Assignment 3

Rory Sarten 301005654

22 September, 2020

Question 1

a)

A Runs Test tests a set of binary variables $X_1, ..., X_n$ to verify if the variables occur randomly.

H_0: variables occur randomly, i.e. knowing $X_1, ..., X_n$ does not help predict X_{n+1}.

H_A: variables are not random, i.e. knowing some part of the sequence can help predict subsequent variables.

As the variables are binary, they will take the value 0 or 1. The number of 0s is n_0 and the number of 1s is n_1, where:

$$n_0 = n - \sum_{i=1}^{n} X_i$$

$$n_1 = \sum_{i=1}^{n} X_i$$

To perform a Runs Test the observations are combined into one collection of $n = n_0 + n_1$ observations and arranged in increasing order of magnitude or observation. They are labeled according to which set they originally came from. A run is a group of two or more sequential values of 0 or 1.

Let R denote the number of runs in the combined ordered sample of $X \in \{0, 1\}$. Under H_0, R can be approximated as a normally distributed random variable, assuming both n_0 and n_1 are sufficiently large.

$$\bar{R} = \frac{2n_0n_1}{n} + 1$$

$$Var(\bar{R}) = \frac{2n_0n_1(2n_0n_1 - n)}{n^2(n-1)}$$

With test statistic $Z = \dfrac{R - \bar{R}}{\sqrt{Var(\bar{R})}}$ where $Z \sim N(0, 1)$

b)

```
## 0 healthy, 1 has disease
X <- "HHHDDDHDDHHHHHDHHHHHHHHHHHHDDDHHHHDDDHHHHDHHHDHH" %>%
  stringr::str_split("") %>% unlist() %>% `==`("D") %>% as.integer()

n_0 <- length(X) - sum(X)
n_1 <- sum(X)

(2 * n_0 * n_1)/length(X) + 1
```

```
## [1] 20.65957
```