Lab Assignment 4

STA 100 | A. Farris | Spring 2020

A pdf copy of your assignment is due at the end of the day (11:59pm PT) Monday, May 11. Submission of the pdf will be through Canvas. Please put in the effort to make it look reasonably professional – you're encouraged to use R Markdown. Note that specific tasks for you are highlighted. You are free to work in groups, but you must submit your own writeup, and run your own code.

Hypothesis testing: A simple test for viral lineage

Introduction

In this case study, we will investigate the implications of a pair of early 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 an 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 (this sequence is referred to as "Wa-1").

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 would have been circulating in Snohomish county undetected in the interim, or some new carrier had 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 ("Wa-2").

Was the second case related to the first? The answer to this question has large public health implications, because it implies the undetected transmission of the disease in Washington for several weeks. However, it is possible as well that the two are not directly related, and that instead the two cases were associated with viruses transmitted independently to Washington in two incidents. If this were to be the case, an undetected outbreak may not be occurring.

Between the two genetic sequences obtained from Washington, and many others obtained by labs in already affected areas around the world, it is possible to investigate genetic variation within the virus as it spreads.

The genome of the SARS-CoV-2 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. It should be possible to track viral lineages due to these differences, but because of the randomness inherent in them, there is inherent uncertainty involved in doing so.

Investigating viral lineage

Imagine for sake of example that an outbreak has occurred in the fictional land of Laputa, where ten virus samples have been sequenced. These viruses have the genomes `GTGGGACC`, `GTGGGACC`, `GTGGGACC`, `GTGGGACC`, `GTCGGACC`, `GTGGGACC`, `GTGGGACC`, `GTCGGACC`, `GTGGGACC`, and `GTCGGACC`.

Around this time, on the 5th of Smarch, we sequence a viral sample in nearby Balnibarbi (we'll call it "Ba-1"), obtained from a traveller who had come from Laputa. This virus has the genome `GTCGGACC`.

[1] https://bedford.io/blog/ncov-cryptic-transmission/

Two weeks later, on the 19th of Smarch, we sequence a second sample (“Ba-2”) from a different patient in Balnibarbi, and find the genome GTCGAACC.

We notice that this genome is similar to the one that we sampled two weeks ago. Are these viruses related? If so, it may be that there has been otherwise undetected transmission of the virus in Balnibarbi, with substantial public health consequences. However, being as there is a known outbreak taking place in Laputa, we need to be careful to rule out the possibility that both cases independently came from Laputa in order to determine that this is really the case.

To analyze them, we read the genomes into memory in R:

```
Genomes <- c("GTCGGACC", # the first Balnibarbi sequence, Ba-1
             "GTCGAACC", # the second Balnibarbi sequence, Ba-2
             "GTGGGACC", # the 10 Laputan sequences:
             "GTGGGACC",
             "GTGGGACC",
             "GTGGGACC",
             "GTCGGACC",
             "GTGGGACC",
             "GTGGGACC",
             "GTCGGACC",
             "GTGGGACC",
             "GTCGGACC")
```

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 Balnibarbi genome sequences:

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

```
[1] 1
```

This tells us that the first two genomes in our list differ in one base pair. Indeed, comparing GTCGGACC with GTCGAACC, it is clear that the difference is in the fifth base pair, which was G for the first and A for the second Balnibarbi sequences.

For sequences with hundreds or thousands of base pairs, it is harder to compare sequences by eye like this: so here is how we might look at this difference programmatically.

```
SplitGenomes <- strsplit(Genomes,split = "") # split into individual base pairs
SplitBa1 <- SplitGenomes[[1]]
SplitBa2 <- SplitGenomes[[2]]
MutationLocation <- which(diag(adist(SplitBa1,SplitBa2))==1) # find the location of the mismatch
MutationLocation # print out the location of the difference
```

```
[1] 5
```

```
substr(Genomes[1:2],MutationLocation-2,MutationLocation+2) # print out the neighborhood of this location
```

```
[1] "CGGAC" "CGAAC"
```

From this we can see again that one base pair (the 5th) in the first Balnibarbi case has G, and the same location for the second case has A. If the two cases are related, then this mutation has occurred over the time that has transpired between the two cases.

To compare Ba-1 and Ba-2 with the genomes from Laputa:

```
DistFromBa1 <- DistanceMatrix[,1]
DistFromBa2 <- DistanceMatrix[,2]
table(DistFromBa1,DistFromBa2)
```

```
           DistFromBa2
DistFromBa1 0 1 2
          0 0 4 0
          1 1 0 7
```

4+1+7=12

This table tells us how many genomes, among all 12 of the genomes in the list, differ from Ba-1 and Ba-2 by different numbers of base pairs. It tells us that, out of 12 sequences, 0 differ both from Ba-1 by 0 base pairs and Ba-2 by 0 base pairs (this is to be expected, as the two differ themselves by one base pair!).

Only one of the twelve differs from Ba-1 by one base pair and by Ba-2 by none (this must be the Ba-2 sequence itself!).

Four of them differ from Ba-1 by none and from Ba-2 by one; so, there must be three Laputan sequences identical to Ba-1 (plus Ba-1 itself).

Of the remaining 7 sequences, all must be Laputan, and they all differ from Ba-1 by one base pair and from Ba-2 by two.

In fact, we can also see that the other 7 sequences from outside of Balibarbi are identical, in that they differ from one another by zero base pairs:

```
DistFromOther <- DistanceMatrix[,3]
table(DistFromOther)
```

```
DistFromOther
0 1 2
7 4 1
```

Thus we find that the Ba-1 sequence is the same as three others from Laputa, and that the Ba-2 sequence differs from these four by one base pair and from all others by two. So, these five sequences share one mutation relative to the other seven.

A hypothesis test for viral lineage

Let us now judge what we can about whether this Ba-1 case was connected to Ba-2. On the one hand, the genome sequences that we have for Ba-1 and Ba-2 share a mutation that is not present in most of the Laputan cases. On the other hand, a few of the Laputan cases do share this mutation, so it is possible that it did arrive in Balnibarbi a second time by chance.

Let's assume that the similarity was, in fact, due strictly to chance, and both Balnibarbi cases were transmitted independently from Laputa. This will be our *null hypothesis*. In this case, we can assume that the second Balnibarbi case was a random selection from the genomes in Laputa. The proportion of Laputan cases that share the mutation that is in common between Ba-1 and Ba-2 is then the *p-value*; if this value is small, it will reflect evidence against the assumption of the null hypothesis. We can use a significance level of $\alpha = 0.05$ to assess this p-value.

Because we don't know the proportion of Laputan cases that share this mutation (because in turn we don't know exactly what all of the viruses in Laputa look like), we do not know the p-value exactly. However, we can estimate it. If we assume viruses sequenced to be sampled independently, we would estimate the proportion of the viruses that share this mutation outside of Balnibarbi to be $\frac{3}{10} = 0.3$, with estimated standard error $\sqrt{\left(\frac{3}{10}\right)\left(1 - \frac{3}{10}\right)\left(\frac{1}{10}\right)} \approx 0.145$. Because the estimated p-value is not smaller than α, we can take it to reflect a lack of evidence against the null hypothesis that the two cases of Covid-19 in Balnibarbi were independent transmissions of the disease from elsewhere. That is, we do not have enough evidence to rule out the possibility that the two cases were associated with independent transmissions from Laputa.

p(1-p)/n

Back to the real world

As Basil Fawlty says, "Back to the world of dreams."

It is now early March in Washington state; an outbreak is occuring in China and elsewhere, but there is no community transmission that has yet been identified locally. But, as we've seen, a new case (Wa-2) has arisen. What can it tell us?

We'll begin our analysis by reading the contents of the file `corona-genomes.txt` into memory:

```
GenomeDataFrame <- read.table("http://www.stat.ucdavis.edu/~affarris/corona-genomes.txt",
                              skip = 2,
                              stringsAsFactors = FALSE)
Genomes <- GenomeDataFrame[[1]] #extract character vector   character vector of length 72
```

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 object `Genomes` contain part of the genome from each of the first and the second cases in Washington (Wa-1 and Wa-2), respectively, much like the object by the same name in the Balnibarbi example above. The remaining rows contain part of the genome from each of seventy other genomes obtained by early March from other parts of the world that were affected earlier, mostly from China and Vietnam.

Replicate the analysis above for the Washington cases:

How do the Wa-1 and Wa-2 sequences differ from one another?

Do you have enough evidence to rule out the possibility that the two Washington cases were transmitted to Washington independently? To answer this, carry out a hypothesis test, using a significance level of 0.05. Briefly state your Null hypothesis, report an (estimated) p-value, and determine whether or not you reject your null hypothesis. Does this suggest that undetected community transmission was occuring in Washington between January and March?

---

[2] https://www.gisaid.org/

Appendix: R Script

```r
Genomes <- c("GTCGGACC", # the first Balnibarbi sequence, Ba-1
             "GTCGAACC", # the second Balnibarbi sequence, Ba-2
             "GTGGGACC", # the 10 Laputan sequences:
             "GTGGGACC",
             "GTGGGACC",
             "GTGGGACC",
             "GTCGGACC",
             "GTGGGACC",
             "GTGGGACC",
             "GTCGGACC",
             "GTGGGACC",
             "GTCGGACC")
DistanceMatrix <- adist(Genomes) #get Levenshtein dist.
DistanceMatrix[1,2] # distance between 1st and 2nd entries
SplitGenomes <- strsplit(Genomes,split = "") # split into individual base pairs
SplitBa1 <- SplitGenomes[[1]]
SplitBa2 <- SplitGenomes[[2]]
MutationLocation <- which(diag(adist(SplitBa1,SplitBa2))==1) # find the location of the mismatch
MutationLocation # print out the location of the difference
substr(Genomes[1:2],MutationLocation-2,MutationLocation+2) # print out the neighborhood of this location
DistFromBa1 <- DistanceMatrix[,1]
DistFromBa2 <- DistanceMatrix[,2]
table(DistFromBa1,DistFromBa2)
DistFromOther <- DistanceMatrix[,3]
table(DistFromOther)
GenomeDataFrame <- read.table("http://www.stat.ucdavis.edu/~affarris/corona-genomes.txt",
                              skip = 2,
                              stringsAsFactors = FALSE)
Genomes <- GenomeDataFrame[[1]] #extract character vector
```