INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

HOMEWORK ASSIGNMENT - 2

INDRAPRASTHA INSITUTE OF INFORMATION TECHNOLOGY, DELHI

COMPUTER SCIENCE AND APPLIED MATHEMATICS

Introduction to Quantitative Biology (BIO213)

Author:
Aditya Chetan
(2016217)

Date: January 18, 2018

1 Gene Expression and Z-Scores

Genes contain information that is decoded and used to synthesize proteins. This information that helps in gene synthesis is called **Gene Expression**. It decides when and what amount of protein is to be synthesized.

A **Z-Score** on the other hand, gives the number of standard deviations away from the mean score that a data point is. It gives us a measure of how far away a score is from the mean of the entire dataset. They come handy in comparing the results from a test to the normal score.

Statistical correlation between Gene Expression Levels and Copy Number Variations

On calculating the correlation between Gene Expression Levels and Copy Number Variations, we found that the correlation was 0.74 which is greater than 0.5. This shows that there is a positive correlation between the two values. That is, if one increases, the other increases too.

2 Code

```
# The following code was submitted as a part
# of Homework Assignment - 2 of BIO213 offered
# at IIITD.
#
# Name              :       Aditya Chetan
# Roll No.          :       2016217
# Python Version    :       Python 3.6+



# Imports necessary for the code

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np




# Reading the data file

df = pd.read_csv("./data/Cnv_detailsThu Jan 18 09_09_18 2018.csv")

# Extract the "Z-Score" records from the dataframe

z_score_data = df[ [ 'Expr Level (Z-Score)' ] ]


# Calculate mean and standard deviation of the data

mean = z_score_data.mean()[0]

stddev = z_score_data.std()[0]

# Printing the mean value and standard deviation to
# console

print("\n\nFor Z-Score:\n1. Mean:\t\t",mean,\
        "\n2. Standard deviation:\t",stddev,"\n")

# Plot the records as a normalised histogram

z_score_data.hist(normed=1)

# Extract the "Copy Number" records from the dataframe

copy_number_data = df[ [ 'Copy Number' ] ]
copy_number_data = copy_number_data.dropna()

# Extract NumPy arrays of the two columns

z_score_array = z_score_data.stack().as_matrix()[:20]
copy_number_array = copy_number_data.stack().as_matrix()


# Calculate the correlation between the two variables

corr = np.corrcoef(z_score_array, copy_number_array)[1][0]

# Print the correlation coefficient to console
print("\n\nThe correlation coefficient of Z-Score and Copy number is"\
        ,corr, "\n")

# Show the plotted figure

plt.show()
```

3 Input and Output

Below is the console output of the above code and the graph obtained from it.

Figure 1: Input and Output of the code

```
(assn2) adsrc@binsrc ➜ two git:(master) ✗ python3 analysis.py


For Z-Score:
1. Mean:                4.162285714285715
2. Standard deviation:   7.044501269071575



The correlation coefficient of Z-Score and Copy number is 0.7405503768648262
```

Figure 2: The probability distribution of the Z-Scores