before)
se_bx2=sqrt(Var_betahat[3,3])

#t statistic
(bx2-Bhyp)/se_bx2
## [1] 5.33

# t critic
qt(.975,9)
## [1] 2.26

#p value
2*(pt(-abs((bx2-Bhyp)/se_bx2),9))
## [1] 0.000478
```

11.1.6 f) Variable Selection

Broadly speaking there are two situations in which multiple regression is used to analyze data.

The first is illustrated by the following example. A social science researcher conducts an extensive literature review, identifies all independent variables relevant to the research questions, collects the data, estimates a model in which all independent variables are included and reports results for this model.

The second is illustrated by an example in which the researcher has data on a very large set of variables and does not know prior to analyzing the data which variables will be included in the final model that will be reported. This might happen because the researcher is working in a relatively new research area and collects data on a wide variety of variables or is conducting a secondary data analysis of a data set with a wide variety of predictors. In either case the researcher may want to begin by conducting variable selection that is using statistical results to select the best subset of many independent variables. There are several approaches to select the best subset of predictors. For example, stepwise regression, backward selection or forward selection is covered in many sources. However, in our experience, when applied to the same data set these three approaches are likely to give different answers.

A convenient approach with R is to study all possible regressions.For introductory purposes, examine the code below given for a simulated data set;

```
#simulate data
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)
```

```r
library(leaps)
formula <- formula(paste("yy ~ ",
     paste(names(simdata[2:4]), collapse=" + ")))
allpossreg <- regsubsets(formula,nbest=3,data=simdata)
aprout <- summary(allpossreg)

#this functions reports more than R-squared and adjusted R-squared
#examine str(aprout)


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

This table reports that intercept and X_2 only model results in an R^2 value of .753. When all predictors included, the R^2 reaches to .890; however, excluding the X1 from the full model reduced the R^2 only by .019. Below is a graphical decpiction.

```r
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) Colliniearity

Collinearity is the degree to which the predictors are correlated among themselves. The correlation between predictors is a concern in regression because the standard errors of the coefficients increase as collinearity increase and therefore collinearity hides the individual contribution of each predictor in the regression equation.

As an illustration suppose there are two independent variables with r = .9. You MIGHT have two types of problems: The regression coefficients become unstable (i.e. they would vary a great deal across different

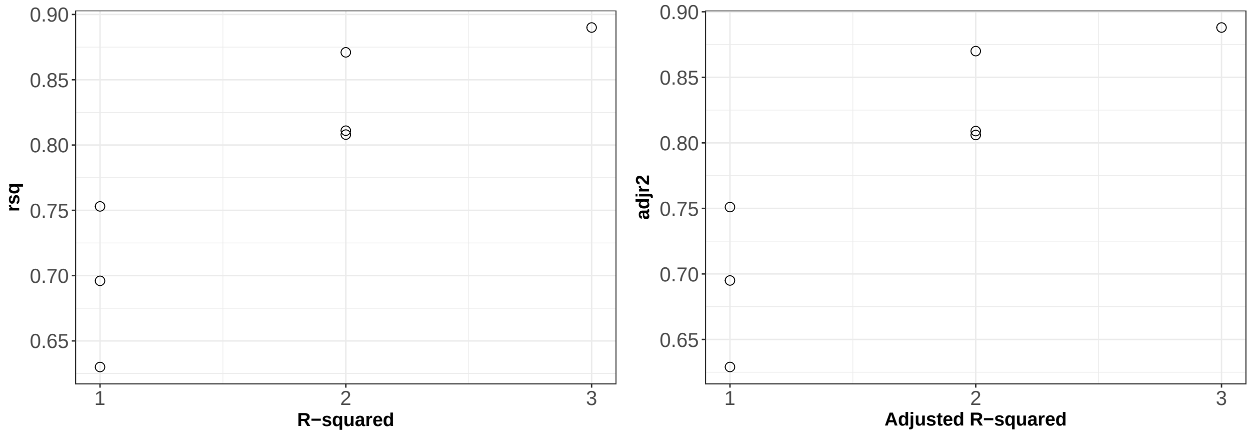


Figure 11.1: All Possible Regressions

samples obtained from the same population).

You may obtain a statistically significant R^2 but not statistically significant regression coefficients.

Variance inflation factor (VIF) is helpful to detect collinearity in regard to a particular independent variables and can be applied in models two or more independent variables. The formula is $VIF_x = \frac{1}{1-R_X^2}$ where R_X^2 is the R^2 when the predictor is regressed on the remaining independent variables. Large VIF values are indicator of possible multicollinearity. Commonly pronounced cut off values are 4 and 10, however VIF values are indirectly affected by sample size and variance (Obrien (2007)). When large VIF values are detected, the researcher should examine the problem. It might be justifiable to (a) leave out one of the highly correlated predictor, (b) combine the two highly correlated variable. The decision should be made cautiously given that the possible solution might be more problematic than a large VIF value, see Obrien (2007). Examine the code below given for a simulated data set;

```
#check correlations among predictors
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

#the largest correlation is.73
#no multicollinearity expected

library(car)
vif(lm(yy~X1+X2+X3,data=simdata))
##   X1   X2   X3
## 2.36 2.50 1.98
# no problematic VIF values
```

11.1.8 h) Non-linearity

In the presence of a non-linear relation between the dependent variable and any given independent variable, ignoring non-linearity is simply a validity concern due to the omitted variable issue. Examining the residuals is helpful to detect non-linearity. Residuals should be plotted against predicted and independent variables. A common practice is to include higher order variables in the model, for example, if the plot indicates a non linear pattern for X_k against residuals, X_k^2 might be needed in the model. The type of the non-linearity , such as quadratic, cubic or quartic should be treated accordingly. Gelman and Hill (2007) , commenting on age variable when the age and dependent variable are not linearly associated, prefers treating the variable as a categorical predictor. Alternatively transformations of the dependent or independent variable may be employed.

11.1.9 i) Correlated errors and nonindependent errors

Errors should not be correlated or more broadly should be independent. When such dependency is not addressed, regression results are invalid. This topic ,however, is well beyond the scope of this introductory material. Correlated errors are likely to distort the standard errors for the beta coefficients. This is not desired. In social sciences, correlated errors might be present when measurements are repeated. Multilevel models and latent growth models has been developed to address appropriate modeling of repeated measure designs. Nesting of participants in subgoups is another common source of non-independent errors in social sciences. Multilevel models are one popular solution to model clustered data.

11.1.10 j) Centering and Scaling

Consider an example in which mother's age at the date of her child's birth (maternal age) is used to predict IQ at age 10. The intercept estimates average IQ for children whose mother's maternal age was zero and cannot be meaningfully interpreted. Centering maternal age around its mean results in an intercept which estimates average IQ for children whose mother's maternal age was at the mean of the sample and can be meaningfully interpreted. Or consider predicting absences from work from an anxiety measure. A score of zero is possible score on the anxiety measure, but does not occur in the sample. The intercept estimates average absences for employees whose anxiety level is outside the range of the data and therefore represents an extrapolation for the data. Centering around the mean for the sample solves this problem. Another approach would be to center around an anxiety score that is in the range of the data and considered high. Or consider a study of income and an index of health. Income is on a scale in which a change of 1 represents a change of 1 dollar in income. The regression coefficient is .001, which represents a trivial change in the index. Dividing X by 1000 so that a change of 1 represents a change of 1000 dollars in income results in a regression coefficient of 1, which is a small but not trivial change in the index may make the results easier to think about.

11.1.11 k) Standardized coefficients

A related topic to linear transformations is to use a z-score for the continious predictors by substracting the mean and dividing by the standard deviation. Depending on the nature of the variable, using the z scores might ease the communication between researchers. Here are interpretation examples; Raw scores: An increase in anxiety of 1 unit is predicted to correspond to an increase of 3 units in achievement, holding the remaining predictors constant. z-scores: An increase in motivation of 1 standard deviation is predicted to correspond to an increase of 0.25 standard deviations in achievement, holding the the remaining predictors constant.

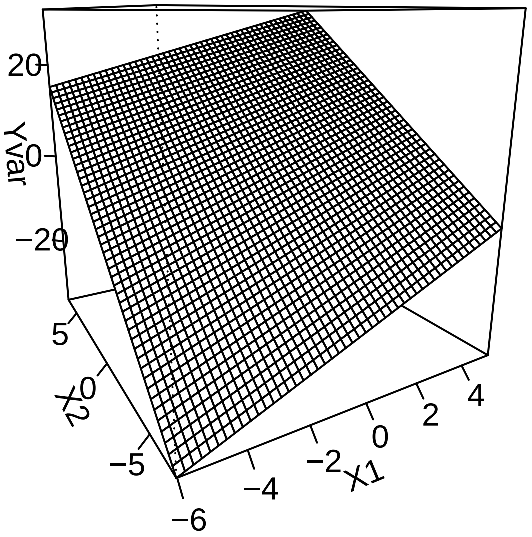
11.1.12 l) Interactions

We covered the basic idea of interaction in our ANOVA section. Ignoring an interaction is an omitted variable problem because an interaction affects the interpretation of main effects. For example, suppose a researcher investigates the relationship between mathematics achievement at the end of the school year (Y), effort measured by voluntary homework completed and submitted during the year (X_1), and mathematics achievement at the end of the preceding year (X_2). Using the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ assumes that the relationship between Y and X_1 does not depend on X_2 and will be misleading if the assumption is false. A common model used to investigate interactions is ;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i \tag{11.2}$$

The slope of the relationship between Y and X_1, for example, is $\beta_1 + \beta_2 X_{i2}$ indicating that the relationship depends on X_2. Similarly the slope relationship between Y and X_2 is $\beta_2 + \beta_1 X_{i1}$. Consideration of $\beta_1 + \beta_2 X_{i2}$ shows that β_1 is the slope of the relationship between Y and X_1 when $X_2 = 0$ and therefore β_1 cannot be meaningfully interpreted if $X_2 = 0$ is not a meaningful score or is outside the range of the data. This problem can be addresses by centering X_1 and X_2 around their respective means. It should be noted that the model in Equation (11.2) assumes that the interaction can be accurately modeled by including $\beta_2 X_{i1} X_{i2}$ in the model. This assumes the relationship between Y and X_1 is linear when is X_2 controlled. Violations of assumption of the model in Equation (11.2) should be investigated.R can be helpful in interpreting interactions by 3-dimension graphs. Examine the code below given for a simulated data set to highlight the use of R package *visreg* (Breheny and Burchett (2016))

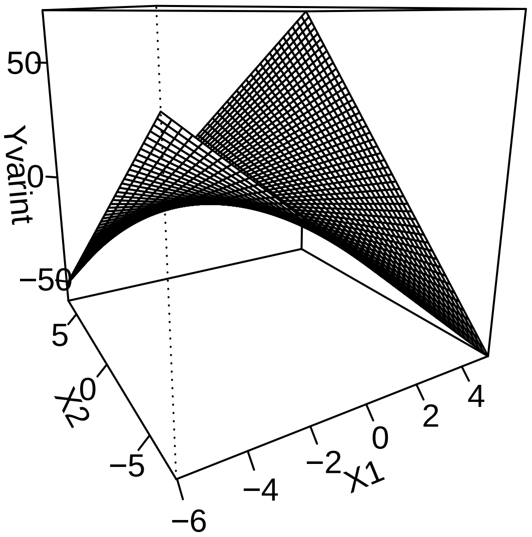
```
## manipulate simdata
## Yvar: dependent variable on no interaction
simdata$Yvar=3+simdata$X1*2+simdata$X2*3+rnorm(nrow(simdata),0,5)

library(visreg)

model=lm(Yvar~X1+X2+X1*X2,data=simdata)
visreg2d(model, "X1", "X2", plot.type="persp")
```



The surface is flat indicating no interaction.

```
## manipulate simdata
## Yvarint: dependent variable on a interaction
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")
```



The surface is no longer flat in the presence of an interaction.

11.1.13 m) Estimators

To be added

11.1.14 n) Robust Regression

To be added

11.1.15 o) Sample size and statistical power

To be added

11.1.16 p) Reliability of variables

To be added

11.1.17 q) The nature of the variables

To be added

11.1.18 r) Multiple dependent variables

To be added

11.1.19 s) Missing variables

To be added

Chapter 12

Useful R codes

```
# Convert numeric to factor

temdata[,2:9] <- lapply(temdata[,2:9], as.factor)

# Convert factor to numeric
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}
temdata[,2:5] <- lapply(temdata[,2:5], as.numeric.factor)

# Have frequencies table for multiple columns

dems=apply(temdata[,5:11], 2, function(x){table(x,temdata$grp)})
library (plyr)
mydems <- ldply (mydems, data.frame)

# Aggregate variables by grp

uncagg=aggregate(. ~ grp, data = temdata, FUN=mean, na.rm=TRUE)

uncaggfaster=temdata[, lapply(.SD, mean,na.rm=T), by = grp]

# Find max in a table
which.max(x)


# Update R
if(!require(installr)) {
  install.packages("installr"); require(installr)} #load / install+load installr
updateR()


# Create dummy variable from a factor
head(temdata)
for(level in unique(temdata$zp)){
  temdata[paste("dummy", level, sep = "_")] <- ifelse(temdata$zp == level, 1, 0)
}
```

```r
# Using semi colon to send multiple input
x=rnorm(10000,5,10)
mean(x);var(x);sqrt(var(x))

# Remove an object
y=rnorm(10)
rm(y)

# Empty the working space
rm(list=ls())

# Remove all but some
rm(list=setdiff(ls(),c("temdata", "temdata2")))


# Integer division
7%/%2

# Modulo = remainder
5%%2

# Define and print
(count=c(25,12,7,4,6,2,1,0,2))

# Read csv by clicking
data=read.csv(file.choose(),header=TRUE,)

#Combine more than 1 csv files
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)


#Multiple QQ plot
  #split screen
layout(matrix(1:9, nc = 3))
sapply(names(temdata)[1:9], function(x) {
  qqnorm(temdata[[x]], main = x)
  qqline(temdata[[x]])
})




#Split for more plots

par(mfrow=c(3,3))


#Double for 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
}


#Missing data
convert -999s to NAs

read.csv("x.csv", na.strings="-999")
temdata[is.na(temdata)] <- 0


#convert NAs to -99s

vector[which(vector== NA)]= (-99)
temdata[is.na(temdata)]= (-99)

#if you are having trouble converting <NA> (but not NA)
temdata=read.csv("temdata.csv",stringsAsFactors=FALSE)


# add group mean

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}


#sort a dataframe by the order of the elements in B
temdata[order(temdata$B),]

#sort the dataframe in reverse order
temdata[rev(order(temdata$B)),]


#create combinations
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))


#setting contrasts
options('contrasts')
options(contrasts=c('contr.sum','contr.poly'))
options(contrasts=c('contr.treatment','contr.poly'))


# delete if all NA
temdata=temdata[apply(temdata,1,function(x)any(!is.na(x))),]


# add group frequency
temdata=ddply(temdata, "grp", transform, cellsize = count(grp)[2])


#create new folder
dir.create("testdir")

#split data frame
```

```r
library(datasets)
head(airquality)
splitdata=split(airquality,airquality$Month)
splitdata
str(splitdata)
splitdata[[2]]


x=list(a=1:5, b=rnorm(10))
x
lapply(x,mean)

# output is always a list

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 generally used for rows or columns

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)
```

```
#missing data proportion percentage
propmiss <- function(temdata) lapply(temdata,function(x) data.frame(nmiss=sum(is.na(x)), n=length(x), pr
propmiss(temdata)


#upper case
temdata$childid=toupper(temdata$childid)



# plot graph individual all variables

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()


# read in upper directory
dd=read.csv("../temdata.csv")
```

12.1 More on the apaStyle package

Here is more details on the apaStyle package;

```
require(pastecs)
require(apaStyle)
library(rJava)
#if this throws an error
Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre1.8.0_111') # for 64-bit version

#define a data set

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
```

12.2 A useful shiny application

Below is a Shiny app example (Figure 12.2) to calculate sample size for an analyses of covariance design;

```
knitr::include_app('https://burakaydin.shinyapps.io/ancovaN/', height = '800px')
```

ANCOVA sample size calculator

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.

Appelbaum, M. I. and Cramer, E. M. (1976). Balancing - analysis of variance by another name. *Journal of Educational Statistics*, 1(3):233–252.

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.

Carlson, J. E. and Timm, N. H. (1974). Analysis of nonorthogonal fixed-effects designs. *Psychological Bulletin*, 81(9):563–570.

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;.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge;New York;.

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.

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.

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.