A Cross-Sectional Study of Bacterial Infection Prevalence on Virtual Islands

Megan Wroolie
Chenghan Yue
Olivia Zang

Abstract

In this study, our objective was to determine whether there's a difference in the prevalence of people who have bacterial infections among the three islands. Our investigators thought the different locations could play an important role in the distribution of the disease because the climate could be different. We performed a Chi-Square test for homogeneity and proved that the location did influence the prevalence of people with bacterial infections. And through a logistic model, we found that the 'age' covariate also had a significant influence on the result while the other covariate 'gender' did not.

Introduction and Background

The research objective that the investigators are interested in is to determine whether there are any associations that geographic location may contribute to one specific diagnosis. But in our case, the specific diagnosis that we used was bacterial infections. In this project, our goal was to determine whether there is a difference in the number of people with the bacterial infection diagnosis on the three different islands. This topic is of interest because the investigators are very interested in health disparity issues. Due to the varying climates within Earth and with the issue of climate change, they want to see how the different climates/locations affect health and how health disparities can be affected by geography. Their idea is based on a previous study which has shown that the distribution of certain diseases is more prevalent in certain areas of the world. To get a glimpse into this topic, the focus of our study was on bacterial infections as the specific diagnosis of interest and how that distribution of the diagnosis varied geographically between the three islands.

Methods and Materials: Clinical

The population of this study was the islanders who went to the hospitals and received a bacterial infection diagnosis, and our primary response was whether the islander got a bacterial infection or not. The data collected from the hospitals assumes that islanders with bacterial infections go to the hospitals. Our investigators collected the data by going to each island's hospital every day and putting all patients and their respective diagnosis into a list. Then after a few days of collecting patient's information from the hospitals, patients were randomly sampled from the hospital list for each island and it was recorded if they

had a bacterial infection or not. Since our focus was just on the relationship between the islands and the prevalence of bacterial infections, there were no restrictions on age, gender, occupation and income when collecting the data. However, some of that information was still recorded and they were treated as covariates when we created logistic regression models.

Methods and Materials: Statistical

For this study, our null hypothesis is that there's no difference in the prevalence of people who have bacterial infections among the three islands. In contrast, our alternative hypothesis is that at least one of these islands is different in the prevalence of bacterial infections. To analyze the hypothesis, we performed the Chi-Square test for homogeneity. This test was designed to determine if two or more populations (or subgroups of a population) have the same distribution of a single categorical variable ("Test of Homogeneity," n.d.). There are three assumptions to this test which are independence, randomness, and the expected value of each of the data points is at least equal to 5 ("Introduction to the chi-square test for homogeneity," n.d.). We had a back-up plan that if the third assumption was violated, then we would proceed to switch to another specific diagnosis instead of bacterial infection. Thus, it is necessary to record other diagnoses from patients when collecting data just in case the data of bacterial infections couldn't meet the third assumption of the Chi-Square test.

Before collecting data, it was important to know how much data should be collected. A built-in function in R called "power.prop.test" was used to calculate sample sizes. This function is created for power calculations for the two-sample test for proportions. Since there are three samples here, we were plugged in all combinations of the two proportions of bacterial infections on each island found in the pilot study and one sample size was computed. Below, are the proportions that were plugged into this function and the resulting sample size listed below: The power was set equal to 0.80, the Type I error to 0.05, and the alternative test to "two-sided."

Table 1: Sample Size Calculated Using Pilot Study Proportions of Bacterial Infections on each Island

	n= sample size
P(North) = 2/13 & P(Center) = 3/16	1961
P(North) = 2/13 & P(South) = 3/11	186
P(Center) = 3/16 & P(South) = 3/11	382

The goal was to pick the largest sample size as the suggested sample size and thus each case could be satisfied. Unfortunately, this size was too large for our investigators. Thus, we suggested to collect 100 patients per island (total of 300 patients) since 100 samples are usually enough to perform well for a Chi-Square test. If our investigators couldn't collect enough data for bacterial infection, we suggested to look at another diagnosis.

Results

1. Data Exploration

Before analysis of the data, we created a demographic table. The investigators collected the demographic statistics of age and sex for each patient in the hospital list along with the sample size that was collected from each island. Table 2 below summarizes this information.

Table 2: Baseline Demographics of the Three Islands

	North Island	**Central Island**	**South Island**
Sample size	44	74	101
Sex (Males : Females)	17 : 18	39 : 28	44: 43
Age (mean ±SE)	65 ± 3.4	61.8 ± 2.7	56.7± 2.7

Table 2 shows that the sample size of these three islands were quite different. As a result, the number of patients who suffered a bacterial infection among the three islands couldn't be compared but instead the proportions should be compared. Also, Table 2 indicates that overall the ratio of male to female is close to 1 to 1 in the North and South Island but the Central Island had more males. Note that the number of males and females does not add up to the sample size in each island. This is because the data did not have sex assigned to some of the observations. Also, notice that there was not much of an age difference among the islands because the mean of them were concentrated from 56.7 to 65 and the standard errors are small. The response of the prevalence of bacterial infections on the three islands can be illustrated in the pie charts below.

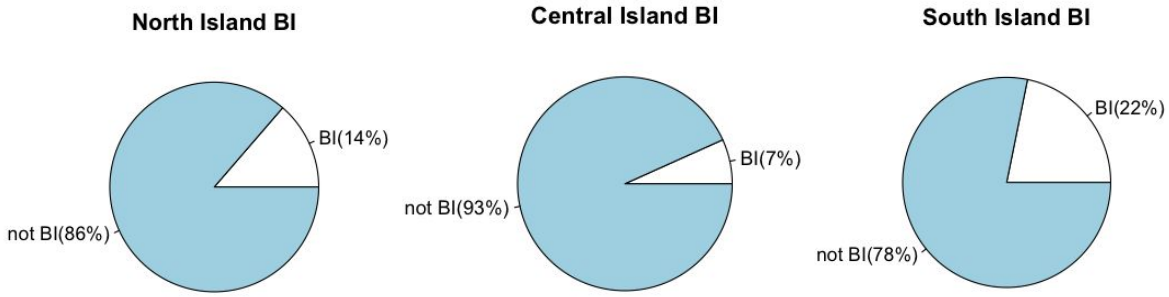

Figure 1: Pie Charts for Bacterial Infections on the Three Islands

Figure 1 indicates that bacterial infections are more prevalent on the South Island compared to the other two islands. The percentage of bacterial infections on the South Island is 8% higher than the North Island and 15% higher than the Central Island. The Central Island has the lowest percentage of bacterial infections with a percentage of 7%.

2. Analysis Method

To determine whether there was a difference between the number of people with the bacterial infection diagnosis on the three different islands, we performed the Chi-Square test for homogeneity. The first assumption for this test was met because we checked for duplicate records of the patients, and any duplicates that we encountered were deleted. As a result, each patient is different and independent from one other. On the other hand, the second assumption was not met. It was intended to go to the hospitals every day to collect all the islanders there, and then randomly collected 100 islanders per island from the whole data. But, we were only able to receive a certain number of patients for each island's hospital which was under the number of patients we wanted to randomly select. As a result, we were not able to randomly select our final patients and we just used what we had. Although a backup plan was made if the third assumption was not satisfied, this would not help the case for the failure for the second assumption. However, the third assumption was met. We calculated the expectation for each of the data points and they were all greater than 5 (outlined in the appendix 1).

If there is a difference between the number of people with the bacterial infection diagnosis on the three different islands, then we can further study those differences with a special version of another test named Tukey's Test to see which two islands are significantly different from one another.

The results from our testing can be generalized to the population of islanders that go into the hospital and have/receive a diagnosis. Since we only sampled from the island's hospitals, we can only apply our inferences to the island's hospitals and not on the whole population of the island.

3. Data Analysis

3.1. Chi-Square Test for Homogeneity and Tukey's Test

Table 3: Bacterial Infections for Patients in the Hospital on the Three Islands

		Bacterial Infection?	
		Yes	No
Islands	North	6	38
	Central	5	69
	South	22	79

Based on Table 3, we calculated the Chi-Square test statistic to be approximately 7.62 with 2 degrees of freedom (outlined in the appendix 2). The calculated p-value was approximately 0.02 which is less than the set significance level of 0.05. That means, if the null hypothesis is true, the probability that we'd observe this test statistic (or even a more extreme test statistic) would be 0.02 which is very low. Therefore, we reject the null hypothesis and conclude that at least one of these islands is different in the prevalence of bacterial infections. Since the null hypothesis was rejected, to see what islands are different in terms of bacterial infection prevalence we used the special version of Tukey's test. Tukey's Test found that there's a statistically significant difference in bacterial infections between the South and Central Island. To get a more specific relationship between the three islands and digitize the result, we created models to do the deeper analysis.

3.2. Logistic Regression Model with one Predictor "Island"

A logistic regression model with no covariates and just the predictor "island" can be analyzed and interpreted.

$$\text{Log odds}(\text{Bacterial.Infection}) = -2.63 + 0.78 \cdot \text{NorthIsland} + 1.35 \cdot \text{SouthIsland}$$

From the summary of this model (outlined in the appendix 6) the coefficients can be interpreted as compared with the Central Island, the South Island increased the log odds of bacterial infection by 1.35. Additionally, compared with the Central Island, the North Island increased the log odds of bacterial

infection by 0.78. Table 8 in appendix 5 interprets this model in the odds scale as well. This model confirms the result from Figure 1 that the South Island has the highest bacterial infection prevalence followed by the North Island and the Central Island.

3.3. Logistic Regression Model with Covariates

Since age and sex may also influence the response, we compared four potential models with and without them. We compared each model using the criterion of AIC with the lower AIC the better the model is at predicting the response.

Table 4: Logistic Regression Models Built in R

Model 1: Log odds(Bacterial.Infection)=$\beta_0 + \beta_1 *$ Island
Model 2: Log odds(Bacterial.Infection)=$\beta_0 + \beta_1 *$ Island+$\beta_2 *$ Age
Model 3: Log odds(Bacterial.Infection)=$\beta_0 + \beta_1 *$ Island+$\beta_2 *$ Gender
Model 4: Log odds(Bacterial.Infection)=$\beta_0 + \beta_1 *$ Island+$\beta_2 *$ Gender +$\beta_3 *$ Age:Gender

The results showed that the model with the lowest AIC (outlined in the appendix 3) is the model with just the covariate of age. This means that age affects whether people get a bacterial infection or not.

$$\text{Log odds(Bacterial.Infection)} = 2.50 + 2.92 \cdot \text{NorthIsland} + 1.96 \cdot \text{SouthIsland} - 0.13 \cdot \text{Age}$$

Using this model adjusted for age (outlined in the appendix 6), coefficients can be interpreted as holding age at a fixed value the log odds of getting a bacterial infection on the North Island increases by 2.92 compared to the Central Island. Additionally, holding age at a fixed value the log odds of getting a bacterial infection on the South Island increases by 1.96 compared to the Central Island. The coefficient for age says that holding island at a fixed value, there will be a 0.13 decrease in the log odds of getting a bacterial infection for a one-year increase in age. Table 9 in appendix 5 interprets this model in the odds scale as well. Again using the special version of Tukey's test, it was found that now the Central Island and the North Island are significantly different from each other in terms of bacterial infection prevalence which is consistent with the result obtained from our model.

Note that from the above results, it is evident that the North Island and the South Island switches on bacterial infection prevalence when the covariate of age is introduced to the model. Now the prevalence of bacterial infections follows this order: Central<South<North. This switch could happen since

age did play an important role in the response as it was significant to the model thus it could change our previous result.

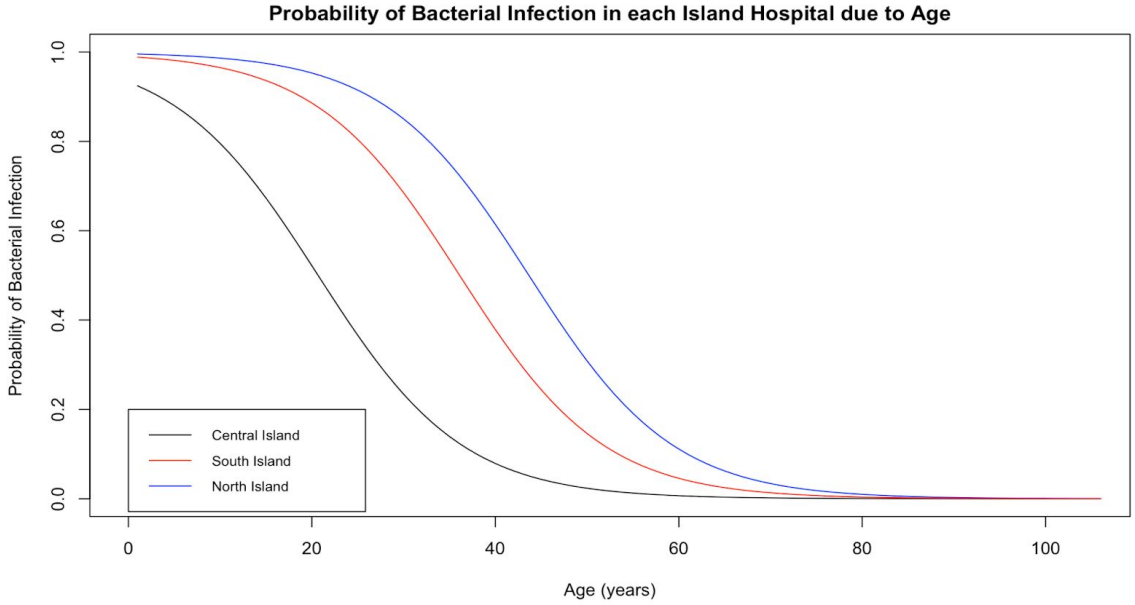

Figure 2: Probability of Bacterial Infection in each Island Hospital due to Age

From Figure 2, we can see that as age increases the probability of an islander going into the hospital and having a bacterial infection decreases. This result was actually surprising because we assumed that older islanders would be more susceptible to bacterial infections due to weaker immune systems. But looking back at the data, we did confirm that when older islanders came to the hospital they had other diagnoses such as heart disease and cancer. Rarely did older islanders have a bacterial infection and almost all younger islanders that came into the hospital had bacterial infections.

All these results can be generalized to the population of the islanders who went to hospitals, rather than the whole populations on the islands. Otherwise, Figure 2 wouldn't make sense since the probability of getting a bacterial infection on the South and North Island was almost 1 for young islanders. This indicates that when young islanders come into the hospital the probability of them having a bacterial infection compared to all other diagnoses is close to 1. However, this does not make sense on the population level because this would mean that every young islander in the entire island population would have a probability of getting a bacterial infection close to 1 which is very impractical.

Discussion and Summary

The results from our Chi-Square test for homogeneity found that there is indeed a difference in

bacterial infection prevalence on the three different islands. Therefore, we can conclude that there is an association between geographic location (the three islands) and one specific diagnosis (bacterial infection) which is consistent with the previous study. Through the analysis of the data with the best model which included the covariate of age, we found the relationship of bacterial infection prevalence among the islands to be Central<South<North. These results are consistent with the previous research that stated that the distribution of certain diseases is more prevalent in certain areas of the world because from our results we concluded that the distribution of bacterial infections is more prevalent on the North Island. From these results, we can imply that the North Island has more bacterial infection cases than the Central Island since we concluded from Tukey's test that the two are significantly different from each other in terms of bacterial infection prevalence.

One limitation to this study is that all our inference on prevalence of bacterial infections is on the hospital level and not on the whole population. This is because we only sampled from the island's hospitals and not the whole population. It would be ideal to conduct the study so the inference would be on the whole island population but this would be more difficult and would require a large sample size.

Here's a few suggestions to improve the study. If the time is adequate, we would suggest collecting more patients on each island's hospital to reach the threshold from which we would be able to randomly sample from. If the data pool is large enough, then we can randomly select 100 patients who went to each island's hospital. Thus, we would get more information about the population and make our estimates more accurate. Additionally, the results we got from this study only focused on the bacterial infection diagnosis. If our investigators want to do some further study about other diagnoses, they could follow our process and apply them to the other diagnoses. We also made tables and pie charts for "Bacterial infection," "Cancer," "Food poisoning", "Heart disease," and "Viral infection" so that they could take a look at the proportion of them among the different islands. Those were placed in the appendix 4.

Looking at our results, we don't necessarily know why when using the covariate model the North Island is significantly different from the Central Island in terms of prevalence of bacterial infections. We could speculate that it could be due to climate differences as the North Island is colder than the Central Island, occupational differences because the North Island has more pig farmers, or some other factor that we are unaware of. Further research would be needed to determine the cause of this difference in cases of bacterial infections between the North Island and the Central Island.

Appendix

1. Method for calculating the expected value of each of the data points

Table 5: Bacterial Infection for Patient in the Hospital on the Three Islands and Expectations

	Bacterial Infections			Row totals:	Expectation	
Islands:	Yes	No			Yes	No
North	6	38		44	6.630136986	37.36986301
Central	5	69		74	11.15068493	62.84931507
South	22	79		101	15.21917808	85.78082192
Column Totals:	33	186		219		

For North Island: E(yes)=(44/219)·33, E(no)= (44/219)·186

For Central Island: E(yes)=(74/219)·33, E(no)= (74/219)·186

For South Island: E(yes)=(101/219)·33, E(no)= (101/219) 186

All these results are greater than 5 (“Introduction to the chi-square test for homogeneity,”n.d.).

2. Method for calculating the Chi-Square statistic

$$\chi^2 = \frac{(6-6.63)^2}{6.63} + \frac{(38-37.37)^2}{37.37} + \frac{(5-11.15)^2}{11.15} + \frac{(69-62.85)^2}{62.85} + \frac{(22-15.22)^2}{15.22} + \frac{(79-85.78)^2}{85.78} = 7.62$$

Df = (#of column-1)·(# of row-1)=(2-1)·(3-1) = 2

3. AIC for each potential model

Table 6: AIC for each Potential Model with the Predictor Island and the Covariates of Age and Sex

	AIC
Model 1 (No covariate)	183.53
Model 2 (Age covariate)	86.25
Model 3 (Sex covariate)	186.61
Model 4 (Age and sex covariate)	88.25

4. Further study about diagnosis

Table 7: Diagnoses for Patients in the Hospitals on the Three Islands

		Diagnoses				
		Bacterial Infection	Heart Disease	Food Poisoning	Cancer	Viral Infection
Islands	North	6	14	4	20	0
	Central	5	42	1	26	0
	South	22	56	1	20	2

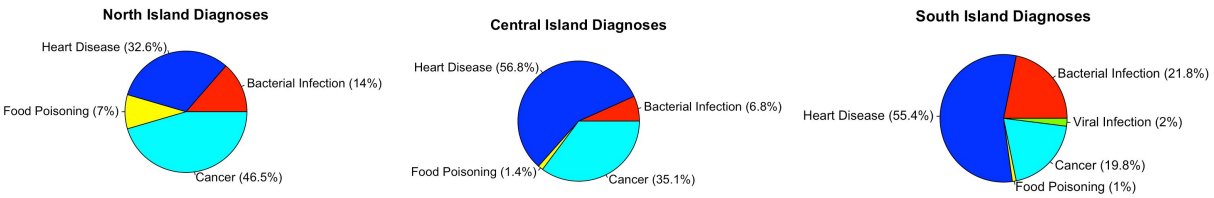

Figure 3: Diagnoses on the Three Islands.

5. Supporting Figures

Table 8: Result Representation (only treat island as predictor in the model)

	Log Odds Scale	Odds Scale
South vs. Central	The South Island increased the log odds of bacterial infections by 1.35	The South Island increased the odds of bacterial infection by e^1.35 = 3.89 (285.7%)
North vs. Central	The North Island Increased the log odds of bacterial infections by 0.78	The North Island increased the odds of bacterial infection by e^0.78 = 2.18 (118.1%)

Table 9: Result Representation (only both island and age as predictors in the model)

	Log Odds Scale	Odds Scale
North vs. Central	Holding age fixed, the North Island increased the log odds of bacterial infection by 2.92	Holding age fixed, the North Island increased the odds of bacterial infection by e^2.92= 18.52 (1754.1%)
South vs. Central	Holding age fixed, the South Island increased the log odds of bacterial infection by 1.96	Holding age fixed, the South Island increased the odds of bacterial infection by e^1.96= 7.10 (609.9%)
Age	Holding island fixed, there will be a 0.13 decrease in the log odds of getting a bacterial infection for a one-year increase in age	Holding island fixed, there will be a e^0.13 = 1.14 (13.9%) decrease in the log odds of getting a bacterial infection for a one-year increase in age

6. R code and analysis

```r
#Chi-Squared test
pchisq(7.622315023, df = 2, lower.tail = FALSE)
```

```
[1] 0.02212256
```

```r
#Reading in data
Data = read.csv("Consulting Data.csv")

#Changing bacterial infection variable to a factor
Data$Bacterial.Infection = as.factor(Data$Bacterial.Infection)
```

```r
#Version of Tukey Test's

#Creating model with just island as a predictor
model.1=glm(Bacterial.Infection~Island,family=binomial, data=Data)


#Performing Tukey's test for generalized linear models
library(multcomp)
summary(glht(model.1, mcp(Island = "Tukey")))
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glm(formula = Bacterial.Infection ~ Island, family = binomial,
    data = Data)

Linear Hypotheses:
                     Estimate Std. Error z value Pr(>|z|)
North - Central == 0   0.7788     0.6383   1.220   0.4363
South - Central == 0   1.3463     0.5221   2.578   0.0263 *
South - North == 0     0.5674     0.5011   1.132   0.4892
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

```r
#Summary of model with no covariates
model.1=glm(Bacterial.Infection~Island,family=binomial, data=Data)
summary(model.1)
```

```
Call:
glm(formula = Bacterial.Infection ~ Island, family = binomial,
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7010  -0.7010  -0.5415  -0.3741   2.3215

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6247     0.4631  -5.667 1.45e-08 ***
IslandNorth   0.7788     0.6383   1.220  0.22242
IslandSouth   1.3463     0.5221   2.578  0.00992 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 185.67  on 218  degrees of freedom
Residual deviance: 177.53  on 216  degrees of freedom
AIC: 183.53

Number of Fisher Scoring iterations: 5
```

```r
model.1=glm(Bacterial.Infection~Island,family=binomial, data=Data)
model.2=glm(Bacterial.Infection~Island + Age,family=binomial, data=Data)
model.3=glm(Bacterial.Infection~Island + Sex,family=binomial, data=Data)
model.4=glm(Bacterial.Infection~Island + Age + Sex,family=binomial, data=Data)

#Comparing models using AIC
data.frame( Model = c("Model 1 (No covariates)", "Model 2 (Age covariate)", "Model
3 (Sex covariate)", "Model 4 (Age and sex covariate)"), AIC = c(AIC(model.1),
AIC(model.2), AIC(model.3), AIC(model.4)))
```

Model <fctr>	AIC <dbl>
Model 1 (No covariates)	183.52730
Model 2 (Age covariate)	86.24766
Model 3 (Sex covariate)	186.61032
Model 4 (Age and sex covariate)	88.24747

4 rows

```{r}
#Summary of model with age covariate
model.2=glm(Bacterial.Infection~Island + Age,family=binomial, data=Data)
summary(model.2)
```

```
Call:
glm(formula = Bacterial.Infection ~ Island + Age, family = binomial,
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2721  -0.3170  -0.1118  -0.0325   3.0426

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.50256    0.90366   2.769  0.00562 **
IslandNorth  2.92280    1.05647   2.767  0.00566 **
IslandSouth  1.96251    0.86461   2.270  0.02322 *
Age         -0.12717    0.02376  -5.352 8.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.022  on 188  degrees of freedom
Residual deviance:  78.248  on 185  degrees of freedom
  (30 observations deleted due to missingness)
AIC: 86.248

Number of Fisher Scoring iterations: 7
```

```{r}
#Version of Tukey Test's

#Creating model with the covariate age
model.2=glm(Bacterial.Infection~Island+Age,family=binomial, data=Data)


#Performing Tukey's test for generalized linear models
library(multcomp)
summary(glht(model.2, mcp(Island = "Tukey")))
```

```
         Simultaneous Tests for General Linear Hypotheses

 Multiple Comparisons of Means: Tukey Contrasts


 Fit: glm(formula = Bacterial.Infection ~ Island + Age, family = binomial,
    data = Data)

 Linear Hypotheses:
                      Estimate Std. Error z value Pr(>|z|)
 North - Central == 0   2.9228     1.0565   2.767   0.0149 *
 South - Central == 0   1.9625     0.8646   2.270   0.0574 .
 South - North == 0    -0.9603     0.7170  -1.339   0.3640
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 (Adjusted p values reported -- single-step method)
```

References

1. Introduction to the chi-square test for homogeneity.(n.d.). Retrieved from https://www.khanacademy.org/math/ap-statistics/chi-square-tests/chi-square-tests-two-way-tables/v/chi-square-test-homogeneity

2. Test of Homogeneity. (n.d.). Retrieved from https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/test-of-homogeneity/