Big data Statistics

Efthymios Ioannis Kavour

May 21, 2023

Contents

1 Exercise 1

Open R and obtain the Leukemia dataset from the *leukemiasEset* package in *Bioconductor*.

```
library("leukemiasEset")
data(leukemiasEset)
x <- exprs(leukemiasEset)
```

The dataset (x) contains expresion data for 20172 genes from 60 bone marrow samples of patients with one of the four main types of leukemia:

- ALL: Acute Lymphoblastic Leukemia
- AML: Acute Myeloid Leukemia
- CLL: Chronic Lymphocytic Leukemi
- CML: Chronic Myeloid Leukemia
- NoL: non-Leukemia

There are 12 samples per class, which can be retrieved using the command

```
leukemiasEset\$LeukemiaType
```

Let j denotes the last digit of your student identification number. We are interested to test which genes are diferentially expressed between the condition c and NoL groups, where

- $c =$ ALL if $j \leq 2$
- $c =$ AML if $3 \leq j \leq 5$
- $c =$ CLL if $6 \leq j \leq 7$
- $c =$ CML if $8 \leq j \leq 9$

So your dataset should consists of a matrix with 20172 rows (gene expression measurements) and 24 columns (12 replicates for each one of the two experimental groups).

1. Explore and visualize the data. Focus on the research question and try to visually describe the variability of the average gene expression between the two groups. Produce some meaningful summaries and descriptive statistics for your dataset.

2. Use PCA in order to visualize the dataset

$$(20172 \times 24)$$

. Project the data on the first few principal components and explain your findings. Do the same when considering the transposed input data (24×20172). Describe what you see.

3. Use two independent samples t-tests (you may assume that the variance is equal between groups) in order to test the null hypothesis per gene. State the null and alternative hypothesis per gene, as well as the assumptions you use to model the data. Plot a histogram (relative frequencies) of the p-values.

4. Can you give a rough estimate of the proportion of true null hypotheses?

5. Report how many genes are diferentially expressed when controlling the $FWER$, FDR and $pFDR$ at $\alpha = 0.01$.

6. Visualize the results obtained in question 5 according to whether the corresponding hypothesis is rejected or not when controlling the FDR at 0.01:

    a Plot a meaningful summary of the data and colour the genes depending on the result of the test (Diferentially Expressed or not Diferentially Expressed when controlling the FDR at the given level). Try to take into account both the mean difference as well the standard deviation per gene. Be creative.

    b using Principal Components projections and explain your findings.

2 Solution

2.1 Exploratory Data Analysis

To begin with the exercise, proceed to the solution and produce some meaningful results, we need to load the libraries and the data required for the needs of this analysis. As a result, we are going to use the following libraries:

- leukemiasEset
- tidyverse
- factoextra
- FactoMineR

Those libraries contain the data useful functions for our analysis as well as (for *tidyverse* package) dependencies to other packages like *ggplot2*, *dplyr*, *tidyr* which will help us manipulate data with the usage of pipes, or produce better plots. Next we load the data *leukemiaEset* as stated in the exercise definition. As my student number ends with 4 (P3622114), as instructed, I selected from the dataset, Acute Myeloid type of Leukemia (*AML*) to proceed with. The first thing one should do when coming across dataset that has never seen before is proceed to Exploratory Data Analysis (*EDA*) in order to understand data as good as possible. Initially, we can see that our dataset has the dimensions of 20172 rows and 24 columns. In the rows are all the different genes, and the columns are difference subjects take from. The initial 12 columns are samples that have been taken from subjects with AML whereas the later 12 are taken from health samples. Next, we calculate the averages per column, and use a plot to see if there is any difference amongst the subjests.

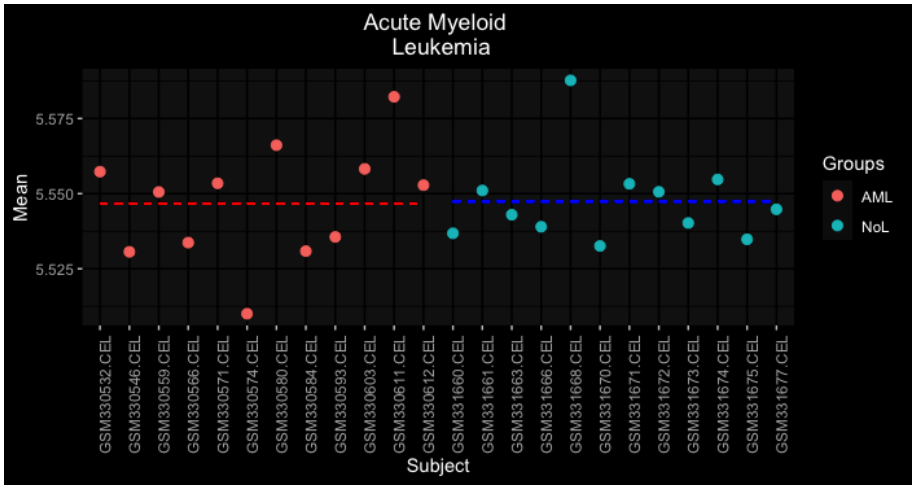


Figure 1: Means of AML and NoL subject

As we can see the means are not that different between the two groups, but we can see tow of the subjects specifically, **GSM330611.CEL** and **GSM331668.CEL** are far away from the rest of their group (AML or NoL), as a result it is worth creating a box plot in order to check for outliers.

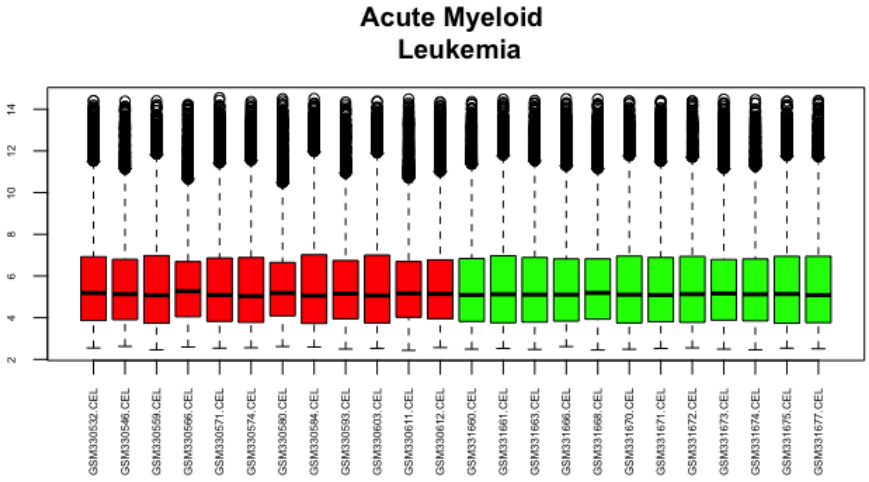


Figure 2: Boxplot of AML and Nol subjects

We see the same result as before, with the addition of the outliers that appear in the upper part of the boxplot. We observe a significant number of outliers located above the upper whisker. These outliers represent data points that deviate significantly from the majority of the values in the dataset. They indicate the presence of potential extreme or unusual observations.

Though it is always important to further investigate these outliers to understand their nature and potential impact on an analysis, we are not going to take that path, as it is out of our scope. In general the existence of outliers may vary. Some of the most frequent reasons for outliers' existence are measurement errors, data entry mistakes, or genuinely rare occurrences.

Given the results, provided above, we decided to continue with a density plot where we are going to take a look at the distribution of the measurements (means) of the subjects.

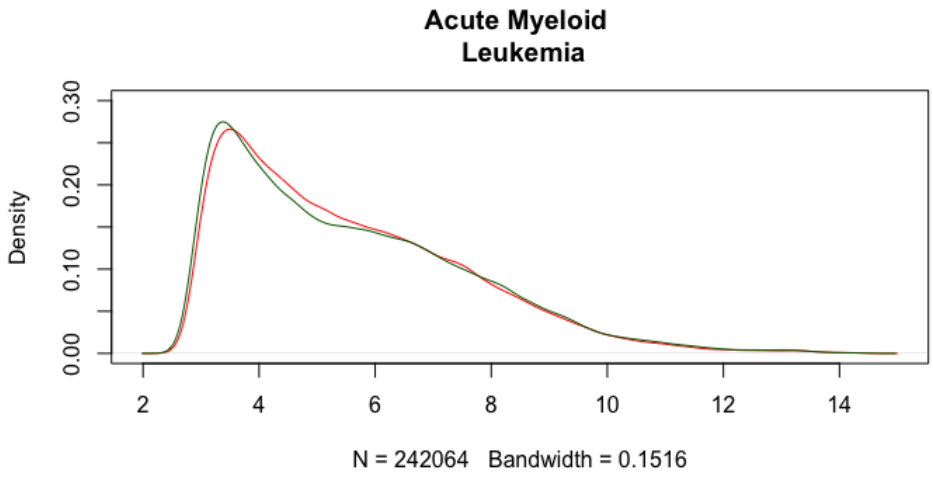


Figure 3: Density plot of AML and NoL

We can see that both groups' means are right skewed, which tells us that the tail of the distribution extends towards higher values, while the majority of the data points are concentrated towards the left or lower values. The mode of the most frequent mean tends to be smaller than the mean and the median per group which is located as shown above towards the left side of the plot where the density is higher. Of course, the median, is smaller than the mean of the values. This is due to the previous fact stated (longer tail towards higher values). Overall using the *summary()* function we can take a look at the details provided earlier:

```
> summary(x[,1:12])
 GSM330532.CEL     GSM330546.CEL     GSM330559.CEL     GSM330566.CEL
 Min.   : 2.549   Min.   : 2.624   Min.   : 2.459   Min.   : 2.592
 1st Qu.: 3.867   1st Qu.: 3.905   1st Qu.: 3.735   1st Qu.: 4.047
 Median : 5.183   Median : 5.132   Median : 5.077   Median : 5.272
 Mean   : 5.557   Mean   : 5.531   Mean   : 5.551   Mean   : 5.534
 3rd Qu.: 6.923   3rd Qu.: 6.801   3rd Qu.: 6.979   3rd Qu.: 6.690
 Max.   :14.407   Max.   :14.384   Max.   :14.403   Max.   :14.234
 GSM330571.CEL     GSM330574.CEL     GSM330580.CEL     GSM330584.CEL
 Min.   : 2.539   Min.   : 2.560   Min.   : 2.618   Min.   : 2.595
 1st Qu.: 3.823   1st Qu.: 3.777   1st Qu.: 4.090   1st Qu.: 3.725
 Median : 5.091   Median : 5.037   Median : 5.185   Median : 5.056
 Mean   : 5.553   Mean   : 5.510   Mean   : 5.566   Mean   : 5.531
 3rd Qu.: 6.866   3rd Qu.: 6.887   3rd Qu.: 6.650   3rd Qu.: 7.020
 Max.   :14.556   Max.   :14.363   Max.   :14.500   Max.   :14.512
 GSM330593.CEL     GSM330603.CEL     GSM330611.CEL     GSM330612.CEL
 Min.   : 2.496   Min.   : 2.530   Min.   : 2.433   Min.   : 2.567
 1st Qu.: 3.945   1st Qu.: 3.746   1st Qu.: 4.022   1st Qu.: 3.951
 Median : 5.153   Median : 5.065   Median : 5.164   Median : 5.142
 Mean   : 5.536   Mean   : 5.558   Mean   : 5.582   Mean   : 5.553
 3rd Qu.: 6.744   3rd Qu.: 7.006   3rd Qu.: 6.696   3rd Qu.: 6.776
 Max.   :14.347   Max.   :14.387   Max.   :14.490   Max.   :14.352
```

Figure 4: Summary of AML group

```
> summary(x[,13:24])
 GSM331660.CEL     GSM331661.CEL     GSM331663.CEL     GSM331666.CEL
 Min.   : 2.491   Min.   : 2.527   Min.   : 2.478   Min.   : 2.616
 1st Qu.: 3.821   1st Qu.: 3.755   1st Qu.: 3.790   1st Qu.: 3.837
 Median : 5.092   Median : 5.121   Median : 5.107   Median : 5.107
 Mean   : 5.537   Mean   : 5.551   Mean   : 5.543   Mean   : 5.539
 3rd Qu.: 6.840   3rd Qu.: 6.969   3rd Qu.: 6.888   3rd Qu.: 6.820
 Max.   :14.362   Max.   :14.475   Max.   :14.423   Max.   :14.500
 GSM331668.CEL     GSM331670.CEL     GSM331671.CEL     GSM331672.CEL
 Min.   : 2.453   Min.   : 2.485   Min.   : 2.525   Min.   : 2.555
 1st Qu.: 3.930   1st Qu.: 3.748   1st Qu.: 3.807   1st Qu.: 3.777
 Median : 5.189   Median : 5.105   Median : 5.090   Median : 5.130
 Mean   : 5.588   Mean   : 5.533   Mean   : 5.553   Mean   : 5.551
 3rd Qu.: 6.822   3rd Qu.: 6.957   3rd Qu.: 6.888   3rd Qu.: 6.939
 Max.   :14.489   Max.   :14.386   Max.   :14.407   Max.   :14.390
 GSM331673.CEL     GSM331674.CEL     GSM331675.CEL     GSM331677.CEL
 Min.   : 2.491   Min.   : 2.461   Min.   : 2.539   Min.   : 2.510
 1st Qu.: 3.884   1st Qu.: 3.850   1st Qu.: 3.734   1st Qu.: 3.763
 Median : 5.162   Median : 5.123   Median : 5.136   Median : 5.085
 Mean   : 5.540   Mean   : 5.555   Mean   : 5.535   Mean   : 5.545
 3rd Qu.: 6.792   3rd Qu.: 6.817   3rd Qu.: 6.946   3rd Qu.: 6.945
 Max.   :14.482   Max.   :14.458   Max.   :14.384   Max.   :14.421
```

Figure 5: Summary of NoL group

2.2 Principal Component Analysis

Now that we have an idea how our data look like, we can move forward to our next goal. Since our dataset has that big of dimentions, we are going to use Principle Component Analysis (*PCA*) in order to see if the reduction of the dimentionality is possible. With no further ado, we use the function *prcromp*, and as a results, we get the following results.

```
> summary(r_pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6
Standard deviation     4.7057 0.72364 0.51931 0.41897 0.35899 0.32773
Proportion of Variance 0.9226 0.02182 0.01124 0.00731 0.00537 0.00448
Cumulative Proportion  0.9226 0.94446 0.95570 0.96301 0.96838 0.97286
                           PC7     PC8    PC9    PC10   PC11    PC12    PC13
Standard deviation     0.31295 0.30409 0.2544 0.24430 0.2297 0.21116 0.20335
Proportion of Variance 0.00408 0.00385 0.0027 0.00249 0.0022 0.00186 0.00172
Cumulative Proportion  0.97694 0.98079 0.9835 0.98598 0.9882 0.99003 0.99176
                          PC14    PC15    PC16    PC17    PC18    PC19
Standard deviation     0.18491 0.17285 0.15395 0.14495 0.13404 0.12706
Proportion of Variance 0.00142 0.00124 0.00099 0.00088 0.00075 0.00067
Cumulative Proportion  0.99318 0.99443 0.99541 0.99629 0.99704 0.99771
                          PC20    PC21    PC22    PC23    PC24
Standard deviation     0.12216 0.11652 0.10285 0.09599 0.08166
Proportion of Variance 0.00062 0.00057 0.00044 0.00038 0.00028
Cumulative Proportion  0.99833 0.99890 0.99934 0.99972 1.00000
```

Figure 6: Summary of PCA

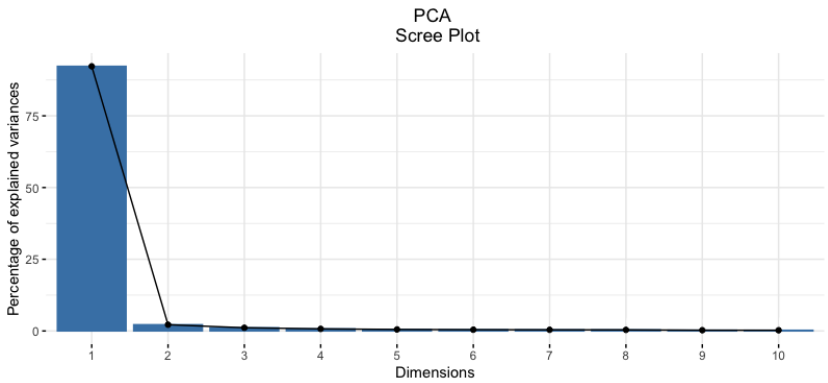


Figure 7: Scree Plot

In Figure 7, the x-axis shows the pricipal components (dimensions), which, in our case, are 10. The y-axis shows the prcentage of the explained variance per principle component. In order to interpret this we are going to use the "elbow method" and state that at the second principal component and beyond the "curve" is flatten. This means that only the first principal component should be take under consideration for this analysis.

Moving forward, we can see the projection of the genes on PCA dimentions. We can notice that the genes that are similar are grouped by potition. Apart from that we can take a look at the projection of subjects on PCA dimensions in Figure 9.

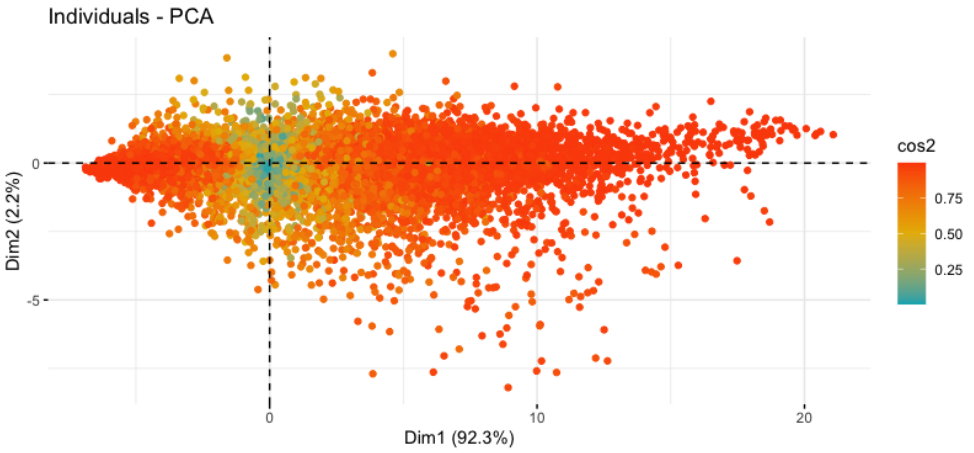


Figure 8: Genes projection on PCA1 and PCA2

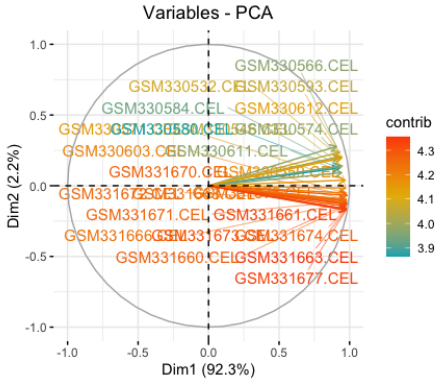


Figure 9: Subjects projection on PCA1 and PCA2

We can see that the genes spread across the dimension of PCA1, whereas the genes seem to all be positive correlated, and are spread in the the right side of PCA1. Moving forward we are going to repeat this process as for the transpose of our data. This means that we will make rows columns and make columns rows. Let us take a closer look at the results provided below. Initially, in Figure 10 the Scree plot of the transposed data are provided. Following the same method, now, it seems that more than just the first dimensions are need to be taken under consideration if we were to continue our analysis using the transposed dataset. To be more specific, the first 4 dimensions, it seems to me that are important to be noted.

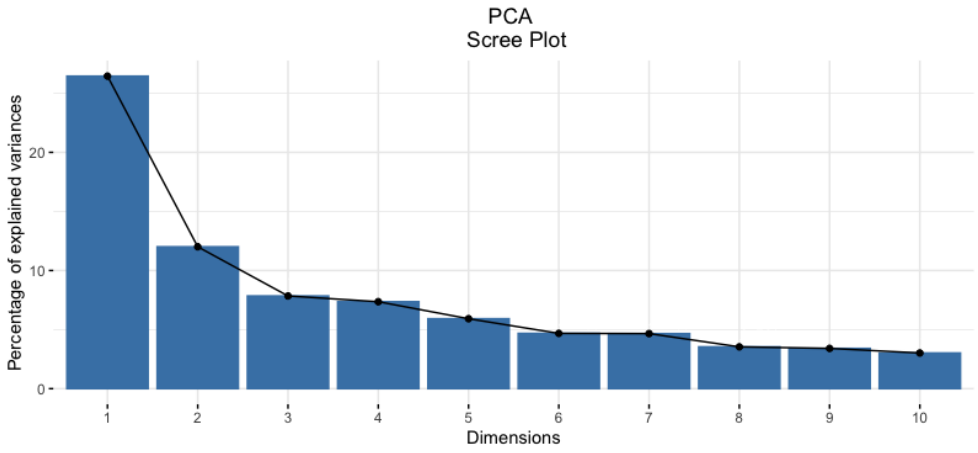


Figure 10: Scree Plot of transposed data

The aforementioned notice about the number of dimensions need to full explain the variability is confirmed by the following data where we should have more than just two or even three dimensions to full understand what is the contribution per gene.

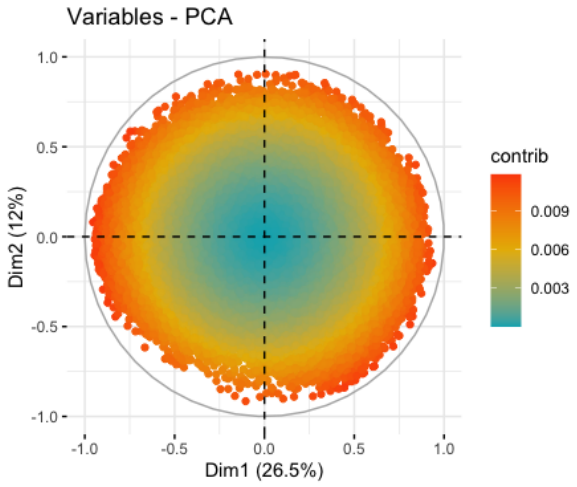


Figure 11: Scree Plot of transposed data

As well as the projection of individuals is displayed in the next plot we can see that the quality of representation for the variables is higher as for most of the variables that are located further from the center of the PC-axis.

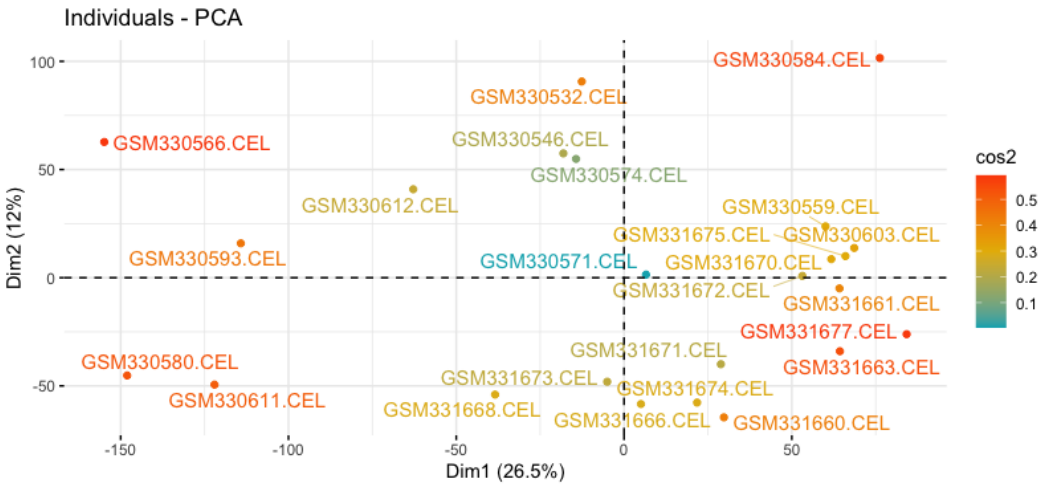


Figure 12: Scree Plot of transposed data

2.3 T-test - Mean similarity

In this section we aim to figure out whether there exist great similarity in the means among the two groups we study, the subjects with Leukemia (AML) and the subjects that are healthy (Nol). In order to figure this out we are going to use **t-test** with the following hypothesis:

H_0: $\mu_{i,1} = \mu_{i,2} \quad \forall\, i = \{1, \ldots, 20172\}$

H_1: $\exists i$ such that $\mu_{i,1} \neq \mu_{i,2}$

It is assumed that the variance is equal between groups and that all the samples are indipendent and identically distributed (iid). In general that the assumptions needed to implement a t-test (Independence, Normality, Homogeneity of Variance) are met. Following we can take a look at the histogram generated

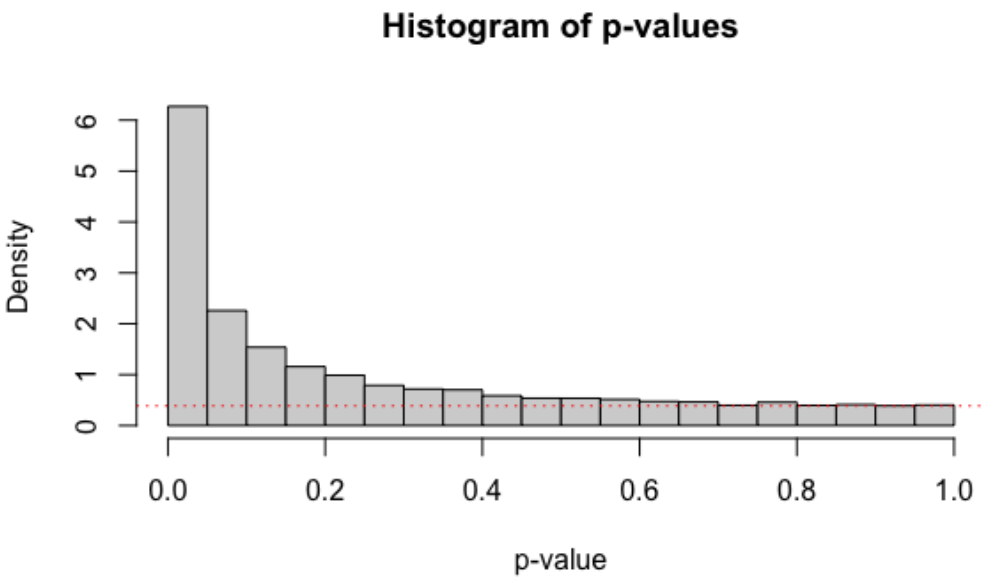


Figure 13: Histogram of p-values of the two-sample t-tests

Where the dotted red line is a rough estimation of the proportion of true null hypotheses. This value is found with the help of the function *qvalue::pi0est* and has the value of approximately 0.39. We can clearly see that after a certain p-value, the histogram tends to be flattened. As a result we can say that the Null hypothesis is established there and that there is no significant differentiation within the two grops.

2.4 Test statistics and p value adjust(ments)

It is time to investigate the number of genes that are differentially expressed within the groups (AML, NoL). All the tests are set to take place for $\alpha = 0.01$. The tests we are going to use are **Family-Wise Error Rate** (*FWER*), **False Discovery Rate** (*FDR*), **Positive False Discovery Rate** (*pFDR*). Our results are presented below:

FWER	92
FDR	953
pFDR	1687

We can clearly the difference amongst the three methods where FWER discovers only 92 genes whereas pFDR manages to find 1687.

2.5 FDR - Further Analysis

Here we adjust our data and add the information acquired by the FDR method as to which genes are differentially expressed amongst the two groups. Moreover, we are going to create some plots in order to investigate whether there is a correlation to the mean differences (per gene, per group) and finally try out Pearson correlation test in order to see if there is any correlation among the rejection-issue and difference of means. Enough with the explanation though. Initially let us remind ourselves what FDR results:

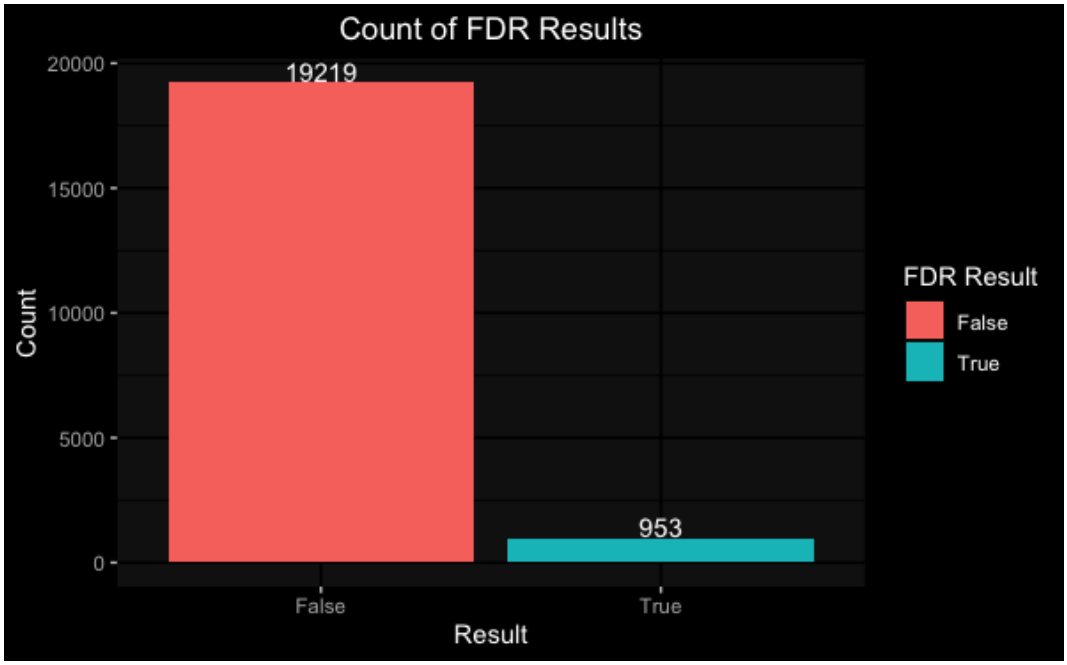


Figure 14: FDR Results

Moving forward using the base *apply* function we managed to find all the mean differences per gene among the 2 groups studied. The result is the following.

Figure 15: T-Test Results per Mean Difference

Figure 16: T-Test Results per Mean Difference

With the help of both those plots, we can say that the results that tend to

be rejected are those with AML group gene mean greatly different from NoL group gene mean per gene.

Moving forward, we will repeat the process for standard deviance per gene per group.

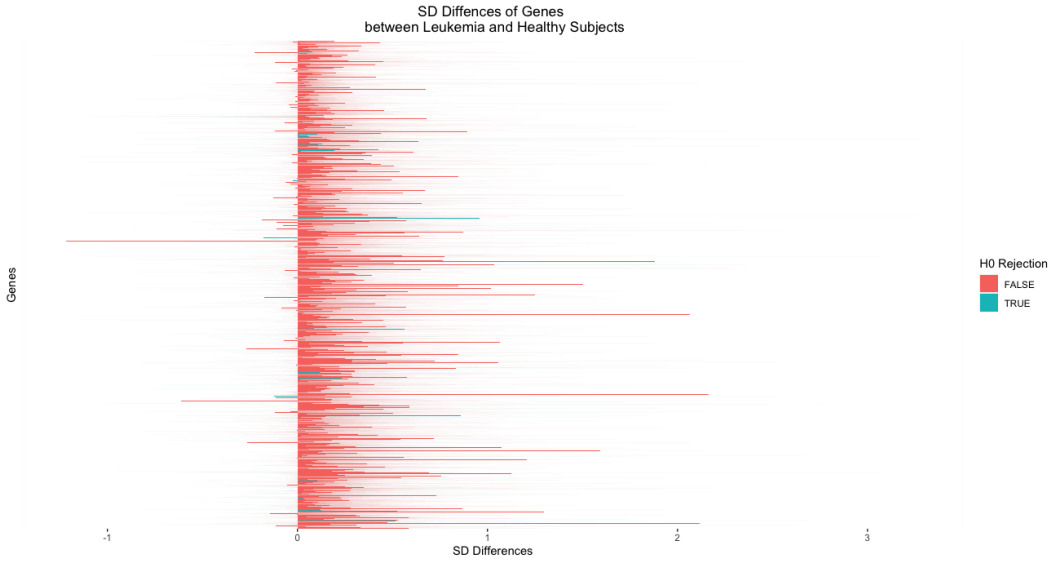


Figure 17: T-Test Results per gene's SD

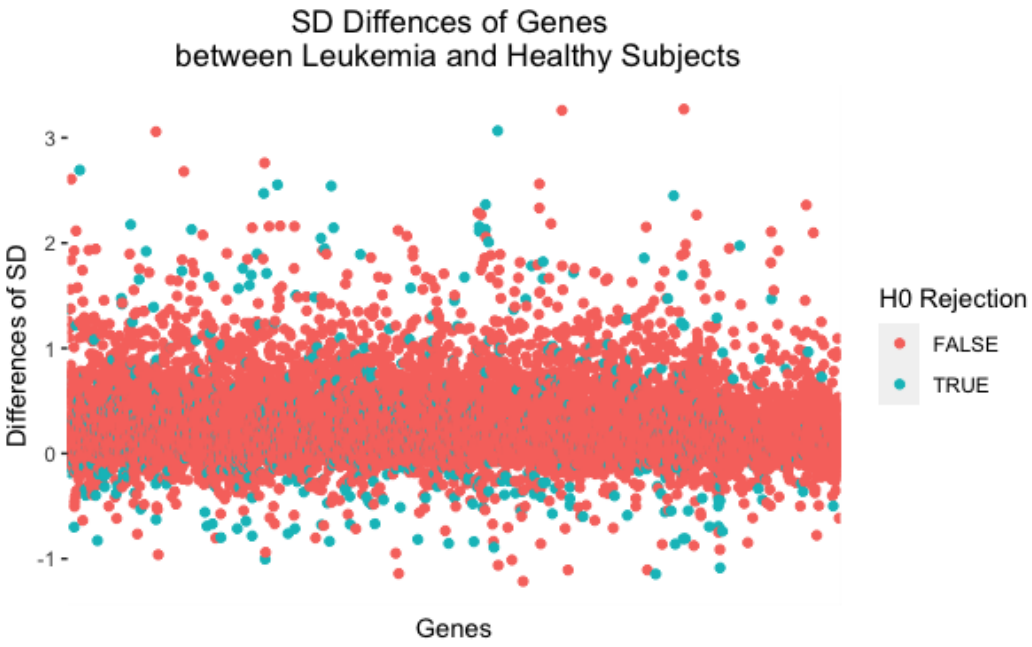


Figure 18: T-Test Results per gene's SD

We can clear see that what is stated above does not apply for this case as well. Gene's sd differences seem to locate within each other (False and True of H_0 rejection).

Case	Pearson Cor
Mean Differences vs H_0 Rejection	-0.4022856
SD Differences vs H_0 Rejection	0.02485962

Finally, we are going to present some statistics of the PCA1 on whether the Null hypothesis is rejected or not.

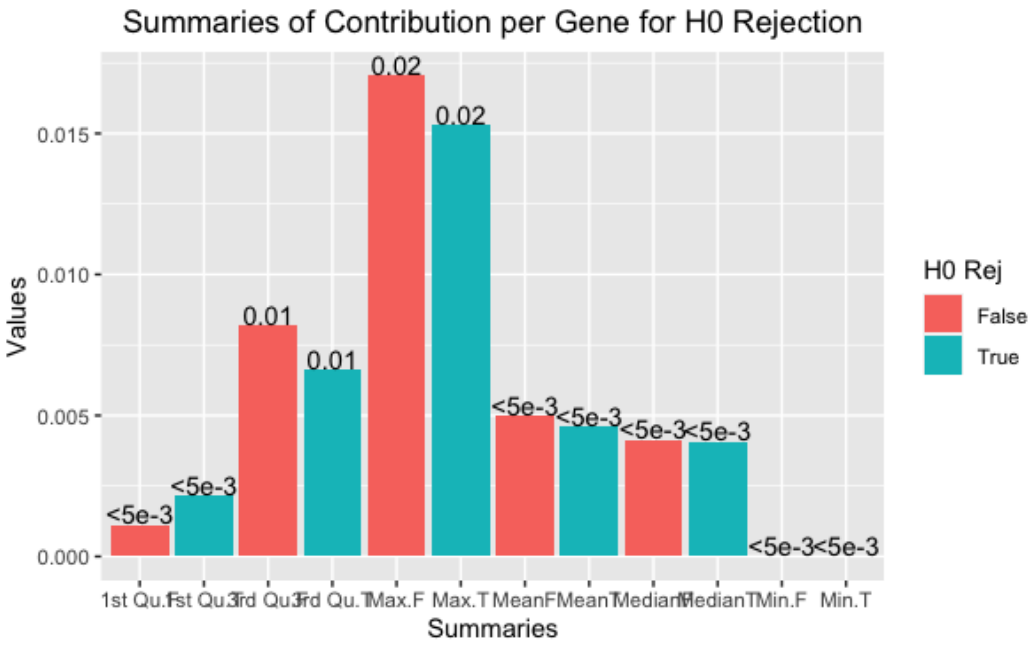


Figure 19: PCA grouped summary

We can see that the rejected genes tend to have less contribution to PCA1 for all statistics produced by *summary* function though the difference is not great to conclude to a concrete conclution. This can be shown as well by the result of Pearson correlation test.

Case	Pearson Cor
Contribution vs H_0 Rejection	-0.01897729

3 Exercise 2

Simulate a synthetic dataset from a normal linear model with $n = 500$ observations and $p = 100$ explanatory variables, as follows:

1. Simulate the explanatory variables from independent normal distributions:

```
x <- matrix(rnorm(n*p),nrow = n, ncol = p)
```

2. Generate the p regression coefficients β1, . . . , βp as follows: b ¡- numeric(p)

```
if( runif(1) < 0.3){ b[1] <- rnorm(1) }
```

This means that $\beta_i = 0$ for all $i \geq 2$, while the first coefficient (β_1) is zero with probability 0.7, while it is different than zero with probability

3. Generate the values of the response variable from a typical normal linear model, that is,

```
y <-  x%*%b + rnorm(n)
```

Repeat Steps 1, 2, 3 for $m = 10000$ times (so you will generate 10000 regression datasets). For each synthetic dataset we are interested to test the hypothesis that the response variable is not linearly depending on any of the p explanatory variables, that is,

$$H_0(j) : \beta_1 = ... = \beta_p = 0 \text{ vs } H_1(j) : \beta_i \neq 0 \text{ for at least one } i = 1, 2, \ldots, p$$

for $j = 1, \ldots, m$. Apply the standard F-test for this purpose. Recall that the p-value of the F-test is returned in the *summary()* method of the *lm()* command. You should extract the p-values for each one of the 10000 synthetic datasets. Since you are generating the data, you know which null hypotheses are true or not. Test all 10000 hypotheses and control the type I error rate using all methods (c("**holm**", "**hochberg**", "**hommel**", "**bonferroni**", "**BH**", "**BY**")) described in the *p.adjust()* command of R, as well as the *q-value*.

1. Report a confusion matrix per method with respect to the ground-truth, when controlling the relevant type I error at the $\alpha\alpha = 0.05$ level. What is the estimated power (proportion of true discoveries with respect to the total number of non-true null hypotheses) for this target value ($\widehat{power}(0.05)$), per method?

2. Plot the points $(\alpha, \widehat{power}(\alpha))$ for a sequence of values $\alpha \in (0, 1)$, that is, the estimated power versus the type I error control-value, for each method (see the relevant plots in the slides of Unit 1). Comment on the ranking of methods.

Advice: be gentle to your machine. There is no need to save 10000 simulated datasets. All you need is the vector of 10000 p-values and the ground-truth per tested hypothesis.

4 Solution

In this case, as stated above, we are going to generate random data under some circumstances, and then try out several methods, and see the results generated per method. The process of generating the data, and the methods we are going to use are stated above. Here we are going to present the confution matrix per method, and finally a plot with which we are going to choose the best possible method for the process discribed above. As a result, we have the following:

- Bonferroni

	FALSE	TRUE
FALSE	6818	146
TRUE	680	2356

- Benjamini Hochberg

	FALSE	TRUE
FALSE	6873	91
TRUE	715	2321

- Holm

	FALSE	TRUE
FALSE	6964	0
TRUE	1037	1999

- Hochberg

	FALSE	TRUE
FALSE	6964	0
TRUE	1037	1999

- Hommel

	FALSE	TRUE
FALSE	6964	0
TRUE	1037	1999

- Benjamini Yekutieli

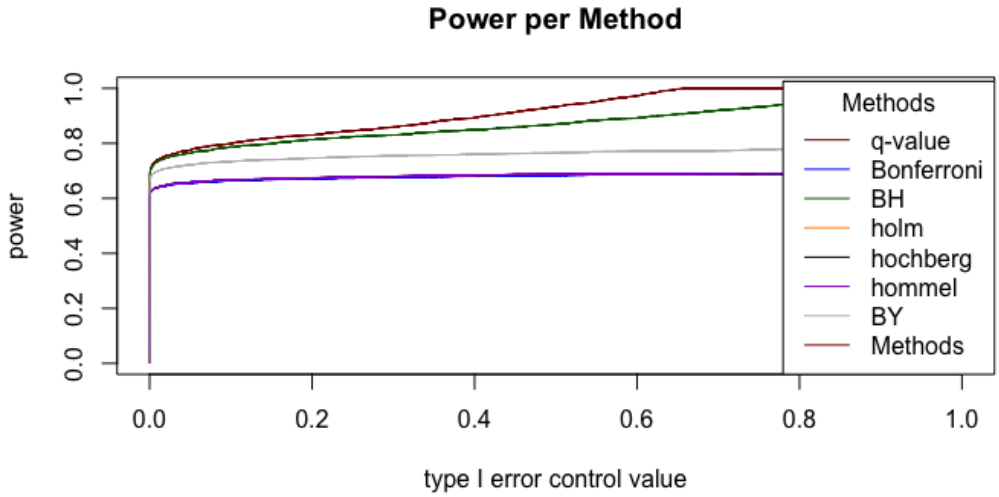
	FALSE	TRUE
FALSE	6956	8
TRUE	843	2193

- q-Value

	FALSE	TRUE
FALSE	6818	146
TRUE	680	2356

Finally, we can see here how every method works along different values of alpha. We also provide the power of every method.

Method	Power
Bonferroni	0
BH	0.0130672
Holm	0
Hochberg	0
Hommel	0
BY	0.001148765
qvalue	0.02096496



We can see that qvalue has the higher power from both, the plot (for various alphas as well as the straight forward table) provided above.