Biostatistics Assignment

Presented to :

Dr / Ibrahim Mohamed Ibrahim

Name	Section	BN
Alaa Tarek Samir	1	12
Amira Gamal Mohamed	1	15
Salma Haitham Fathy	1	38
Nouran Khaled Soliman	2	40

Introduction

In this paper we illustrate our work in conducting data Analysis on gene expression data of two sets one of them for healthy tissues and the other one is for cancerous tissues. We first built some insights about the data sets by compute the correlation coefficient between healthy sample and cancerous for each gene to get the highest positive and the lowest negative among them. As the correlation shows to what degree are a pair of variables correlated / linearly related.

As well as we conducted a hypothesis testing for two assumptions to determine how close are, they to be true.

- assumption#1 (the genes don't change when diseased)
- assumption#2 (the genes change when diseased)

The assumptions are applied on two datasets (paired and independent)

Methods

* Reading Datasets:

We used pandas software library to read the two text files into pandas.dataframe using **pd.read_csv** , the data sets had two indexing columns so we used the names of the genes as our indexing columns and dropped the genes id column form the dataframes.

* Exploring and Cleaning data

We started off with exploring the datasets, then cleaning and filtering our data by checking each row in both files if a row had more than 25 columns (total number of data columns is 50) with the value zero then we drop the whole row, to make sure that the rows in both files are balanced we started with the data for healthy dataset then we deleted that rows that justify the condition in both files and repeated the same procedure for the cancer dataset, we ended up with 17337 gene.

* Computing Correlation between the normal samples and the diseased samples

In this stage we started off by computing spearman's cc to gain some insights about montonicity between the data we used **scipy.stats.spearmanr** .Then for assessing linearity between the datasets we used **scipy.stats.pearsonr** to compute pearson's correlation coefficient between healthy sample and cancerous for each gene to find the then we sorted the coefficients ascendingly and we plotted a scatterplot of the expression levels for both genes using **matplotlib.pyplot.**

* Hypothesis Testing

We have two hypothesizes the null hypothesis is that gene expression value doesn't change in case of healthy tissue or cancerous tissue. The alternative hypothesis is that the gene expression value does change in case of being diseased.

We used **st.ttest_rel** to compute the p value for each gene in both cases (paired or independent) we choose alpha = 0.05 (confidence level= 95%) which means that out of total number of genes 5%

will be false positive so we used the FDR to control the number of false positives using **statsmodels.stats.multitest.multipletests**

To check whether we are going to accept the null hypothesis or reject it we compare the p value before and after FDR with alpha/2 = 0.025:

If P value < 0.025 then the null hypothesis is rejected on the other hand if P value > 0.025 then the null hypothesis is accepted.

in order to get the DEG we search for the areas where we reject the null hypothesis.

Results and Discussion:

The files had originally 19648 rows and 52 columns after cleaning and filtering of the data sets, we had 17337 rows.

```
[3]: df_h.info()

<class 'pandas.core.frame.DataFrame'>
Index: 17337 entries, HIST3H2A to FUT2
Data columns (total 50 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   TCGA-43-7657   17337 non-null  float64
 1   TCGA-58-8386   17337 non-null  float64
 2   TCGA-22-5478   17337 non-null  float64
 3   TCGA-22-5472   17337 non-null  float64
 4   TCGA-43-5670   17337 non-null  float64
 5   TCGA-60-2709   17337 non-null  float64
 6   TCGA-22-5489   17337 non-null  float64
 7   TCGA-77-8007   17337 non-null  float64
 8   TCGA-22-5471   17337 non-null  float64
 9   TCGA-22-4609   17337 non-null  float64
 10  TCGA-22-5482   17337 non-null  float64
 11  TCGA-56-8082   17337 non-null  float64
 12  TCGA-22-5483   17337 non-null  float64
 13  TCGA-56-8623   17337 non-null  float64
 14  TCGA-33-4587   17337 non-null  float64
 15  TCGA-56-7579   17337 non-null  float64
 16  TCGA-43-3394   17337 non-null  float64
 17  TCGA-34-8454   17337 non-null  float64
 18  TCGA-77-7338   17337 non-null  float64
 19  TCGA-43-6143   17337 non-null  float64
```

```
df_h.head(10)
```

Hugo_Symbol	TCGA-43-7657	TCGA-58-8386	TCGA-22-5478	TCGA-22-5472	TCGA-43-5670	TCGA-60-2709	TCGA-22-5489	TCGA-77-8007	TCGA-22-5471	TCGA-22-4609	...	TCGA-85-7710	TCGA-56-7580	TCGA-43-6647	TCGA-90-6837	TCGA-56-8083	TC
HIST3H2A	62.12	130.60	33.06	35.50	73.03	60.39	92.05	66.65	54.33	15.56	...	90.77	59.55	40.07	22.92	29.91	82
LIN7B	185.11	283.05	119.26	169.07	165.57	161.02	131.51	198.47	175.07	147.06	...	185.11	119.26	102.97	123.50	264.03	194
LXN	909.17	819.30	412.00	743.43	1340.84	607.87	1709.26	1709.26	603.67	555.41	...	813.63	2400.97	543.96	2193.99	540.19	521
CNKSR2	41.81	18.29	40.93	67.12	54.72	29.27	20.26	23.76	28.04	39.22	...	34.51	70.01	57.49	57.89	67.12	34
SCML1	133.36	214.27	108.14	109.66	190.34	211.31	96.01	208.38	120.10	239.52	...	251.48	209.84	120.10	109.66	155.50	162
AC024592.12	25.91	32.13	17.38	13.72	21.01	28.86	23.59	34.02	36.01	22.75	...	37.32	20.56	32.13	28.86	11.30	18
GSDMD	1733.13	2835.70	1508.65	1936.53	1819.35	2502.97	2090.03	2434.50	1757.34	1844.76	...	2720.15	1950.00	2018.80	2004.85	2417.67	2209
AKR1C1	1088.92	947.83	684.02	860.08	1096.50	2090.03	1369.04	1151.06	1119.56	703.28	...	511.00	660.68	1051.79	743.43	961.07	1242
C3orf62	122.64	181.28	181.28	150.17	183.82	117.60	106.63	150.17	100.83	112.77	...	194.36	183.82	120.94	84.63	130.60	166
CRISPLD2	848.22	536.45	620.67	1216.75	1832.01	2287.20	693.58	836.53	312.00	1674.06	...	277.20	579.04	2956.17	1233.75	1769.57	1781

```
df_c.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 17337 entries, HIST3H2A to FUT2
Data columns (total 50 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   TCGA-43-7657  17337 non-null  float64
 1   TCGA-58-8386  17337 non-null  float64
 2   TCGA-22-5478  17337 non-null  float64
 3   TCGA-22-5472  17337 non-null  float64
 4   TCGA-43-5670  17337 non-null  float64
 5   TCGA-60-2709  17337 non-null  float64
 6   TCGA-22-5489  17337 non-null  float64
 7   TCGA-77-8007  17337 non-null  float64
 8   TCGA-22-5471  17337 non-null  float64
 9   TCGA-22-4609  17337 non-null  float64
 10  TCGA-22-5482  17337 non-null  float64
 11  TCGA-56-8082  17337 non-null  float64
 12  TCGA-22-5483  17337 non-null  float64
 13  TCGA-56-8623  17337 non-null  float64
 14  TCGA-33-4587  17337 non-null  float64
 15  TCGA-56-7579  17337 non-null  float64
 16  TCGA-43-3394  17337 non-null  float64
 17  TCGA-34-8454  17337 non-null  float64
 18  TCGA-77-7338  17337 non-null  float64
 19  TCGA-43-6143  17337 non-null  float64
 20  TCGA-43-6773  17337 non-null  float64
 21  TCGA-51-4080  17337 non-null  float64
```

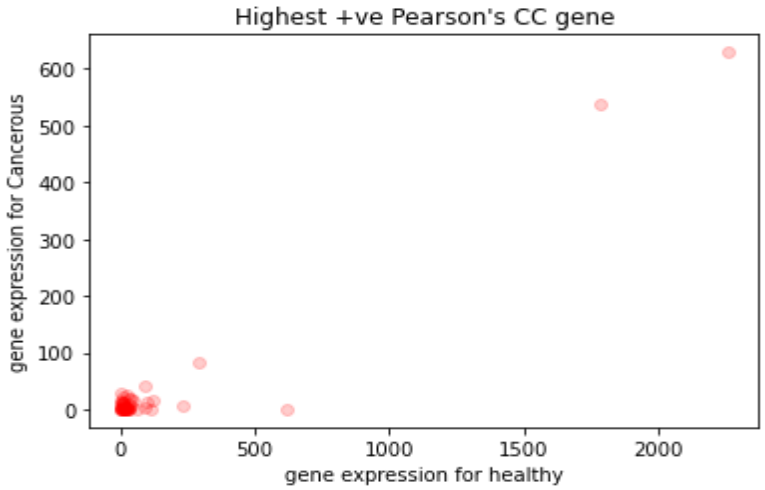
```
df_c.head(10)
```

Hugo_Symbol	TCGA-43-7657	TCGA-58-8386	TCGA-22-5478	TCGA-22-5472	TCGA-43-5670	TCGA-60-2709	TCGA-22-5489	TCGA-77-8007	TCGA-22-5471	TCGA-22-4609	...	TCGA-85-7710	TCGA-56-7580	TCGA-43-6647	TCGA-90-6837	TCGA-56-8083
HIST3H2A	336.79	500.46	703.28	287.01	486.75	70.51	145.02	14.03	397.93	318.57	...	3.06	420.68	109.66	106.63	1233.75
LIN7B	105.15	212.78	102.25	212.78	172.65	244.57	105.89	152.28	258.57	218.79	...	135.24	135.24	151.22	395.18	295.11
LXN	848.22	236.21	271.48	759.08	61.25	620.67	329.84	599.49	587.13	638.15	...	688.78	204.07	438.59	503.95	3039.30
CNKSR2	32.59	8.51	45.85	6.16	49.21	11.91	12.27	15.00	1.38	8.71	...	1.38	6.62	6.11	1.66	33.54
SCML1	84.63	74.58	67.12	57.89	102.97	132.44	66.65	57.08	336.79	171.45	...	165.57	119.26	87.65	53.57	232.94
AC024592.12	17.13	25.91	16.88	63.45	27.84	23.08	46.50	27.05	38.95	36.53	...	8.58	32.13	32.59	21.63	5.19
GSDMD	1551.09	1427.22	1674.06	1685.71	3124.78	2133.97	2451.44	2240.11	1467.37	1477.58	...	2075.59	1135.20	1832.01	1208.34	1883.54
AKR1C1	9945.68	723.08	1023.00	1242.34	136.19	40621.74	660.68	84.04	5366.37	1111.82	...	884.29	1175.27	1143.10	656.11	334.46
C3orf62	82.29	111.21	59.97	100.83	98.04	112.77	52.08	137.14	79.45	76.17	...	62.56	129.69	48.18	167.90	80.01
CRISPLD2	162.14	297.17	518.15	220.32	185.11	1015.93	1059.11	591.22	241.19	1709.26	...	480.04	262.20	2319.15	1023.00	287.01

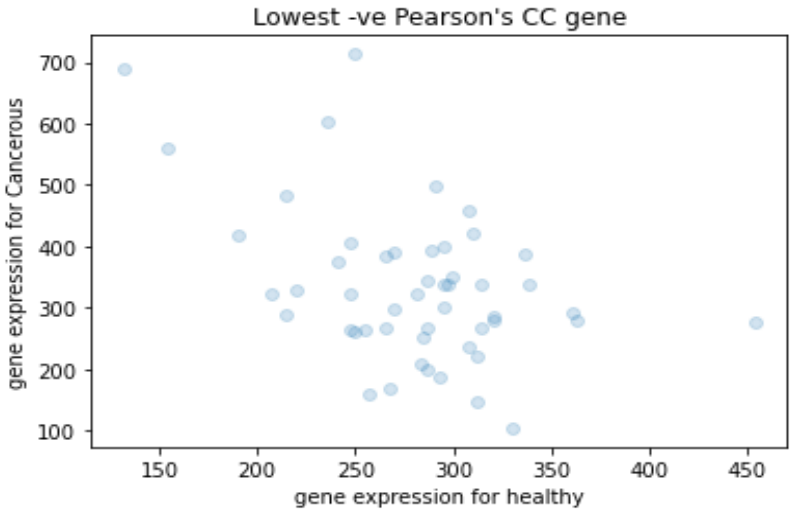
After computing correlation coefficients and sorting them we found out that the:

1- Pearson's Correlation Coefficient

- The gene with the highest positive coefficient is **AREGB** with correlation coefficient = 0.969044



- The gene with the Lowest Negative coefficient is **FAM222B** with correlation coefficient = -0.452807



As for the hypothesis we applied all the statistical test on two pairing cases and comparing DEGs in both cases we got the following results (refer to the attached spreadsheets):

- the number of common genes between DEGs sets (paired and independent) = 12241 genes.
- the number of distinct genes between DEGs sets (in DEGs independent but not in DEGs paired) = 79 genes .
- the number of distinct genes between DEGs sets (in DEGs independent but not in DEGs paired) = 169 genes.

Conclusion:

By reviewing the results, we found out that the gene AREGB which has the highest +ve CC showcase that both the healthy sample and the diseased sample move in the same direction on the other hand the gene FAM222B which has the lowest -ve showcase that the healthy sample and the diseased sample move in opposite directions.

The null hypothesis was rejected and the alternative hypothesis was accepted which means that the genes are affected (changed) when diseased.

Team contribution:

Nouran Khaled: Reading the files, correlation.

Alaa Tarek and Amira Omar: Hypothesis testing and comparing DEGs sets.

Salma Haitham: FDR correction.