**TÜRKİYE CUMHURİYETİ**
**YILDIZ TEKNİK ÜNİVERSİTESİ**
**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**


# İstatistiksel Veri Analizi Ödev Raporu

17011604 – Havvanur Dervişoğlu

Danışman
Dr.Öğr.Üyesi Zeyneb KURT

## 1. Sağlıklı ve hastalıklı örneklerin her geni için standart sapma değeri ve %90 güven aralığındaki ortalama değeri:

**1-a:** "normal_skin_8genes_10samples.txt": Resim 1'de veri seti gösterilmektedir.

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| M7731 | 4.8397077 | 4.9357426 | 4.7402800 | 5.870636 | 0.02147973 | 0.03984026 | 4.1725675 | 0.000000 |
| M7874 | 2.2630344 | 2.3572705 | 2.2960168 | 5.212025 | 2.40408565 | 3.89549652 | 3.8309651 | 2.067295 |
| M7715 | 3.4121040 | 3.4220984 | 1.4451970 | 2.756169 | 1.24671210 | 2.47690008 | 2.8239531 | 0.000000 |
| M7785 | 0.0000000 | 0.0000000 | 1.8282254 | 3.184280 | 2.81003173 | 4.50639883 | 2.5112150 | 2.034216 |
| M5247 | 0.0000000 | 0.0000000 | 1.6735564 | 3.602766 | 3.09305305 | 4.70642004 | 2.6899700 | 0.000000 |
| M9000 | 0.0000000 | 0.0000000 | 5.1392239 | 6.284644 | 0.01006368 | 0.01435529 | 4.1525079 | 0.000000 |
| M7682 | 0.0000000 | 0.0000000 | 0.8463930 | 1.153805 | 3.71292588 | 5.31589223 | 1.2485348 | 2.238175 |
| M5305 | 0.2016339 | 0.2028878 | 0.8883049 | 2.447844 | 3.19408705 | 4.71182526 | 2.0352720 | 0.000000 |
| M5742 | 0.0000000 | 0.0000000 | 3.3593806 | 5.332242 | 0.01863417 | 0.04264434 | 4.2132695 | 0.000000 |
| M7767 | 0.0000000 | 0.0000000 | 0.8559897 | 1.624803 | 2.52907130 | 4.09744225 | 0.8890841 | 2.541019 |

**Resim 1:** normal_skin_8genes_10samples.txt görünümü

- Resim 2'de her gen değeri için standart sapma değeri görülmektedir. (gen_1:GAGE4,gen_2:GAGE5,gen_3:KRTAP7-1,gen_4:KRTAP8-1,gen_5:TSIX,gen_6:XIST,gen_7:KRTAP19-3,gen_8:MIR2117)

| | |
|---|---|
| gen_1 | 1.78708052335504 |
| gen_2 | 1.81811652325525 |
| gen_3 | 1.5872634865434 |
| gen_4 | 1.8209931668122 |
| gen_5 | 1.45027150323326 |
| gen_6 | 2.16470660719587 |
| gen_7 | 1.22387279546278 |
| gen_8 | 1.15427976419809 |

**Resim 2:** Her gen için standart sapma değerleri

- Her gen için %90 güven aralığında ortalama değeri: Şekil 3'de her gen için summary statistics değerleri görülmektedir.

```
> summary(data_df)
     GAGE4             GAGE5             KRTAP7-1          KRTAP8-1
 Min.   :0.000   Min.   :0.000   Min.   :0.8464   Min.   :1.154
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.0275   1st Qu.:2.525
 Median :0.000   Median :0.000   Median :1.7509   Median :3.394
 Mean   :1.072   Mean   :1.092   Mean   :2.3073   Mean   :3.747
 3rd Qu.:1.748   3rd Qu.:1.819   3rd Qu.:3.0935   3rd Qu.:5.302
 Max.   :4.840   Max.   :4.936   Max.   :5.1392   Max.   :6.285
      TSIX              XIST           KRTAP19-3         MIR2117
 Min.   :0.01006   Min.   :0.01435   Min.   :0.8891   Min.   :0.0000
 1st Qu.:0.32779   1st Qu.:0.65121   1st Qu.:2.1543   1st Qu.:0.0000
 Median :2.46658   Median :3.99647   Median :2.7570   Median :0.0000
 Mean   :1.90401   Mean   :2.98072   Mean   :2.8567   Mean   :0.8881
 3rd Qu.:3.02230   3rd Qu.:4.65641   3rd Qu.:4.0721   3rd Qu.:2.0590
 Max.   :3.71293   Max.   :5.31589   Max.   :4.2133   Max.   :2.5410
```

**Resim 3:** Her gen için istatistik değerleri

Resim 4'te her gen için standart error değerleri gösterilmiştir.(SE=s/sqrt(n))

| | |
|---|---|
| se_gen_1 | 0.565124481592766 |
| se_gen_2 | 0.574938926507307 |
| se_gen_3 | 0.501936786429717 |
| se_gen_4 | 0.575848601072948 |
| se_gen_5 | 0.458616117585336 |
| se_gen_6 | 0.684540334475438 |
| se_gen_7 | 0.387022559997977 |
| se_gen_8 | 0.365015311190804 |

**Resim 4:** Her gen için SE değerleri

Resim 5'de degree of fredom değeri 10-1=9 ve (1-0.9)%2=0.05 değeri için t-dist değeri gösterilmiştir. T dist. Değerinin kullanılmasının sebebi popülasyon standart sapma değerinin bilinmiyor olması elimizdeki 10 sample'dan elde ettiğimiz standart sapma değeri ile güven aralığını hesaplıycaz.

```
> qt(0.05,9)
[1] -1.833113
```

**Resim 5:** t-dist değeri

Resim 6'da her genin "t_dist*[s/sqrt(n)]" değeri görülmektedir.

| | |
|---|---|
| m_e_gen1 | 1.03593699576835 |
| m_e_gen2 | 1.05392798166804 |
| m_e_gen3 | 0.920106814580225 |
| m_e_gen4 | 1.05559551787882 |
| m_e_gen5 | 0.840695136270274 |
| m_e_gen6 | 1.25483974005175 |
| m_e_gen7 | 0.709456059962017 |
| m_e_gen8 | 0.669114287561404 |

**Resim 6:** Margin of error değerleri

Resim 7'da her bir gen için aralık değerleri görülmektedir. (aralık:[left_geni,right_geni], left_geni=mean_geni-m_e_geni, rigth_geni=mean_geni+m_e_geni şeklinde hesaplanmıştır.)

| | | | |
|---|---|---|---|
| left_gen1 | 0.0357110052347871 | right_gen1 | 2.10758499677149 |
| left_gen2 | 0.0378719508569667 | right_gen2 | 2.14572791419304 |
| left_gen3 | 1.38714995394189 | right_gen3 | 3.22736358310234 |
| left_gen4 | 2.69132585050148 | right_gen4 | 4.80251688625913 |
| left_gen5 | 1.06331929835129 | right_gen5 | 2.74470957089184 |
| left_gen6 | 1.72588177064501 | right_gen6 | 4.23556125074851 |
| left_gen7 | 2.14727785811528 | right_gen7 | 3.56618997803932 |
| left_gen8 | 0.218956193290334 | right_gen8 | 1.55718476841314 |

**Resim 7:** %90 güven aralığında mean değerleri

Resim 7'da her satır bir gen için aralık değerini göstermektedir.

**1-b:** “psoriasis_skin_8genes_10samples.txt” hastalıklı genler. Resim 8’de veri setine ait görüntü vardır.(Not: Veri setleri sütunlar genler ve satırlar örnekler olacak şekilde ayarlandı)

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| M5385 | 0.10433666 | 0.10433666 | 3.9549407 | 4.512164 | 0.021479727 | 0.042644337 | 3.8364291 | 1.3015876 |
| M4368 | 0.00000000 | 0.00000000 | 1.4388251 | 1.874600 | 1.954196310 | 3.359661722 | 2.8057052 | 1.3103401 |
| M4310 | 5.84962403 | 5.95891226 | 0.0000000 | 0.000000 | 0.020057652 | 0.045442971 | 0.0000000 | 0.0000000 |
| M8554 | 0.00000000 | 0.00000000 | 1.0628125 | 1.674913 | 2.601458821 | 4.121678557 | 1.1196882 | 1.0614306 |
| M4294 | 2.00934717 | 2.16124276 | 0.1724875 | 0.000000 | 2.910540801 | 4.330773498 | 0.0000000 | 0.0000000 |
| M777 | 0.00000000 | 0.00000000 | 0.9891390 | 1.088142 | 2.380452407 | 3.722028778 | 1.0250288 | 1.9414820 |
| M4298 | 0.00000000 | 0.00000000 | 0.0000000 | 1.060739 | 3.189350170 | 4.461986859 | 1.6154163 | 0.6590112 |
| M4390 | 0.05797007 | 0.05866284 | 0.0000000 | 0.000000 | 3.370931467 | 5.008854364 | 0.0000000 | 2.0450934 |
| M5478 | 0.00000000 | 0.00000000 | 2.8268027 | 4.275082 | 0.004321606 | 0.005759269 | 2.8771553 | 0.0000000 |
| M8530 | 0.00000000 | 0.00000000 | 0.0000000 | 1.016496 | 2.725741157 | 4.366881697 | 0.6544355 | 1.0454430 |

**Resim 8:** Hastalıklı genler veri seti

Resim 9’da her gen için ortalama değeri %90 güven aralığı görülmektedir. Her satır bir genin aralığını temsil ediyor. Şekil 10‘da ise her gen için 10 örnekten alınmış değerlerin ortalama değerleri görülmekte. Bu veri setinde de 1-a da yaptığımız adımlar yapılmıştır.

| | | | |
|---|---|---|---|
| left_gen1 | -0.288120463105785 | right_gen1 | 1.89237604923296 |
| left_gen2 | -0.287273768122186 | right_gen2 | 1.94390467101482 |
| left_gen3 | 0.249435642439814 | right_gen3 | 1.83956586712697 |
| left_gen4 | 0.598361462306014 | right_gen4 | 2.50206556229697 |
| left_gen5 | 1.12330782484159 | right_gen5 | 2.71239819896509 |
| left_gen6 | 1.75335124732808 | right_gen6 | 4.13979116308526 |
| left_gen7 | 0.60150385205629 | right_gen7 | 2.18526782749215 |
| left_gen8 | 0.493801731060318 | right_gen8 | 1.37907585131695 |

**Resim 9:** Her gen için %90 güven aralığı, ortalama değeri için

| | |
|---|---|
| mean_gen1 | 0.802127793063586 |
| mean_gen2 | 0.828315451446315 |
| mean_gen3 | 1.04450075478339 |
| mean_gen4 | 1.55021351230149 |
| mean_gen5 | 1.91785301190334 |
| mean_gen6 | 2.94657120520667 |
| mean_gen7 | 1.39338583977422 |
| mean_gen8 | 0.936438791188634 |

**Resim 10:** Her gen için ortalama değeri

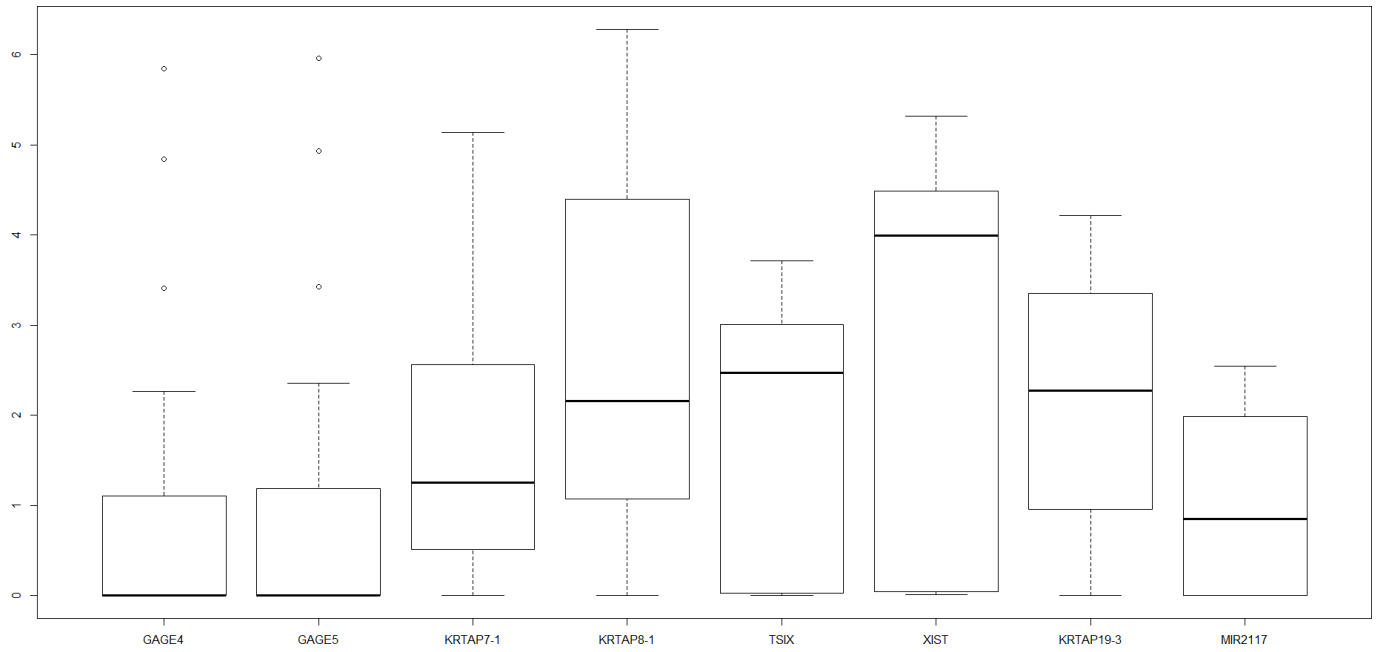
## 2. İki veri setini birleştirip boxplot elde edilecek. Her gen için bir boxplot.

Resim 11'de 8 gen için sağlıklı ve hastalıklı örneklerin birleştirmiş hali vardır.

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| M5385 | 0.10433666 | 0.10433666 | 3.9549407 | 4.512164 | 0.021479727 | 0.042644337 | 3.8364291 | 1.3015876 |
| M4368 | 0.00000000 | 0.00000000 | 1.4388251 | 1.874600 | 1.954196310 | 3.359661722 | 2.8057052 | 1.3103401 |
| M4310 | 5.84962403 | 5.95891226 | 0.0000000 | 0.000000 | 0.020057652 | 0.045442971 | 0.0000000 | 0.0000000 |
| M8554 | 0.00000000 | 0.00000000 | 1.0628125 | 1.674913 | 2.601458821 | 4.121678557 | 1.1196882 | 1.0614306 |
| M4294 | 2.00934717 | 2.16124276 | 0.1724875 | 0.000000 | 2.910540801 | 4.330773498 | 0.0000000 | 0.0000000 |
| M777 | 0.00000000 | 0.00000000 | 0.9891390 | 1.088142 | 2.380452407 | 3.722028778 | 1.0250288 | 1.9414820 |
| M4298 | 0.00000000 | 0.00000000 | 0.0000000 | 1.060739 | 3.189350170 | 4.461986859 | 1.6154163 | 0.6590112 |
| M4390 | 0.05797007 | 0.05866284 | 0.0000000 | 0.000000 | 3.370931467 | 5.008854364 | 0.0000000 | 2.0450934 |
| M5478 | 0.00000000 | 0.00000000 | 2.8268027 | 4.275082 | 0.004321606 | 0.005759269 | 2.8771553 | 0.0000000 |
| M8530 | 0.00000000 | 0.00000000 | 0.0000000 | 1.016496 | 2.725741157 | 4.366881697 | 0.6544355 | 1.0454430 |
| M7731 | 4.83970770 | 4.93574260 | 4.7402800 | 5.870636 | 0.021479727 | 0.039840265 | 4.1725675 | 0.0000000 |
| M7874 | 2.26303441 | 2.35727048 | 2.2960168 | 5.212025 | 2.404085654 | 3.895496519 | 3.8309651 | 2.0672945 |
| M7715 | 3.41210404 | 3.42209841 | 1.4451970 | 2.756169 | 1.246712101 | 2.476900083 | 2.8239531 | 0.0000000 |
| M7785 | 0.00000000 | 0.00000000 | 1.8282254 | 3.184280 | 2.810031728 | 4.506398832 | 2.5112150 | 2.0342157 |
| M5247 | 0.00000000 | 0.00000000 | 1.6735564 | 3.602766 | 3.093053048 | 4.706420039 | 2.6899700 | 0.0000000 |
| M9000 | 0.00000000 | 0.00000000 | 5.1392239 | 6.284644 | 0.010063683 | 0.014355293 | 4.1525079 | 0.0000000 |
| M7682 | 0.00000000 | 0.00000000 | 0.8463930 | 1.153805 | 3.712925879 | 5.315892234 | 1.2485348 | 2.2381754 |
| M5305 | 0.20163386 | 0.20288783 | 0.8883049 | 2.447844 | 3.194087052 | 4.711825257 | 2.0352720 | 0.0000000 |
| M5742 | 0.00000000 | 0.00000000 | 3.3593806 | 5.332242 | 0.018634174 | 0.042644337 | 4.2132695 | 0.0000000 |
| M7767 | 0.00000000 | 0.00000000 | 0.8559897 | 1.624803 | 2.529071300 | 4.097442248 | 0.8890841 | 2.5410192 |

**Resim 11:** Birleştirme matrisi

Resim 12'de her gen için boxplot vardır.




**Resim 12:** Boxplot

## 3. Korelasyon matrisi :

- **Sağlıklı veri seti için :** Resim 13'de gösterilmiştir.

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| GAGE4 | 1.0000000 | 0.9999064 | 0.3338204 | 0.3397682 | -0.4270771 | -0.3886911 | 0.4214235 | -0.2606302 |
| GAGE5 | 0.9999064 | 1.0000000 | 0.3366772 | 0.3450594 | -0.4256279 | -0.3876765 | 0.4250562 | -0.2553420 |
| KRTAP7-1 | 0.3338204 | 0.3366772 | 1.0000000 | 0.9199632 | -0.8656501 | -0.8904471 | 0.8555133 | -0.4824370 |
| KRTAP8-1 | 0.3397682 | 0.3450594 | 0.9199632 | 1.0000000 | -0.7902748 | -0.7997521 | 0.9634352 | -0.4922408 |
| TSIX | -0.4270771 | -0.4256279 | -0.8656501 | -0.7902748 | 1.0000000 | 0.9935357 | -0.7943990 | 0.5650919 |
| XIST | -0.3886911 | -0.3876765 | -0.8904471 | -0.7997521 | 0.9935357 | 1.0000000 | -0.7960535 | 0.5806983 |
| KRTAP19-3 | 0.4214235 | 0.4250562 | 0.8555133 | 0.9634352 | -0.7943990 | -0.7960535 | 1.0000000 | -0.5732245 |
| MIR2117 | -0.2606302 | -0.2553420 | -0.4824370 | -0.4922408 | 0.5650919 | 0.5806983 | -0.5732245 | 1.0000000 |

**Resim 13:** Sağlıklı genler veri seti korelasyon matrisi

- **Hastalıklı veri seti için:** Resim 14'de gösterilmiştir.

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| GAGE4 | 1.0000000 | 0.9998318 | -0.32817516 | -0.43044156 | -0.3974110 | -0.4125014 | -0.46604299 | -0.56146743 |
| GAGE5 | 0.9998318 | 1.0000000 | -0.33113189 | -0.43493810 | -0.3907263 | -0.4060869 | -0.47087863 | -0.56714854 |
| KRTAP7-1 | -0.3281752 | -0.3311319 | 1.00000000 | 0.95178107 | -0.6797860 | -0.6670427 | 0.88911646 | 0.03187441 |
| KRTAP8-1 | -0.4304416 | -0.4349381 | 0.95178107 | 1.00000000 | -0.6344229 | -0.6258330 | 0.92328528 | -0.04743878 |
| TSIX | -0.3974110 | -0.3907263 | -0.67978599 | -0.63442294 | 1.0000000 | 0.9948699 | -0.52555891 | 0.40949568 |
| XIST | -0.4125014 | -0.4060869 | -0.66704273 | -0.62583304 | 0.9948699 | 1.0000000 | -0.51189760 | 0.43938325 |
| RTAP19-3 | -0.4660430 | -0.4708786 | 0.88911646 | 0.92328528 | -0.5255589 | -0.5118976 | 1.00000000 | 0.05847581 |
| MIR2117 | -0.5614674 | -0.5671485 | 0.03187441 | -0.04743878 | 0.4094957 | 0.4393832 | 0.05847581 | 1.00000000 |

**Resim 14:** Hastalıklı genler veri seti korelasyon matrisi

- **Hastalıklı + sağlıklı için:** Resim 15'de gösterilmiştir.

| | GAGE4 | GAGE5 | KRTAP7-1 | KRTAP8-1 | TSIX | XIST | KRTAP19-3 | MIR2117 |
|---|---|---|---|---|---|---|---|---|
| GAGE4 | 1.00000000 | 0.99985850 | 0.04837480 | 0.01372367 | -0.4108532 | -0.3982195 | -0.01018420 | -0.3750647 |
| GAGE5 | 0.99985850 | 1.00000000 | 0.04626541 | 0.01143931 | -0.4067797 | -0.3946114 | -0.01375295 | -0.3744825 |
| KRTAP7-1 | 0.04837480 | 0.04626541 | 1.00000000 | 0.93577253 | -0.7149630 | -0.7161532 | 0.88749484 | -0.2778493 |
| KRTAP8-1 | 0.01372367 | 0.01143931 | 0.93577253 | 1.00000000 | -0.6002816 | -0.5933687 | 0.95428422 | -0.2825039 |
| TSIX | -0.41085323 | -0.40677966 | -0.71496302 | -0.60028160 | 1.0000000 | 0.9940692 | -0.56485650 | 0.4976741 |
| XIST | -0.39821952 | -0.39461142 | -0.71615324 | -0.59336871 | 0.9940692 | 1.0000000 | -0.55229648 | 0.5176052 |
| KRTAP19-3 | -0.01018420 | -0.01375295 | 0.88749484 | 0.95428422 | -0.5648565 | -0.5522965 | 1.00000000 | -0.2667652 |
| MIR2117 | -0.37506471 | -0.37448253 | -0.27784929 | -0.28250386 | 0.4976741 | 0.5176052 | -0.26676523 | 1.0000000 |

**Resim 15:** Hastalıklı + sağlıklı genler veri seti korelasyon matrisi

4. **HA: ”Hastalıklı geneX için ortalama değeri, sağlıklı aynı gen için ortalama değeri birbirinden farklıdır.”(sd: 1.5) Bu hipotezi sağlıklı ve hastalıklı veri setlerindeki her gen için test edip, p-değerini buluyoruz.**

- **GEN1:**

X1=mean value of hastalıklı geneX
X2=mean value of sağlıklı geneX
HA:x12=!0
H0:”GENX için her iki veri setinde ortalama değerleri birbirine eşittir.” X12=0

SD12=sqrt((sd/n1)+(sd/n2))=sqrt(2*1.5/10)=0.55
X12=0,27
Z_score=0,27/0,55=0,49

Pobs=2*P(Z>=abs(Z_score))=2*0.31=0.62

Sonuç significant value’lerden büyük olduğu için H0’ı reddemeyiz. Yani hastalıklı genx ortalaması ile sağlıklı genex ortalamaları birbirlerine benzerdir diyebiliriz.

Resim 16’da tüm genler için gen1’ yapılan işlemlerin sonucu gösterilmektedir.

| | |
|---|---|
| p_obs1 | 0.622666801698647 |
| p_obs2 | 0.630477660944584 |
| p_obs3 | 0.0211404602272276 |
| p_obs4 | 6.05591308422285e-05 |
| p_obs5 | 0.979843056712846 |
| p_obs6 | 0.950284396579766 |
| p_obs7 | 0.00754690203715447 |
| p_obs8 | 0.929631841933354 |

**Resim 16:** Her gen için p-value değerleri

Sonuç olarak gen3, gen4 ve gen7 için HA kabul edilirken(sağlıklı ve sağlıksız genx ortalama değerleri birbirlerinden farklıdır.).Ama diğer için genler için H0 reddedilemez yani HA reddedilir.(alpha=0.1)