## 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.

$$R = 1 + \sum_{i=2}^{n} I_{(X_i, X_{i-1})}$$
, where  $I_{(X_i, X_{i-1})} = 0$  if  $X_i = X_{i-1}$  and  $I_{(X_i, X_{i-1})} = 1$  if  $X_i \neq X_{i-1}$ 

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

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

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

b)

A small number of runs (a small value for R) would indicate that  $X_i$  is more likely to be the same as  $X_{i-1}$ . A large number means that X is fluctuating regularly between values and X is less likely to be the same as  $X_{i-1}$ .

c)

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

n <- length(X)
n_0 <- n - sum(X)
n_1 <- sum(X)</pre>
```

```
R \leftarrow 1 + sum(X[2:n] != X[1:n-1])
R_{est} \leftarrow 1 + 2*n_0*n_1/n
R_{var} \leftarrow (2*n_0*n_1*(2*n_0*n_1 - n))/(n^2*(n - 1))
Z <- (R - R_est)/sqrt(R_var)</pre>
p <- pnorm(Z)</pre>
cbind(R, R_est, R_var, Z, p)
          R
                R_est
                          R_var
## [1,] 15 20.65957 7.974767 -2.004125 0.02252834
xs <- data.frame(x = seq(-4, 4, length.out = 100)) %>% dplyr::mutate(density = dnorm(x))
ggplot(xs, aes(x, density)) +
  geom_area(fill = "firebrick") +
  gghighlight(x < Z)</pre>
   0.4 -
   0.3 -
density
   0.1 -
   0.0 -
                                                                          2
                               -2
                                                     Ö
                                                                                               4
          -4
                                                     Χ
```

With p = 0.0225 we reject  $H_0$  at the 5% level. The values are not randomly ordered.

d)