P8130 Biostatistical Methods Homework 5

Emil Hafeez (eh2928)

11/13/2020

Problem 1

```r
#Read the CSV data into a dataframe
antibodies_df <- read.csv("./data/Antibodies.csv")

# Make AgeCategory, Smell, and Gender appopriate datatypes. Helps to ensure we know all unique values,
unique(antibodies_df$AgeCategory)
    antibodies_df <- antibodies_df %>%
        mutate(AgeCategory = factor(AgeCategory, labels = c("18-30", "31-50", "51+") ))
unique(antibodies_df$Smell)
    antibodies_df <- antibodies_df %>%
        mutate(Smell = factor(Smell, levels = c("Normal","Altered", "Unanswered/Others")))
    antibodies_df = antibodies_df %>% filter(Smell != "Unanswered/Others")
unique(antibodies_df$Gender)
    antibodies_df <- antibodies_df %>%
        mutate(Gender = factor(Gender, levels = c("Male","Female")))
```

In order to assess the difference in IgM levels between the two smell factor groups (Normal vs Altered) given non-normal distributions, we opt for a non-parametric test called the Wilcoxon Rank-Sum test. It's the nonparametric equivalent of the two-sample independent t-test, and examines if the medians of the two populations are equal versus not equal:

$H_0 =$ the medians of the two groups' IgM levels are equal, and $H_A =$ the medians of the two groups' IgM levels are not equal. The decision rule is that we reject $H_0 =$ if $T > z_{1-(\alpha/2)}$.

The test statistic is computed, with a continuity correction, one of two ways. We first combine the data from the two groups, order the values from lowest to highest, assign ranks to the individual values (1 to n), and if ties, assign the average rank. Then, select a group and compute the sum of ranks T_1 for the first group, and then use the appropriate test statistic formula.

With no ties (referring to two equally ranked values once the values are listed), the test statistic is

$$T = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)(n_1+n_2+1)}}$$

and with ties, the test statistic is

$$T = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)\left[(n_1+n_2+1) - \sum_{i=1}^{g} t_i\left(t_i^2-1\right)/(n_1+n_2)(n_1+n_2-1)\right]}}$$

where t_i refers to the number of observations with the same absolute value in the i^{th} group and g is the number of tied groups.

In our case, the test statistic calculated by R is slightly different, since it does not by default add the $n_1(n_1+1)/2$ term (and is denoted by W).

The p-value under the normal approximation, with n_1 and $n_2 \geq 10$ is described by $2 * [1 - \Phi(T)]$.

```r
antibodies_df2 =
  antibodies_df %>%
  pivot_wider(
        names_from = Smell,
        values_from = Antibody_IgM
    )

wilcox.test(antibodies_df2$Normal, antibodies_df2$Altered, mu = 0)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  antibodies_df2$Normal and antibodies_df2$Altered
## W = 5836, p-value = 0.01406
## alternative hypothesis: true location shift is not equal to 0
```

In context, and ignoring missing values and the unanswered smell category, we find evidence to reject the null hypothesis and conclude that the true location shift between the Normal and Altered smell categories is not equal to zero (in other words, the median IgM values are different for the two groups).

Problem 2

Let's come back to this.

Problem 3