Statistics and Bioinformatics Assignment Part 1

Bryan Mensah

21 November 2018

Note   
Anova Hypothesis for all the anova tests conducted in the report.  
Null Hypothesis- The mean is the same for all groups.  
Alternative Hypothesis- The mean isn’t the same for all groups.

## Question 1

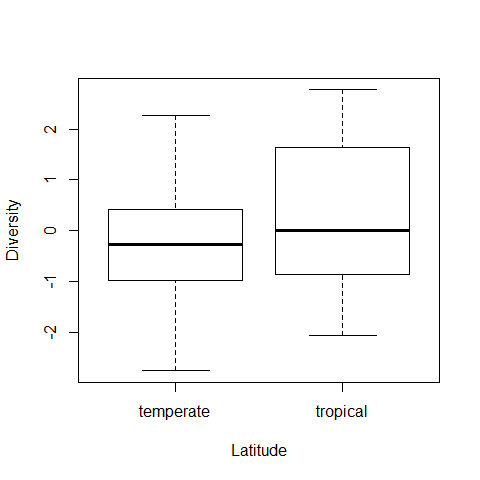


Figure 1: Boxplot of the marine microbial diversity and the latitude

Analysis of Variance Table 1: Analysing the effects of differing latitude on the Microbial marine diversity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Latitude | 1 | 3.508 | 3.508 | 1.862 | 0.1804 |
| Residuals | 38 | 71.58 | 1.884 | NA | NA |

Referring to figure 1, the boxplots have a similar median. However, the tropical boxplot is skewed to the right which suggest a potential bias in our data favouring larger values. Tropical also has a higher interquartile range. These differences suggest variation between our latitude data. There are no outliers.

**ANOVA analysis**

Referring to table 1, p > 0.05 we accept the null and reject the alternative. The mean is the same for all groups. So, there is no statistically significant difference between the two means.  
From the results of our analysis it suggests:  
Microbial Diversity had no large effect on the latitude (tropical and temperate) so we could say that any changes in the microbial diversity may not be dependent on the latitude and any changes to suggest otherwise would be due to chance. However, given our sample size we should consider further analysis before drawing any conclusion.

## Question 1 part 2



Figure 2:Boxplot of the microbial diversity and the Season

Analysis of Variance Table 2:Analysing the effects of differing time of years (season) on the Microbial marine diversity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Season | 1 | 0.4898 | 0.4898 | 0.2495 | 0.6203 |
| Residuals | 38 | 74.59 | 1.963 | NA | NA |

Referring to figure 2, boxplots have a similar median. January has a higher upper quartile range and interquartile range. These differences suggest variation between our data.

**ANOVA analysis**

Referring to table 2, P > 0.05 we accept the null and reject the alternative.  
The mean is the same for all groups. So there is no statistically significant difference between the two means we can conclude that those differences may be due to chance. From the results of our analysis we found: Microbial Diversity had no large effect on the season. so, we could say that any changes in the microbial diversity may not be dependent on the season and any changes to suggest otherwise would be due to chance. However, given our sample size we should consider further analysis before drawing any conclusion.

## Question 1 Part 3

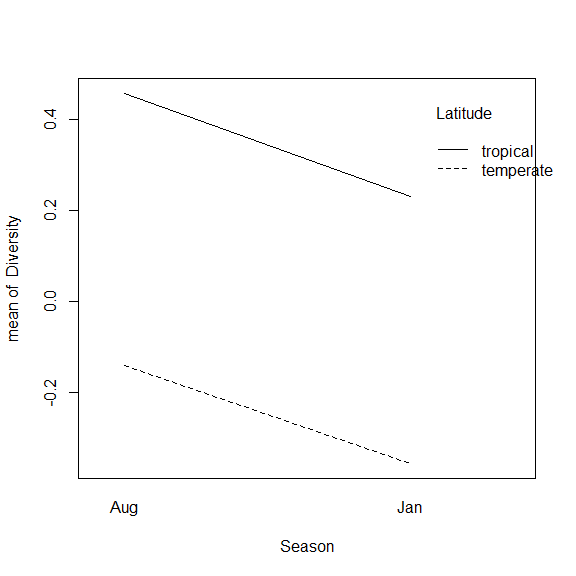


Figure 3:Interaction plot showing the interactions of the 3 variables (Season, Latitude and Diversity)

Analysis of Variance Table 3:Analysing the effects of differing latitude on the season

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Season | 1 | 6.281e-33 | 6.281e-33 | 2.387e-32 | 1 |
| Residuals | 38 | 10 | 0.2632 | NA | NA |

The interaction plot show that the season and location don’t interact. However, as season increases the latitude decreases.

**ANOVA analysis**

Referring to table 3,P > 0.05, so we accept the null and reject the alternative.  
The mean is the same for all groups.

From the results of our analysis we found:

Season had no large effect on the latitude- Any changes to suggest otherwise would be due to chance. However, given our sample size we should consider further analysis before drawing any conclusion. For our model’s appropriate tests were conducted and passed to ensure assumptions were valid.

## Question 2

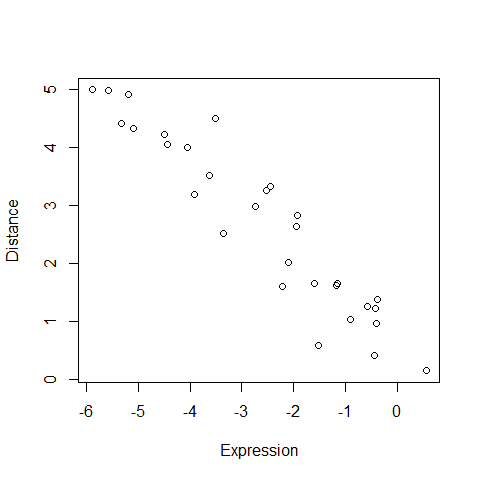


Figure 6:Scatter plot of the relationship between gene expression and genetic distance

Pearson’s product-moment correlation: Expression and Distance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test statistic | df | P value | Alternative hypothesis | cor |
| -14.96 | 28 | 6.924e-15 \* \* \* | two.sided | -0.9428 |

Analysis of Variance Table 4:Analysing the effects of differing gene expression on the genetic distance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Distance | 1 | 86.22 | 86.22 | 223.9 | 6.924e-15 |
| Residuals | 28 | 10.78 | 0.385 | NA | NA |

Referring to table 4, P < 0.05 we reject the null and accept the alternative. The mean isn’t the same for all groups. This means that the expression does have a significant on the genetic distance.

## Part 2

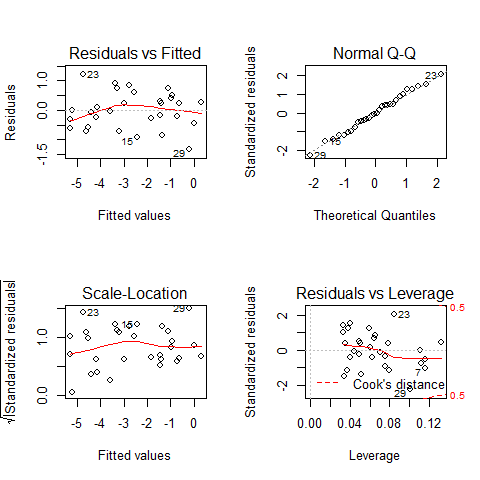


Figure 7:Diagnostic plots of anova model of Expression and Distance

The model assumption all valid, referring to figure 7 the graphs look fine with no random or unusual outlier values.

## Part 3

The effect that responsible for the relationship is natural selection (evolution).Changes in the DNA sequences cause changes in the expression level of the gene (Wu, no date). Depending on the type of natural selection can cause an increase or decrease in the frequency of gene (Herrera, Garcia-Bertrand and Salzano, no date). This then cause the expression level to be altered depending on the intensity of the natural selection (Akutsu, 2010). So as the gene expression decrease we are see less variation in our gene meaning those genes are closely related.

# Question 3

## Model Selection

When selecting the model, I decided to use the AIC value as my selection criteria. I used a stepwise regression. The forwards, backward and both searches were performed. The (both) model was selected as it had the lowest AIC value and also the model was simple enough while also providing worthwhile information.  
Model selected: Viral ~ CD4 + Shannon Diversity + Genetic Distance   
Using the model selected in method selection. Further analysis was carried out.

Analysis of Variance Table 5: Anova table of 3 models (From top to bottom.Backward,Forward and Both)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 26 | 92.57 | NA | NA | NA | NA |
| 34 | 133.1 | -8 | -40.51 | 1.422 | 0.2341 |
| 36 | 140 | -2 | -6.963 | 0.9779 | 0.3895 |

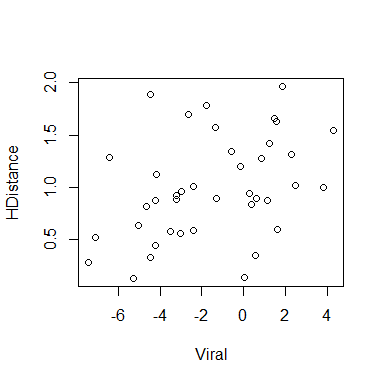


Figure 8:Scatter plot of the relationship between viral population size and Shannon diversity

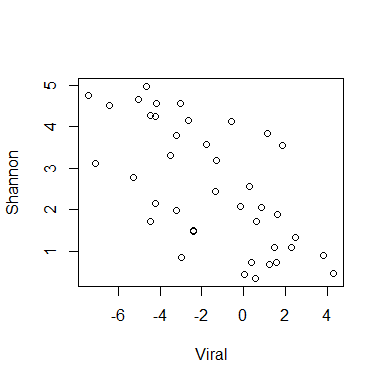


Figure 9:Scatter plot of the relationship between viral population size and Genetic distance

Analysis of Variance Table 6 : Analysing the effect of differing CD4,Shannon diversity and genetic distance on viral population size

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| CD4 | 1 | 3.399 | 3.399 | 0.8739 | 0.3561 |
| Shannon | 1 | 150.9 | 150.9 | 38.8 | 3.44e-07 |
| HDistance | 1 | 58.41 | 58.41 | 15.02 | 0.0004333 |
| Residuals | 36 | 140 | 3.89 | NA | NA |

Referring to figure 8, the graph shows a weak negative correlation. This suggests a possible relationship. As the viral population size decreases, the Shannon diversity decrease. Referring to figure 9, the graph shows a weak positive correlation. This suggests a possible relationship. As the viral population size increases, the genetic distance increase.

**ANOVA analysis**

Referring to table 6, P > 0.05 for CD4 we accept the null and reject the alternative. The mean is the same for all groups. Referring to the table 6, P < 0.05 for Shannon population diversity and genetic distance for we reject the null and accept the alternative. The mean isn’t the same for all groups. This means that both the Shannon population diversity and genetic distance do have a very significant effect on the viral load. So as the viral population size changes the Shannon population diversity and genetic distance changes enough to affect the viral population size.

From the results of our analysis we found: A positive relationship between Viral load s and genetic distance. This could be due to the high mutation rate of RNA viruses. Which lead to high genetic variability within the population (Jridi et al., 2006).This variability causes an increase in the genetic distance.

No relationship between Viral load and CD4. This could suggest they don’t directly impact each other.

A negative relationship between Viral load and Shannon diversity. Shannon diversity is measure of diversity(Van Dyke, 2008). So as the population decrease there is less variation and so less diversity.

**References**

Akutsu, T. (2010) Genome informatics 2010 : the 10th Annual International Workshop on Bioinformatics and Systems Biology (IBSB 2010) : Kyoto University, Japan, 26-28 July 2010. Imperial College Press. Available at: <https://books.google.co.uk/books?id=5roz_KaMC1oC&pg=PA154&dq=gene+expression+and+genetic+distance+selection+pressure&hl=en&sa=X&ved=0ahUKEwipl9yosYHfAhXlLcAKHVYmCMIQ6AEIXTAJ#v=onepage&q=gene> expression and genetic distance selection pressure&f=false (Accessed: 2 December 2018).

Van Dyke, F. (2008) Conservation biology : foundations, concepts, applications. Springer. Available at: <https://books.google.co.uk/books?id=Evh1UD3ZYWcC&pg=PA90&dq=SHANNON+DIVERSITY&hl=en&sa=X&ved=0ahUKEwiw0LSCx4HfAhVqSBUIHbz3BpUQ6AEIOzAE#v=onepage&q=SHANNON> DIVERSITY&f=false (Accessed: 2 December 2018).

Herrera, R. J., Garcia-Bertrand, R. and Salzano, F. M. (no date) Genomes, evolution, and culture : past, present, and future of humankind.

Jridi, C. et al. (2006) ‘Distinct viral populations differentiate and evolve independently in a single perennial host plant.’, Journal of virology. American Society for Microbiology (ASM), 80(5), pp. 2349–57. doi: 10.1128/JVI.80.5.2349-2357.2006.

UniProt (2018) P08659. Available at: <http://www.pnas.org/cgi/doi/10.1073/pnas.84.10.3264> (Accessed: 21 November 2018).

Wu, J. Y. (no date) Posttranscriptional gene regulation : RNA processing in eukaryotes.