

FBES03 UYGULAMALI VERİ ANALİZİ ARA SINAV RAPORU

Melike BEKTAŞ - 20424011001

18.04.2021

Özet

İstatistiksel analiz yöntemleri shapiro, t-test ve Anova testi gibi test yöntemleri çoğu verinin analiz edilmesi için sıklıkla kullanılmaktadır. Bu test yöntemleri sonucunda elde edilen değerlerin yorumlanması önemli veri analizleri ve veri ilişkileri bulmamızı sağlamaktadır. Bu çalışmada farklı senaryolar üzerinden Shapiro, t-test ve Anova testi yapılmıştır. Elde edilen sonuç ve gözlemlere sonuçlar ve tartışma bölümünde yer verilmiştir.

1. GİRİŞ

Bu çalışmada ilgili veri setlerinin analiz edilebilmesi için istatistiksel test yöntemlerinden shapiro test, t-test ve Anova testi yapılmıştır. Bunlara ek olarak t-test ve Anova testine power analizi eklenmiş ve bunların sonuçları yorumlanmıştır. Shapiro test için öğrencilerin İngilizce sınav notlarının olduğu bir veri seti kullanılmıştır ve öğrencilerin almış oldukları sınav notlarının normal dağılıp dağılmadığı kontrol edilmiştir. T-testler'den one-sample ve two sample t-test için bir sarkacın belirli aralıklardaki uzunluk ölçümlerinin ortalamalarının istatistiksel olarak aynı olup olmadığı kontrol edilmiş, paired t-test için ise birbirileri ile ilişkili olan taban malzemelerinin ortalamaları istatistiksel olarak karşılaştırılmıştır. Anova testi için bir ilacın içerisinde bulunan A, B, C ve D maddelerinin kullanım miktarlarının Covid-19 hastalığına yakalanmış kişilerin iyileşmesi üzerindeki etkisi incelenmiş ve bu etken maddelerin ortalamalarının istatistiksel olarak aynı olup olmadığı incelenmiştir.

2. PROBLEMLERİN TANIMI

Problem 1: Öğrencilerin İngilizce sınavından almış oldukları puanların normal dağılıp dağılmadığının bulunması, normal dağılmıyor ise transforme edilmesi

```
ExamScores
<dbl>
0
39
75
79
85
90
.
```

Şekil 1. Bir grup öğrencinin İngilizce sınavından almış oldukları notlar

Problem 2: Bir sarkacın belirli aralıklardaki uzunluk ölçümlerinin ortalamalarının istatistiksel olarak aynı olup olmadığının tespit edilmesi

```
# A tibble: 6 x 2
  length time10periods
<dbl>         <dbl>
1     30           17.5
2     28           16.8
3     26           16.3
4     24           15.8
5     22           15.0
6     20           14.3
```

Şekil 2. Bir sarkacın belirli aralıklardaki uzunluğunun ölçüm değerleri

Problem 3: Birbirileri ile ilişkili olan taban malzemelerinin belirli tabanlar için kullanım miktarı ortalamalarının istatistiksel olarak karşılaştırılması

```
# A tibble: 6 x 2
  `soleMaterial_A` soleMaterial_B
<dbl>         <dbl>
1     13.2         14
2      8.2         8.8
3    10.9         11.2
4    14.3         14.2
5    10.7         11.8
6      6.6         6.4
```

Şekil 3. Birbirileri ile ilişkili olan taban malzemelerinin belirli tabanlar için kullanım miktarı

Problem 4: Bir ilacın içerisinde bulunan A, B, C ve D maddelerinin kullanım miktarlarının Covid-19 hastalığına yakalanmış kişilerin iyileşmesi üzerindeki etkisinin incelenmesi ve ortalamalarının istatistiksel olarak aynı olup olmadığının bulunması

```
# A tibble: 5 x 4
  A     B     C     D
<dbl> <dbl> <dbl> <dbl>
1  60.8  78.7  92.6  86.9
2   67   77.7  84.1  82.2
3  64.6  76.3  90.5  83.7
4  61.7  79.8  92.4  90.3
5   65   73.7  88.1  82.2
```

Şekil 4. Bir ilacın içerisinde bulunan A, B, C ve D maddelerinin farklı hastalar tarafından kullanım miktarları

3. METOT

Bu çalışmada bir grup öğrencinin İngilizce sınavından almış oldukları notların normal dağılıp dağılmadığı shapiro test ile, bir sarkacın belirli bir zaman aralığındaki uzunluk ölçümlerinin ortalamasının istatistiksel olarak aynı olup olmadığı two-sample t-test ve one-sample t-test ile, birbirileri ile ilişkili olan taban malzemelerinin kullanım miktarlarının ortalamalarının istatistiksel olarak aynı olup olmadığı paired t-test ile ve bir ilacın içerisinde bulunan A, B, C ve D maddelerinin farklı hastalar tarafından kullanım miktarlarının ortalamalarının istatistiksel olarak aynı olup olmadığı Anova ile test edilecektir.

Yapılan her bir test için null hipotez (H0) ve alternatif hipotez (HA) belirlenmiştir.

Shapiro test için;

H0: Ideal normal dağılım eğrisi ile bizim verimizin normal dağılım eğrisi arasında bir fark yoktur.

HA: Ideal normal dağılım eğrisi ile bizim verimizin normal dağılım eğrisi birbirinden farklıdır.

T-testler için;

H0: $\mu_1 - \mu_0 = 0$ (İki grup arasında istatistiksel olarak fark yoktur.)

HA: $\mu_1 - \mu_0 \neq 0$ (İki grup istatistiksel olarak farklıdır.)

Anova için;

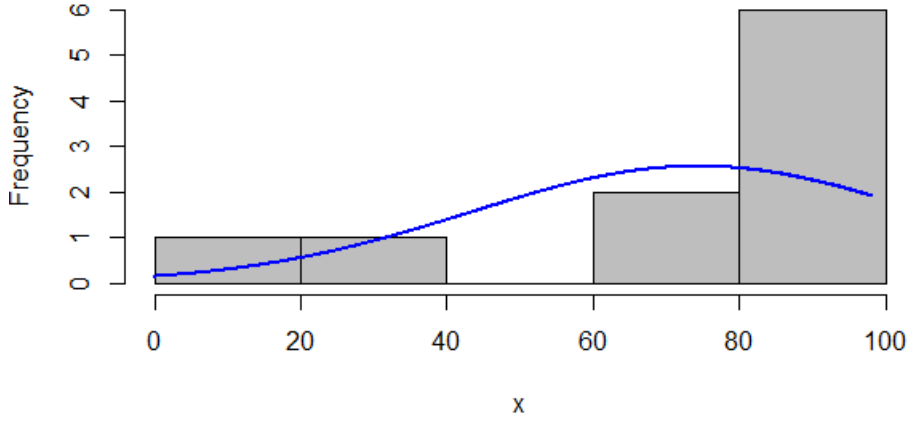
H0: $\mu_1 = \mu_2 = \mu_3 \dots = \mu_1$ (gruplar arasında fark yoktur)

HA: En az bir değişkenimizin ortalaması diğer grupların ortalamasından farklıdır.

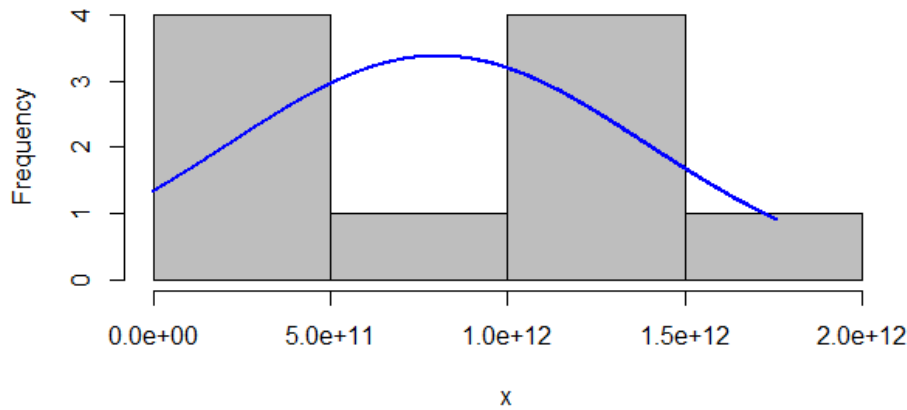
İncelemelere her bir veri setinin excel'den çekilmesi ile başlanmıştır. Yapılan istatistiksel analizler için R programlama dili kullanılmıştır. Verilerin excel dosyasından alınabilmesi için “readxl” kütüphanesinin yüklenerek verilerin erişilebilir olması için “attach” komutu

kullanılmıştır. Veriler attach yapıldıktan sonra her bir verinin histogram üzerinde normal dağılıp dağılmadığı kontrol edilmiştir.

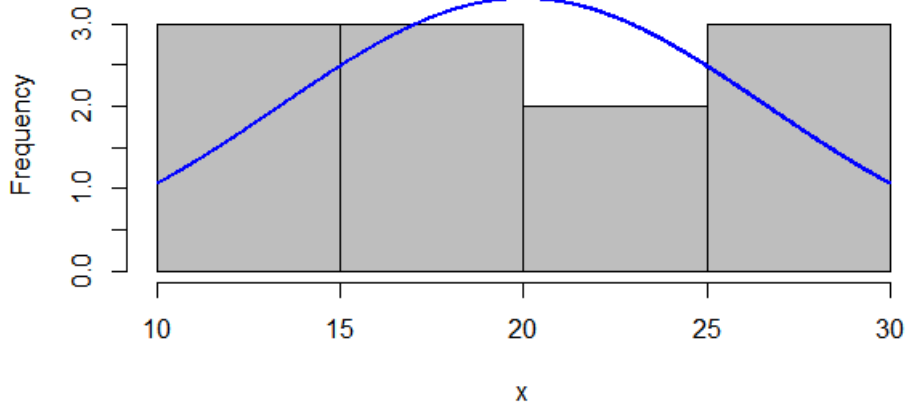
Shapiro test için;



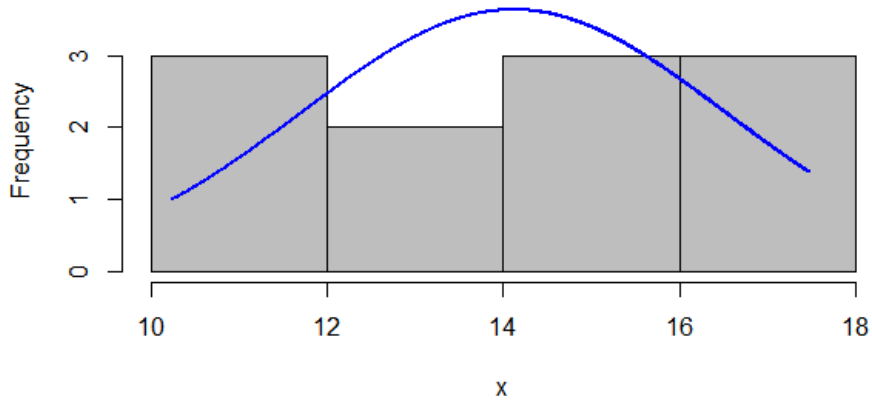
Şekil 5. Öğrencinin İngilizce sınavından almış oldukları notları gösteren dağılım eğrisi İngilizce notlarının normal dağılımını test etmek için shapiro testi uygulanmıştır. İngilizce sınav notlarının p-value = 0.001909 olarak bulunmuştur. P-value değeri 0.05’den küçük olduğu için “tansformTukey” komutu kullanılarak Veriler transforme edilmiştir.



Şekil 6. Öğrencinin İngilizce sınavından almış oldukları notlarının transforme edildikten sonraki normal dağılım eğrisi Transforme edildikten sonra p-value = 0.6128 olarak bulunmuştur. T-testler içinde veri setinde bulunan verilerin normal dağılıp dağılmadığının bulunması için shapiro test uygulanmıştır.



Şekil 7. T-test için length değerinin normal dağılım eğrisi
Length verilerinin p-value = 0.8698 olarak bulunmuştur.

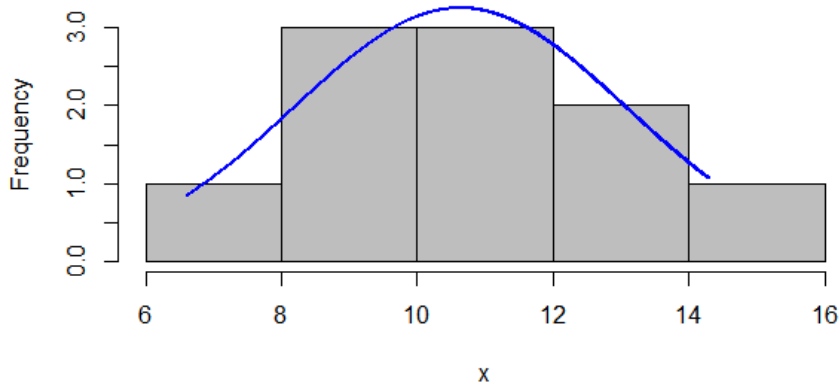


Şekil 8. T-test için time10periods değerinin normal dağılım eğrisi
time10periods verilerinin p-value = 0.7842 olarak bulunmuştur.

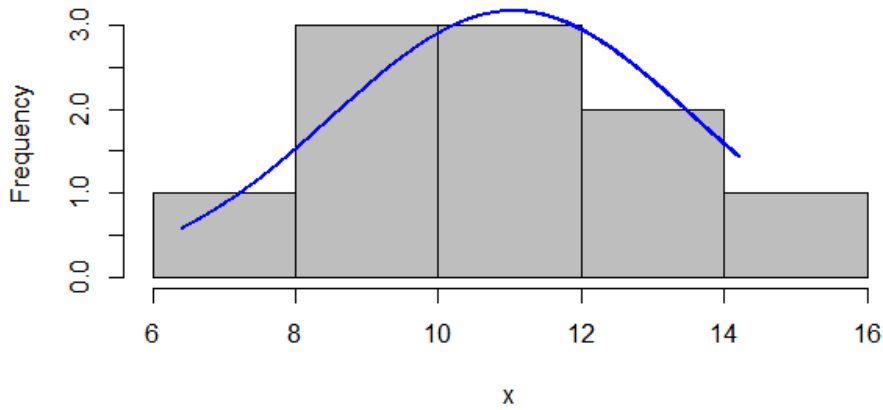
Her iki veri grubunda normal dağıldığı için t-test için uygundur. Bu aşamadan sonra varyans testi uygulanmıştır.

Varyans testi sonucunda p-value = 0.003643 bulunmuştur ve varyansların eşit olmadığı görülmüştür. Bu sebeple t-test'ler için `var.equal = FALSE` syntax'ı eklenmiştir.

Paired t-test için normal dağılım eğrileri;



Şekil 9. Paired t-test için SoleMaterial _A değerinin normal dağılım eğrisi
SoleMaterial _A verilerinin p-value = 0.8129 olarak bulunmuştur.

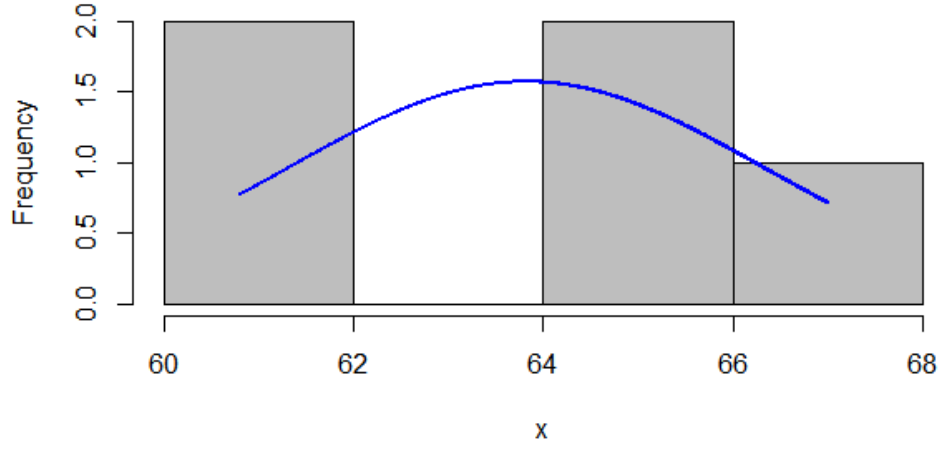


Şekil 10. Paired t-test için SoleMaterial _B değerinin normal dağılım eğrisi
SoleMaterial _B verilerinin p-value = 0.6467 olarak bulunmuştur.

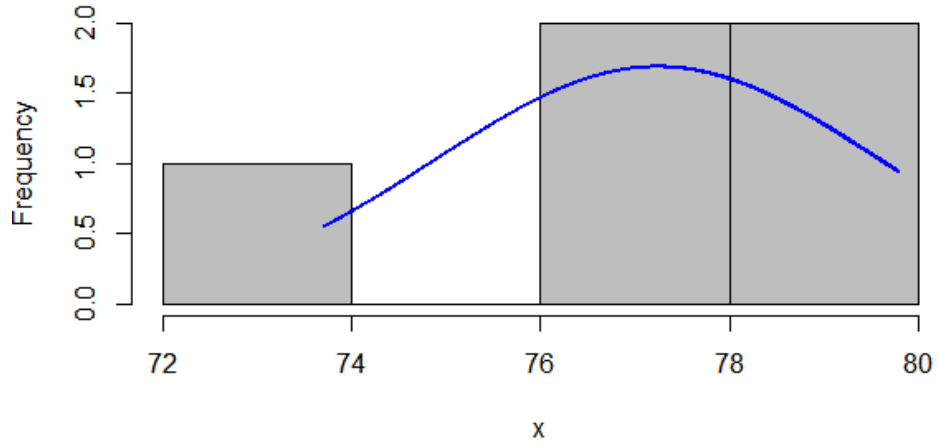
Her iki veri grubunda normal dağıldığı için t-test için uygundur. Bu aşamadan sonra varyans testi uygulanmıştır.

Varyans testi sonucunda p-value = 0.9372 bulunmuştur ve varyansların eşit olduğu görülmüştür. Bu sebeple t-test'ler için var.equal = TRUE syntax'ı eklenmiştir.

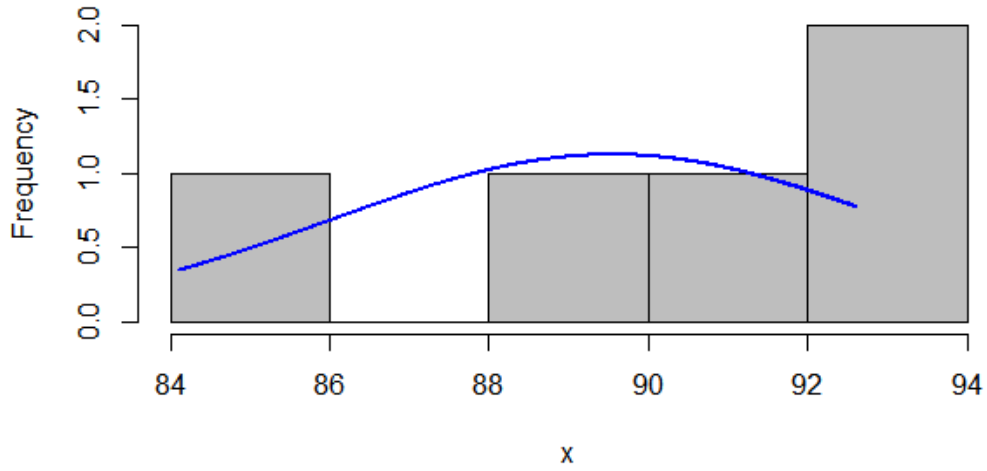
Anova testi içinde veri setinde bulunan verilerin normal dağılıp dağılmadığının bulunması için shapiro test uygulanmıştır.



Şekil 11. Anova testi için A değerinin normal dağılım eğrisi
A verilerinin p-value = 0.6646 olarak bulunmuştur.

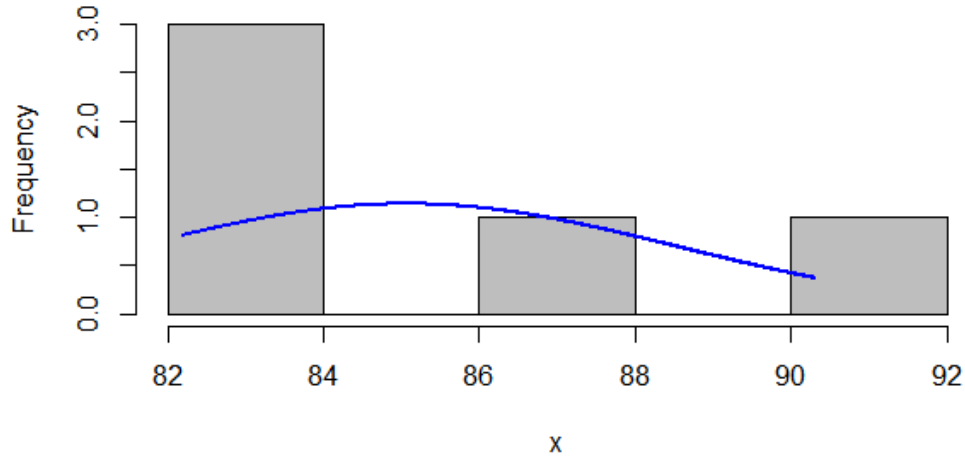


Şekil 12. Anova testi için B değerinin normal dağılım eğrisi
B verilerinin p-value = 0.8236 olarak bulunmuştur.



Şekil 13. Anova testi için C değerinin normal dağılım eğrisi

C verilerinin p-value = 0.3675 olarak bulunmuştur.



Şekil 14. Anova testi için D değerinin normal dağılım eğrisi

D verilerinin p-value = 0.251 olarak bulunmuştur.

A, B, C ve D verilerinin p-value değerleri 0.05'den büyük olduğu için normal dağılmışlardır ve Anova testi için uygundurlar. Bu aşamadan sonra varyans kontrolü yapılmıştır. A verilerinin standart sapması 2.536139, B verilerinin standart sapması 2.361779, C verilerinin standart sapması 3.540198, D verilerinin standart sapması ise 3.501857 olarak bulunmuştur. Elde edilen varyans değeri sonuçları veri setimizin Anova testi için uygun olduğunu göstermektedir.

R Console ekran çıktıları:

Shapiro.test

```
> shapiro.test(ExamScores)

      shapiro-wilk normality test

data:  ExamScores
W = 0.72767, p-value = 0.001909

> transforme_sinav <- transformTukey(ExamScores)

      lambda      w shapiro.p.value
647    6.15 0.9453          0.6128

if (lambda > 0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda < 0){TRANS = -1 * x ^ lambda}
```


t-test two sample

```
> mean(length)
[1] 20
> mean(time10periods)
[1] 14.11727
> shapiro.test(length)
```

shapiro-wilk normality test

```
data: length
W = 0.96839, p-value = 0.8698
```

```
> shapiro.test(time10periods)
```

shapiro-wilk normality test

```
data: time10periods
W = 0.96101, p-value = 0.7842
```

```
> var.test(length,time10periods)
```

F test to compare two variances

```
data: length and time10periods
F = 7.5624, num df = 10, denom df = 10, p-value = 0.003643
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.03465 28.10776
sample estimates:
ratio of variances
 7.56237
```

```
> t.test(length, time10periods, var.equal = FALSE)
```

welch Two sample t-test

```
data: length and time10periods
t = 2.7643, df = 12.599, p-value = 0.01649
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.270273 10.495182
sample estimates:
mean of x mean of y
20.00000 14.11727
```

```
> mptest <- pwr.t.test(d =(mean(length)-mean(time10periods))/sqrt(((sd(length)^2+sd(time10periods)^2))/2), power = 0.8, type = "two.sample")
> mptest
```

Two-sample t test power calculation

```
      n = 12.33728
      d = 1.178689
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

t-test one sample

```
> t.test(time10periods, mu=19.1)

One Sample t-test

data:  time10periods
t = -6.8512, df = 10, p-value = 4.452e-05
alternative hypothesis: true mean is not equal to 19.1
95 percent confidence interval:
 12.49679 15.73775
sample estimates:
mean of x
 14.11727

>
> pwr.t.test(n = 11 , d = 0.4, sig.level = NULL, power =0.8 , type = "one.sample")

one-sample t test power calculation

          n = 11
          d = 0.4
sig.level = 0.565658
  power = 0.8
alternative = two.sided
```

t-test paired

```
> shapiro.test(`soleMaterial _A`)

Shapiro-Wilk normality test

data:  soleMaterial _A
W = 0.9624, p-value = 0.8129

> shapiro.test(Sole_Material_B)

Shapiro-Wilk normality test

data:  Sole_Material_B
W = 0.94815, p-value = 0.6467

> var.test(Sole_Material_B, `soleMaterial _A`) #varyanslar eŝit

F test to compare two variances

data:  Sole_Material_B and soleMaterial _A
F = 1.0555, num df = 9, denom df = 9, p-value = 0.9372
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2621782 4.2495488
sample estimates:
ratio of variances
 1.055528
```

```
> t.test(`soleMaterial _A`, Sole_Material_B, paired = TRUE, var.equal = TRUE)

Paired t-test

data:  soleMaterial _A and Sole_Material_B
t = -3.3489, df = 9, p-value = 0.008539
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6869539 -0.1330461
sample estimates:
mean of the differences
      -0.41

> mptest <- pwr.t.test(d =(mean(`soleMaterial _A`)-mean(Sole_Material_B))/sqrt(((sd(`soleMaterial _A`)^2+sd(Sole_Material_B)^2))/2), power = 0.8, type = "paired")
> mptest
```

Paired t test power calculation

```
      n = 290.2862
      d = 0.1649818
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number of *pairs*

Anova Testi

```
> shapiro.test(A)
```

shapiro-wilk normality test

```
data:  A
W = 0.93981, p-value = 0.6646
```

```
> shapiro.test(B)
```

shapiro-wilk normality test

```
data:  B
W = 0.96226, p-value = 0.8236
```

```
> shapiro.test(C)
```

shapiro-wilk normality test

```
data:  C
W = 0.89204, p-value = 0.3675
```

```
> shapiro.test(D)
```

shapiro-wilk normality test

```
data:  D
W = 0.8661, p-value = 0.251
```

```
> sd(A)
[1] 2.536139
> sd(B)
[1] 2.361779
> sd(C)
[1] 3.540198
> sd(D)
[1] 3.501857
```

```

> bpdata<-stack(FEED_ANOVA)
> attach(bpdata)
The following objects are masked from bpdata (pos = 17):

    ind, values

> model<-values~ind
> library(car)
> leveneTest(model)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.2373  0.869
      16
> anova<-aov(model)
> summary(anova)
              Df Sum Sq Mean Sq F value    Pr(>F)
ind              3 1906.6    635.5    69.07 2.25e-09 ***
Residuals       16  147.2       9.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> myTukey<-TukeyHSD(anova)
> myTukey
    Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = model)

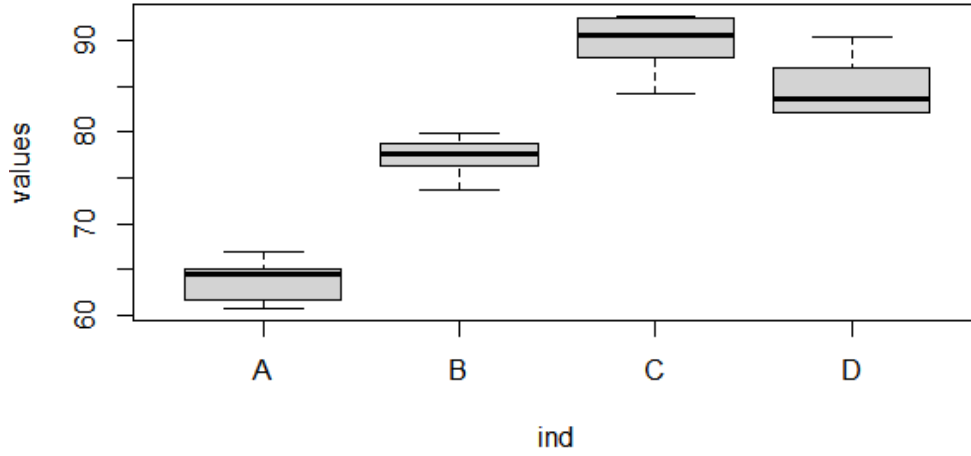
$ind
      diff      lwr      upr      p adj
B-A 13.42  7.931165 18.908835 0.0000164
C-A 25.72 20.231165 31.208835 0.0000000
D-A 21.24 15.751165 26.728835 0.0000000
C-B 12.30  6.811165 17.788835 0.0000464
D-B  7.82  2.331165 13.308835 0.0043873
D-C -4.48 -9.968835  1.008835 0.1313348
> pwr.anova.test(k = 4 , n = 5, f = 0.4, sig.level =NULL , power = 0.8)

    Balanced one-way analysis of variance power calculation

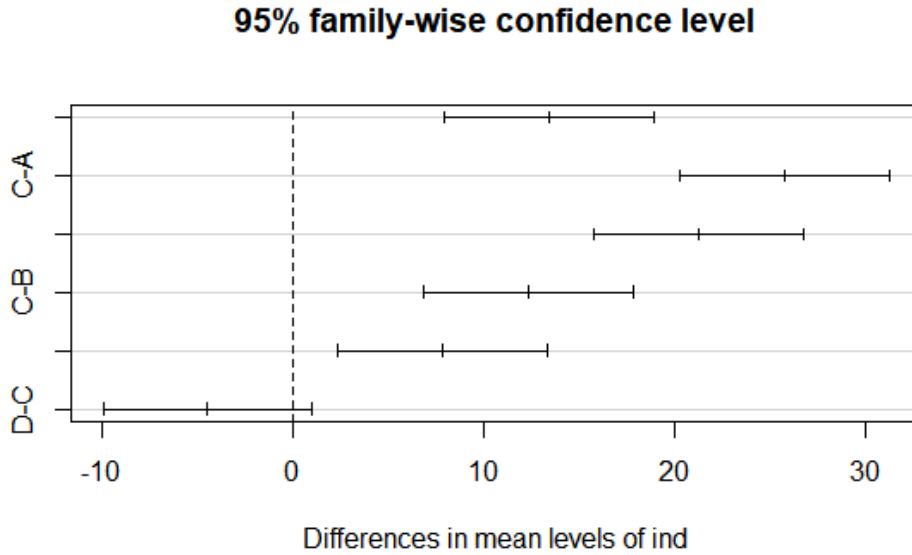
            k = 4
            n = 5
            f = 0.4
sig.level = 0.4985656
power = 0.8

NOTE: n is number in each group

```



Şekil 15. Anova testi için ortalama değerlerinin bar grafiği ile gösterimi



Şekil 16. Anova testi için ortalama değerlerinin TukeyHSD komutu ile gösterimi

4. SONUÇLAR VE TARTIŞMA

Bu çalışmada bir grup öğrencinin İngilizce sınav notlarının ortalamasının normal dağılıp dağılmadığı shapiro testi kullanılarak incelenmiştir. Test sonucunda $p\text{-value} = 0.001909$ olarak bulunmuştur. Elde edilen $p\text{-value}$ değeri 0.05’den küçük olduğu için veriler normal dağılmamıştır. Veriler transforme edildikten sonra $p\text{-value} = 0.6128$ bulunmuştur. Elde edilen yeni $p\text{-value}$ değerimiz 0.05’den büyük olduğu için ortalamamız verilerimizi temsil eder hale gelmiştir. Bir sarkacın belirli aralıklardaki uzunluk ölçümlerinin ortalamalarının istatistiksel olarak aynı olup olmadığının tespit edilmesi için two sample t-test yapılmıştır. Testin öncesinde verilerin normal dağıldığı fakat varyans testi sonucunda $p\text{-value} = 0.003643$

olduğu ve varyanslarının eşit olmadığı gözlemlenmiştir. Bu sebeple varyans eşitsizliği t-testin içerisinde $\text{var.equal} = \text{FALSE}$ olarak belirtilmiştir. Two sample t-test sonucunda elde edilen 0.01619 p-value değeri farklı zamanlarda ölçülen sarkaç uzunluklarının istatistiksel olarak birbirinden farklı olduğunu göstermiştir. Yapılan power analizi sonucunda ise 0.8 power elde edebilmek için olması gereken örnek sayısının 11 değil 12 olduğu gözlemlenmiştir. Aynı veri setinin bir ölçümünün sarkaç uzunluğu değeri ile 19.1 ortalama değeri one sample t-test kullanılarak karşılaştırılmış ve elde edilen 4.452×10^{-5} p-value değeri bu iki ortalamanın istatistiksel olarak birbirinden farklı olduğunu göstermiştir. Paired t-test için ise birbiri ile ilişkili olan taban malzemelerinin ortalama değerleri Shapiro ve varyans testinin ardından karşılaştırılmıştır. Shapiro testinin sonucunda verilerin normal dağıldığı, varyans testinden sonra ise varyanslarının eşit olduğu görülmüştür. T-testinin sonucunda ise 0.008539 p-value değeri ortalamaların istatistiksel olarak birbirinden farklı olduğunu göstermiştir. Yapılan power analizi sonucunda 0.8 power için olması gereken örnek sayısının 290 olduğu sonucunda ulaşılmıştır. Anova testi için ise bir ilacın içerisinde bulunan A, B, C ve D maddelerinin farklı hastalar tarafından kullanım miktarlarının istatistiksel olarak aynı olup olmadığı karşılaştırılmak istenmiştir. Bu test sonucunda p-value değeri 2.25×10^{-9} olarak bulunmuştur. Bu değer kullanılan A, B, C ve D maddelerinin ortalamalarının en az birinin istatistiksel olarak birbirinden farklı olduğunu göstermektedir. Modelin bar diyagramı incelendiğinde D ve C maddelerinin istatistiksel olarak birbiri ile aynı olduğu fakat A ve B maddelerinin miktarlarının istatistiksel olarak çok farklı olduğu sonucuna ulaşılmıştır. Anova testi power analizi sonucunda ise 0.4985656 significant level sonucuna ulaşılmıştır.