4. Assignment NMST 539

Frantisek Helebrand

March 2022

1 Exploratory analysis

For this assignment we will us *ushealth* data available in the SMS package. Primarly we will focus on region. In this paper we distinguish two main regions - West (Midwest and West) and Other (South and Northest). We will manly focus on number of cardiovascular cases, number of cancer cases and number of pneumonia flu cases.

In table 1 we can see sample means for west region. On the other hand table 3 shows sample means for other regions. We can also plot boxplots with respect to the region. These boxplots can be seen in figures 1, 2, 3.

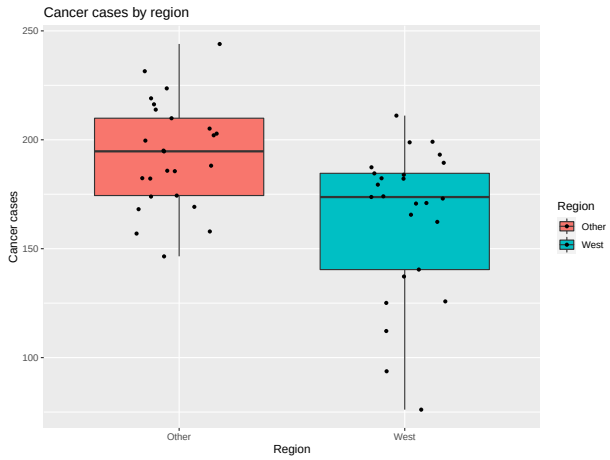

Figure 1: Bowplots of cancer cases with respect to region.

Cancer	Cardiovascular disease	Pneumonia flu
163.68	21.62	366.99

Table 1: Means for West region.

Cancer	Cardiovascular disease	Pneumonia flu
193.13	20.46	430.07

Table 2: Means for Other regions.

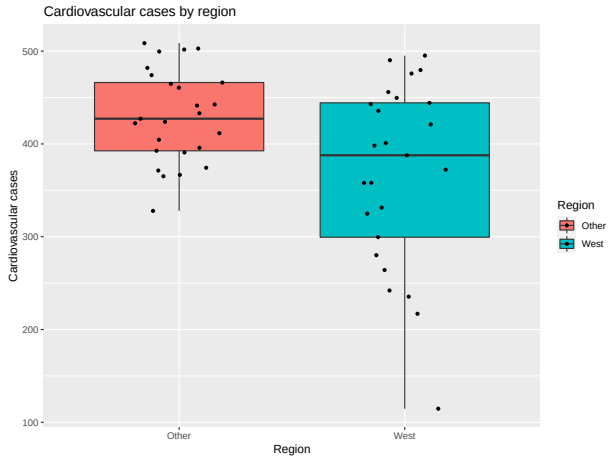

Figure 2: Bowplots of cardiovascular diseases with respect to region.

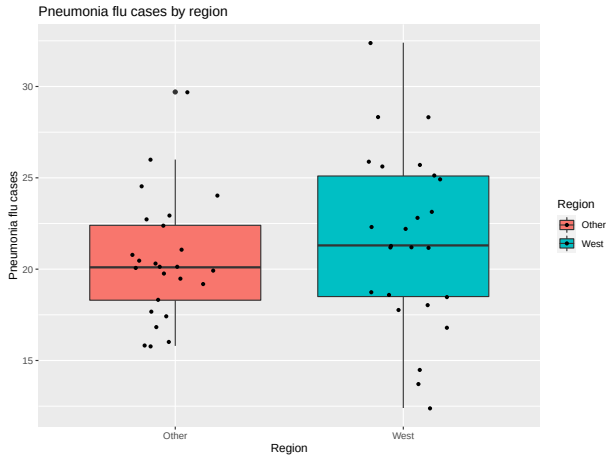

Figure 3: Bowplots of pneumonia flu cases with respect to region.

2 Second part

We assume that we have two multivariate samples $\boldsymbol{X_1}, \ldots, \boldsymbol{X_{25}} \sim N_3(\boldsymbol{\mu_X}, \Sigma)$ for west region and $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_{25}} \sim N_3(\boldsymbol{\mu_Y}, \Sigma)$ for east region.

We are interested in testing hypothesis that the means are equal i. e null hypothesis and alternative are

$$H_0 : \boldsymbol{\mu_X} = \boldsymbol{\mu_Y}, \quad H_1 : \neg H_0$$

We will use the two-sample test for two multivariate. The test statistic is

$$\frac{nm(n+m-p-1)}{p(n+m)^2} \left(\overline{\boldsymbol{X}}_{25} - \overline{\boldsymbol{Y}}_{25}\right)^T \mathbf{S}^{-1} \left(\overline{\boldsymbol{X}}_{25} - \overline{\boldsymbol{Y}}_{25}\right)$$

where $n = m = 25$ and $p = 3$. The $\mathbf{S} = \frac{1}{2}(S_1 + S_2)$ where S_1, S_2 are covariance matrices of multivariate sample $\boldsymbol{X_1}, \ldots, \boldsymbol{X_{25}}$, $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_{25}}$ respectivelly. The coresponding $p-$value is 0.00041 and we can reject null hypothesis. So we can say that the meeans are not equal and the number of diseases ´depends on the region.

For the difference (Other-West) we can construct simultaneous confidence intervals. We will obtain

Disease	Lower Boundary	Upper Boundry	Mean Estimate
Cancer	16.04	42.86	29.45
Pneumonia	-3.79	1.46	-1.16
Cardiovascular	19.71	106.44	63.07

Table 3: Simultaneus confidence intervals.

We can see that two out of three intervals does not overlap zero. The widest interval is for cardiovascular diseases.