Assignment 5

Anant Vishwakarma

5/11/2022

Our task was to investigate and analyze two cases of corona virus in Washington and to check if both are related to each other or are cases of independent transmission. Our null hypothesis states that the two cases Wa-1 and Wa-2 were transmitted to Washington independently.

Firstly we can see how the two cases differ from each other. We can do this by using Levenshtein distance.

```
## [1] 1
```

We can see that the two Washington cases Wa-1 and Wa-2 differ from each other by one base pair. We can now look at the distance table of the cases to get the proportion of cases that have common genome sequences or mutations.

```
##            DistFromWa2
## DistFromWa1  1  2
##           0  2  0
##           1  0 68
```

There are two cases which differ from Wa-2 by 1 base pair and from Wa-1 by 0 base pairs (one of these would be Wa-1 itself). The remaining 68 cases must all be from outside Washington as they differ from Wa-2 by 2 base pairs and Wa-1 by 1 base pair.

In fact, we can also see that the other 68 sequences from outside of Washington are identical, in that they differ from one another by zero base pairs:

```
## OtherDist
##  0  1
## 68  2
```

Thus if we independently sample cases the estimated proportion p is $1/69 = 0.0145$.

We can take the significance level alpha as, $\alpha = 0.05$ If the null hypothesis is true and the two Washington cases were transmitted independently then the probability that they both have the same mutation is given by the binomial distribution.

```
## [1] 0.00021025
```

From this we get the p-value to be 0.000210. This means that if we independently sample two viruses then there is a 0.02% probability that they have the mutation. We can see that $0.0002 < 0.05$. So the p-value is much smaller than the significance level. In this case we can reject the null hypothesis.

Discussion of Results

Based on our sample and estimated p-value, we do not have enough evidence to say that the second case is independent from the first case i.e it is very unlikely that the two cases are independent. It is possible that the second one was infected/caused by the first. This might suggest that an undetected community transmission was occuring in Washington between January and March

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(echo = FALSE)
GenomeDataFrame <- read.table("http://www.stat.ucdavis.edu/~affarris/corona-genomes.txt",
                              skip = 2,
                              stringsAsFactors = FALSE)
Genomes <- GenomeDataFrame[[1]] #extract character vector

DistanceMatrix <- adist(Genomes)

# the distance between the two WA viruses
DistanceMatrix[1,2]
knitr::opts_chunk$set(echo = FALSE)
# the distance from the other viruses compared to the two WA
DistFromWa1 <- DistanceMatrix[3:72,1]
DistFromWa2 <- DistanceMatrix[3:72,2]
table(DistFromWa1,DistFromWa2)
knitr::opts_chunk$set(echo = FALSE)
# check the other viruses compared to each other
OtherDist <- DistanceMatrix[3:72,3]
table(OtherDist)
knitr::opts_chunk$set(echo = FALSE)
dbinom(x = 2, size = 2, prob = 0.0145)
```