Computing Assignment 1

STA 32, Farris | James Junaidi | Winter 2020

This is the work for computing assignment one for Statistics 32, Farris, Section A01. The work after highlighted portions done by James Junaidi. This contains some analysis of COVID-19 genomes, specifically from the Washington cases and delves into some data on how those particular genomes are differing from the genomes throughout the rest of the world.

A genomic interlude

Introduction

In this case study, we will investigate a pair of recent cases of Covid-19 from Washington state. Here we will offer an analysis following that of Trevor Bedford, a researcher at the Fred Hutchinson Cancer Research Center in Seattle[1].

On January 19, 2020, a male patient in Snohomish County, WA was given a oropharyngeal swab. He had travelled in recent days from Wuhan, China, and was showing symptoms of Covid-19. From this swab, the genetic sequence of the infecting virus was obtained, and this was quickly made public by the CDC.

On February 24, a suspected second patient in Snohomish County was identified for the disease, and given a nasal swab. This patient had not travelled to a known, affected area, which drew attention to the possibility of undetected transmission of the disease. As a consequence of this new case, either the virus has been circulating in Snohomish county undetected in the interim, or some new carrier has brought it into the county for the second time.

Genetic material from this swab was quickly sequenced, the results of which were distributed on March 1.

Between these genetic sequences and many others that have been obtained by labs around the world, it is possible to investigate genetic variation within the virus as it spreads.

The genome of the virus consists of approximately 30,000 base pairs. Genetic variation is driven by mutations, which typically affect only one base pair at a time; moreover, these mutations are passed on to consecutive generations, so that as the virus spreads around the world, distinct viral lineages slowly start to accumulate differences.

Computing differences between genomes

We'll begin our analysis by downloading the file `coronas.txt` from Canvas into the same folder that contains our `.Rmd` file, and reading its contents into memory:

```
GenomeDataFrame <- read.table("coronas.txt",
                              skip = 2,
                              stringsAsFactors = FALSE) # read in data
Genomes <- GenomeDataFrame[[1]] # extract character vector
```

This file contains some genomic data that has been obtained from various labs and collected with GISAID[2]. Because of limitations on the permitted dissemination of the data, only a small portion appears here; instead of the whole genome, we will look at a snippet taken from each one, each consisting of a few hundred base pairs. The first two rows of the data here contain part of the genome from each of the first and the second cases in Washington, respectively. The remaining rows contain part of the genome from each of seventy other genomes obtained during the outbreak from around the world.

[1] https://bedford.io/blog/ncov-cryptic-transmission/
[2] https://www.gisaid.org/

In order to compare these genomes, we can use the *Levenshtein distance*. This measures the 'distance' between two base pair sequences by the number of mutations required to obtain one from the other.

To do this, we can use the function `adist`:

```
DistanceMatrix <- adist(Genomes) # get Levenshtein dist.
```

We can use this to, for example, check the similarity between the first and second Washington genome sequences:

```
DistanceMatrix[1,2] # distance between 1st and 2nd entries
```

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

Thus we find that there is exactly one base pair difference in the part of the genome that we are looking at differentiating the first and second Washington genomes. We can find which base pair differs by splitting the sequences into individual letters and then comparing them, each pair at a time. Finally, we can look at the neighborhood of the difference in each of the two sequences:

```
SplitGenomes <- strsplit(Genomes,split = "") # split into individual base pairs
SplitWa1 <- SplitGenomes[[1]]
SplitWa2 <- SplitGenomes[[2]]
MutationLocation <- which(diag(adist(SplitWa1,SplitWa2))==1) # find the location of the mismatch
substr(Genomes[1:2],MutationLocation-4,MutationLocation+4) # print out location's neighborhood
```

```
## [1] "GAATATGAC" "GAATGTGAC"
```

From this we can see that a one base pair (the 47th) in the first Washington case has A, and the same location for the second case has G.

How does our available sequence from Wa-1 differ from the other genomes?

```
DistFromWa1 <- DistanceMatrix[,1]
table(DistFromWa1) # by how many bp's do other seq's differ from Wa-1
```

```
## DistFromWa1
##  0  1
##  3 69
```

We find that three sequences in total are the same as Wa-1, and 69 differ by one base pair.

To compare both Wa-1 and Wa-2 simulataneously with the others:

```
DistFromWa2 <- DistanceMatrix[,2]
table(DistFromWa1,DistFromWa2) # by how many bp's do other seq's differ from Wa-1 and Wa-2
```

```
##            DistFromWa2
## DistFromWa1  0  1  2
##           0  0  3  0
##           1  1  0 68
```

We see that, out of 72 sequences, only one of them differs from Wa-1 by one base pair and one from Wa-2 by none (this must be Wa-2 itself). Three of them differ from Wa-1 by none and from Wa-2 by one; so, there must be two other sequences identical to Wa-1, aside from itself. Of the remaining 68 sequences, all differ from Wa-1 by one base pair and from Wa-2 by two.

How many unique sequences are present from the other 70 genomes available? Which of these sequences is most common, how does it relate to the Washington sequences, and exactly how common is it in this sample? How do the less common, unique sequences relate to the Washington sequences?

To see how many unique sequences are present:

```
duplicated(DistanceMatrix) # returns False for unique elements, True for duplicate elements
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
## [37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [49]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [61]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
unique(Genomes) # prints out the Genomes that are unique in our set
```

```
## [1] "TTGGGACTACCAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATATTCACTCAAACCACTGAAACAGCTCACTCTTGTAATGTA
## [2] "TTGGGACTACCAACTCAAACTGTTGATTCATCACAGGGCTCAGAATGTGACTATGTCATATTCACTCAAACCACTGAAACAGCTCACTCTTGTAATGTA
## [3] "TTGGGACTACCAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATATTCACTCAAACCACTGAAACAGCTCACTCTTGTAATGTA
```

We see that there are only three unique genomes, because the first three are False, but the rest are True, showing us that the rest of them after the first three are duplicates of one of the three. But now we want to find out which ones are duplicates of which. Additionally, when we print out the unique genomes from the original list, we see that only three are given.

Now let's figure out which genomes exactly were unique or duplicated:

```
unique(DistanceMatrix) # allows us to identify the duplicates
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    0    1    1    1    1    1    1    1    1     1     1     1     1     1
## [2,]    1    0    2    2    2    2    2    2    2     2     2     2     2     2
## [3,]    1    2    0    0    0    0    0    0    0     0     0     0     0     0
##      [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
## [1,]     1     1     1     1     1     1     1     1     1     1     1     1
## [2,]     2     2     2     2     2     2     2     2     2     2     2     2
## [3,]     0     0     0     0     0     0     0     0     0     0     0     0
##      [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38]
## [1,]     1     1     1     1     1     1     1     1     1     1     1     1
## [2,]     2     2     2     2     2     2     2     2     2     2     2     2
## [3,]     0     0     0     0     0     0     0     0     0     0     0     0
##      [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
## [1,]     1     1     1     1     1     1     1     0     1     1     1     1
## [2,]     2     2     2     2     2     2     2     1     2     2     2     2
## [3,]     0     0     0     0     0     0     0     1     0     0     0     0
##      [,51] [,52] [,53] [,54] [,55] [,56] [,57] [,58] [,59] [,60] [,61] [,62]
## [1,]     1     1     1     1     1     1     1     1     0     1     1     1
## [2,]     2     2     2     2     2     2     2     2     1     2     2     2
## [3,]     0     0     0     0     0     0     0     0     1     0     0     0
##      [,63] [,64] [,65] [,66] [,67] [,68] [,69] [,70] [,71] [,72]
## [1,]     1     1     1     1     1     1     1     1     1     1
## [2,]     2     2     2     2     2     2     2     2     2     2
## [3,]     0     0     0     0     0     0     0     0     0     0
```

Now we see the unique elemets of the DistanceMatrix, and we can see that genome two, which was Wa-2, is unique and genomes 1, 46, and 59 are the same (proving our earlier conclusion that there were two other sequences that are identical to Wa-1) and the other 68 are actually the same. There are no sequences that are the same as Wa-2, so it is truly unique. However, an interesting observation here is that there are two other genomes that are the same as Wa-1. However, it is not that common in comparison with the 70 other genomes from around the world. However, this does lead us to know that around the world, the genome is consistent, while in Washington, only the more unique genomes appears.

Analyzing the differences between genomes

Let us now judge what we can about whether this Wa-1 was connected to Wa-2. On the one hand, the the genome sequences that we have for Wa-1 and Wa-2 both differ from the most common sequence among the 70 overseas cases. On the other, this is not the case when comparing the less common sequences to the Washington sequences.

Let's assume that the similarity was, in fact, due strictly to chance. This will be our *null hypothesis*. In this case, we can assume that the Washington cases were both random selections from the rest of the world. If this is the case, then what is the probability distribution of the number of differences from the common sequence above in a sequence

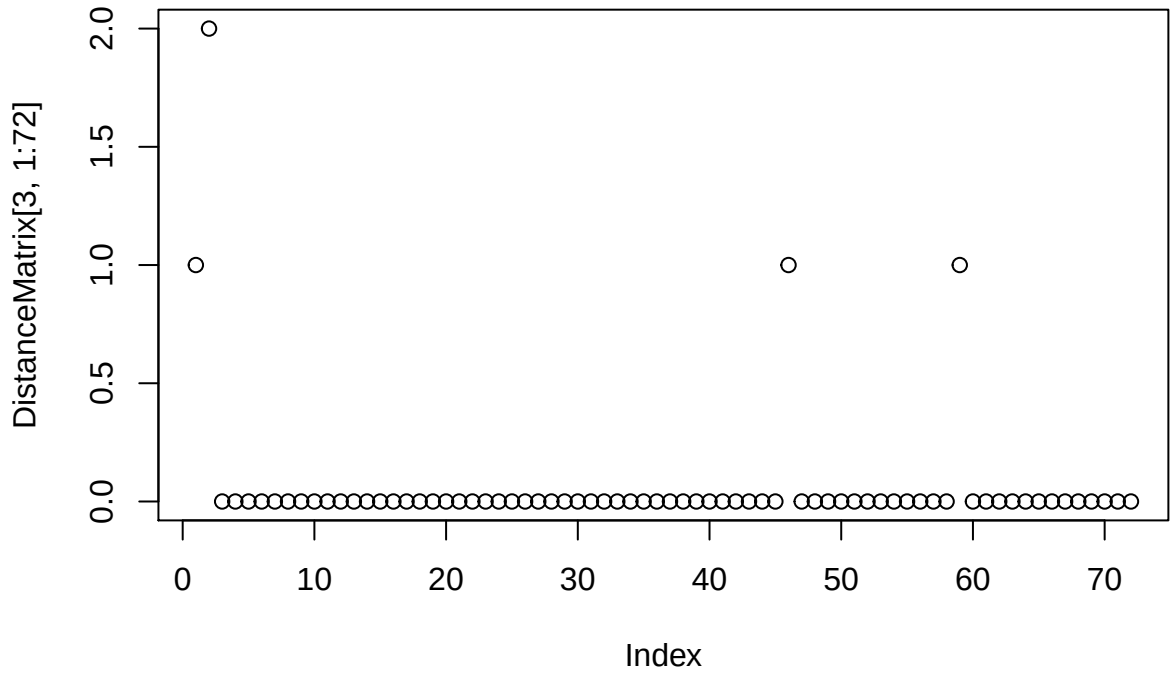
that is chosen at random the world? (this would give us the *null distribution*). The probability assigned by this distribution to the possibility of differing from the common sequence by one or more base pairs would be the *p-value*; if this value is small, it will reflect evidence against the assumption of the null hypothesis.

Because we don't know this null distribution exactly (because in turn we don't know exactly what the viruses around the world look like), we do not know the p-value exactly. However, we can estimate it. If we assume viruses entering Washington to be chosen independently, we would estimate the proportion of the virus that is one or more base pairs away from the common sequence outside of Washington to be $\hat{p}$, the sample proportion among the other sequences.

What is the value of the estimate here, and what do you estimate to be its standard error? Taking this proportion to be the p-value, would you judge it to be small or not? What does this tell you about whether or not the Wa-1 and Wa-2 cases arrived in Washington independently?

The probability distribution of the number of differences from the common sequence (the third sequence) can be seen here:

```
plot(DistanceMatrix[3,1:72]) #number of differences from the common sequence everywhere else
```



We can actually see that of all the sequences only five do not change from the common sequence (partly the reason why it is the common sequence) and that of these five sequences, three of them differ by only one base pair, while just one of them differs by two base pairs.

Here we can get the average number of base pairs that differs from the common sequence, the third sequence, but summing the total differences and dividing it by the number of base pairs, 72:

```
sum(DistanceMatrix[3,1:72]) / 72
```

```
## [1] 0.06944444
```

Here we can estimate a $\hat{p}$ which is the sample proportion of viruses that are one or more base pairs away from the common sequence outside of Washington. We know that four of them differ by one or more base pairs and that there are 72 total base pairs. So our proportion is 4/72 or:

```
phat <- 4/72 #storing our sample proportion
print(phat) #printing the phat in decimal form
```

```
## [1] 0.05555556
```

This $\hat{p}$ value is relatively small, saying that about 5.56% of coronovirus genomes differ from the most common genome. It would be considered uncommon to have a differing genome from the most common one.

Now we can calculate the standard error:

```
standardErr <- sqrt((phat*(1-phat)) / 72)
print(standardErr)
```

```
## [1] 0.02699515
```

With this standard error, we can also compute a 95% confidence interval:

```
left <- (phat - 2*(standardErr)) #gets the left endpoint of our CI
right <- (phat + 2*(standardErr)) #gets the right endpoint of our CI

#printing our confidence interval nicely (C-style printf statement)
sprintf("Confidence Interval: (%f, %f)", left, right)
```

```
## [1] "Confidence Interval: (0.001565, 0.109546)"
```

What this confidence interval tells us is that we are 95% confident that the true proportion of coronavirus cases that differ from the most common base pair configuration is between .1565% and 10.95%.

Now looking back to the case of Wa-1 and Wa-2 and their arrivals in Washington, they most likely did not arrive independently.

If we take a look at the differences in the strands again:

```
DistanceMatrix[1,2] # number of base pair differences between Wa-1 and Wa-2
```

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

We can see that there is only one base pair difference between Wa-1 and Wa-2

Now when we compare both of these with the third genome sequence (the common one)

```
DistanceMatrix[1,3] # number of base pair differences between Wa-1 and the common sequence
```

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

```
DistanceMatrix[2,3] # number of base pair differences between Wa-2 and the common sequence
```

```
## [1] 2
```

We can see that there is only one base pair difference between Wa-1 and the common sequence, and two base pairs that are different between the Wa-2 and the common sequence. Logically, because there are two differences between Wa-2 and the common sequence, and one difference between Wa-1 and the common sequence, and one difference between the Wa-1 and Wa-2 sequnces, Wa-1 and Wa-2 share one base pair that is different from the common sequence.

This allows us to come to a logical conclusion that Wa-1 and Wa-2 are not independent of each other. Also, seeing as there were no genomes that were equilvalent from Wa-2 from around the world, but there were 2 other genomes that were equivalent to Wa-1, this allows us to logically make an assumption that Wa-2 had to have originated from Wa-1, where Wa-1 itself mutated to have an additional one base pair difference from the common sequence.

Appendix: R Script

```
GenomeDataFrame <- read.table("coronas.txt",
                              skip = 2,
                              stringsAsFactors = FALSE) # read in data
Genomes <- GenomeDataFrame[[1]] # extract character vector
DistanceMatrix <- adist(Genomes) # get Levenshtein dist.
DistanceMatrix[1,2] # distance between 1st and 2nd entries
SplitGenomes <- strsplit(Genomes,split = "") # split into individual base pairs
SplitWa1 <- SplitGenomes[[1]]
SplitWa2 <- SplitGenomes[[2]]
MutationLocation <- which(diag(adist(SplitWa1,SplitWa2))==1) # find the location of the mismatch
substr(Genomes[1:2],MutationLocation-4,MutationLocation+4) # print out location's neighborhood
DistFromWa1 <- DistanceMatrix[,1]
table(DistFromWa1) # by how many bp's do other seq's differ from Wa-1
DistFromWa2 <- DistanceMatrix[,2]
table(DistFromWa1,DistFromWa2) # by how many bp's do other seq's differ from Wa-1 and Wa-2
duplicated(DistanceMatrix) # returns False for unique elements, True for duplicate elements
unique(Genomes) # prints out the Genomes that are unique in our set
unique(DistanceMatrix) # allows us to identify the duplicates
plot(DistanceMatrix[3,1:72]) #number of differences from the common sequence everywhere else
sum(DistanceMatrix[3,1:72]) / 72
phat <- 4/72 #storing our sample proportion
print(phat) #printing the phat in decimal form
standardErr <- sqrt((phat*(1-phat)) / 72)
print(standardErr)
left <- (phat - 2*(standardErr)) #gets the left endpoint of our CI
right <- (phat + 2*(standardErr)) #gets the right endpoint of our CI

#printing our confidence interval nicely (C-style printf statement)
sprintf("Confidence Interval: (%f, %f)", left, right)
DistanceMatrix[1,2] # number of base pair differences between Wa-1 and Wa-2
DistanceMatrix[1,3] # number of base pair differences between Wa-1 and the common sequence
DistanceMatrix[2,3] # number of base pair differences between Wa-2 and the common sequence
```