STAT0004, 2021: cover page

All group members contributed equally throughout.

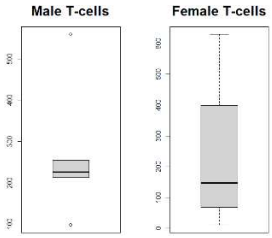
STAT0004, 2021: written report

Task 1

In this report, patient data from 23 donors was provided out of which cell data was given for 12 of them. This report looks at the varying relationships between age, gender, cell location and type of cell as well as provide different critical characteristics evident in the data. All analysis done is strictly on the sample rather than on the whole population. Furthermore, all correlation calculations are done using the Spearman method.

Most of the donors found in the data are young with only two being above 25 years old. Due to these extreme values, the mean age is greater than the median. This is also evident from the strong positive skewness in age (1.700). There are exactly 6 male and 6 female donors.

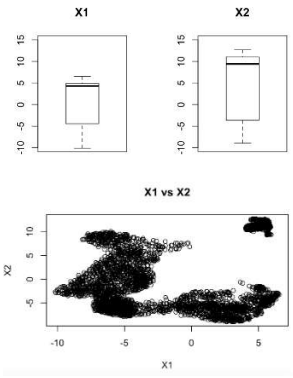
For the whole sample, the mean number of T cells (248.2) and B cells (246.58), per person are very similar, implying there is an almost equal number of T and B cells per person. The high positive skewness (1.440) of B cells compared to a more moderate positive skewness (.7724) in T cells shows that it is more likely for a person to have a low number of B cells rather than a low number of T cells. T-cells and age have a small positive correlation (0.3958) which might suggest a slight increase in the number of T-cells as age increases, however it is not definitive.



When separating the sample by gender, females(F) and males(M) have a higher median for T cells (F:147.5, M:226.5) than B cells (F: 34.00, M:139.50), indicating that both genders have more T cells than B cells. The boxplot shows that both genders have extreme values which might affect the summary statistics and alter the actual skewness of the data. B cells for both genders are highly positively skewed (F: 1.346, M: 1.147). This implies that regardless of the gender, a higher proportion of people have a low number of B cells.

Female B-cells have a strong positive correlation with age (0.7537). From face value, we may deduce that as the age of female donors increases, the number of B-cells increases. However, there is an extreme point which is affecting the correlation. When excluding this value, it is seen that there is a more moderate correlation of 0.5642. Males also have a strong positive correlation for B-cells with (0.8857) or without (0.8000) the extreme value. Since the values are very similar this might suggest that the extreme value is not altering any summary statistics. This may also suggest that this value is not extreme after all. Also, males have a larger variance than females for B cells (M: 226800, F:167100), whereas females have a much larger variance than males for T cells (M:2406, F:55110). T cells have a larger variance than B cells overall.

X1 and X2



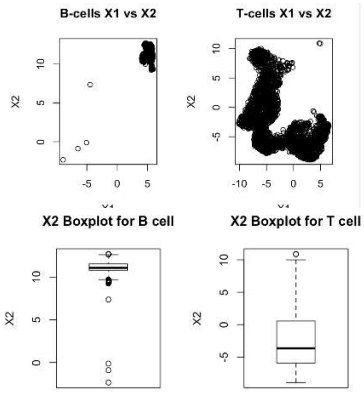
For the whole population X2 values are higher than X1 values. X1 has a median of 4.307 and a mean of 1.064. X2 has a median of 9.429 and a mean of 4.471. It can also be seen that X2 has a greater range (21.629) and IQR (14.737) than the range (16.691) and IQR (9.326) of X1. In the box plot, the bottom 50% of the data is much more spread out than the top 50%. In addition to this, the median being greater than the mean for X1 and X2, can also imply negative skewness. The results that are obtained prove this, with -0.6908 skewness for X1 and -0.4162 for X2. It is noted that although the correlation value (0.6597) suggests that as the X1 values increase X2 values increase, the scatter plot does not show any clear relationship between X1 and X2.

When separating by gender, the medians of X1 and X2 are higher than the means. The median and mean of Male X1 (median: 4.432, mean: 1.269) and X2 (median: 10.733, mean: 5.004) is greater than that of Female (X1: median: 4.1737, mean: 0.7955, X2: median:6.725, mean: 3.770). This shows that

on average males have higher X1 and X2 values than females. Males also have a more negative skewness in X1(-0.7392) and X2 (-0.5085) than Females, X1(-0.6284), X2(-0.3049). This indicates that X1 and X2 values of females are more concentrated towards zero than those of males

.

When separating by cell type, there is a clear difference between B cells and T cells summary statistics for X1 and X2. This can be seen through the scatter plots. For the X1 against X2 scatter plot, cell type B forms a cluster in the top right corner, while cell type T forms an L shape in the bottom left corner. The X1 and X2 coordinates of B cells are positively correlated, while for T cells they are negatively correlated. The median of X1 for B-cells is 4.774 and for T-cells is -4.423. The median of X2 for B cells is 11.064 and for T cells is -3.658. These results show that the values of X1 and X2 are generally higher for B cells than for T cells. In addition, the medians and quartiles also display that the values for X2 are higher than those of X1 when separating by either B-cells or T-cells.



The difference between the mean and median indicates a positive skewness for both cell types. However, when calculating the skewness, X1 and X2 for T-cells have positive values (X1: 0.6005, X2: 0.8901), whereas B-cells have negative values (X1: -8.651, X2: -4.681). This could be due to the presence of extreme values seen in the plot, affecting one measure of skewness more than the other. The variances of B-cells are X1(0.3679) and X2(0.5418) and of-T cells are X1(18.82) and X2 (23.48). The large differences suggest that X1 and X2 coordinates for T cells are more spread out than that of B-cells.

Task 2

A contingency table was created.

	Male	Female	Total
T cells	1577	1402	2979
B cells	1797	1162	2959
Total	3374	2564	5938

The biologist claims that the ratios of T/B cells are not equal between males and females. In order to test that, a null hypothesis is set where it is assumed that the difference between the T/B cell ratios of males and females is zero. This is then tested against the alternative hypothesis that the difference is not equal to zero. A rejection region was created at the 5% level of significance. A rejection region is a range of values within which the null hypothesis is rejected. The test statistic shows how different the observed data behave from what was expected under the initial hypothesis. The sample test statistic was -.9640 and this value falls outside the rejection region. As a result, under the initial assumption, there is no sufficient evidence to reject the null hypothesis at the 5% level, to suggest that there is a difference between the two ratios. We are thus unable to confirm the biologist's claim.

Task 3

The claim made by the biologist is that there is a monotonic relationship between the percentage of B-cells and the age. First, a null hypothesis of no monotonic relationship is set. This is tested against an alternative hypothesis that there is a monotonic relationship. Assuming the null hypothesis is true, a Spearman correlation test is conducted, which tests whether such a relationship exists. When performing the test, a p-value of 0.001 was calculated. This means that the probability of getting the correlation that was found or a more extreme value, given the assumptions, is 0.1%. The p-value is less than 0.05, so the assumption of no monotonic relationship can be rejected at the 5% significance level. The data therefore supports the biologist's claim. The percentage of B cells for each donor are:

F22	A16	F67	C34	T03	C40	F64	T06	T07	C41	A43	F74
1.449	61.99	14.378	0	28.141	7.317	4.436	32.41	51.89	2.703	92.65	1.389