Cancer Incidence Analysis

Andrew Diaz, Edwin Jaison, Jenny Dang, Zena Yacout

PROBLEM:

Does cancer incidence differ among ethnicities in Canada?

INTRODUCTION:

The topic of cancer has become more and more of a highly prevalent topic in contemporary society. This could be due to the overall increase in cancer cases, as a result of a growing and aging population worldwide, and a variety of other factors. The National Cancer Institute defines cancer as "a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body." As cancer is a disease which can affect anyone and everyone, it is imperative that society as a whole understands the disease. Many fields of academia can be applied towards understanding cancer better. The science of statistics can be applied towards cancer research by contributing to understanding the trends behind cancer. Understanding parameters such as cancer incidence, survival, and mortality allow for statistical inferences to be made, allowing for overall trends to be studied by cancer researchers. The subject of cancer incidence is highly important in the field of cancer research, as this parameter can help identify emerging patterns for the disease, as well as assess the progress and efforts that have been made to help combat the disease. Canada as a country is known as a multicultural and multi-ethnic country in the global community, being inhabited by numerous different cultures and ethnicities. Therefore, it made it interesting to see a comparison between different ethnicities in Canada and statistics covering their cancer incidence. The problem that will therefore be addressed in this paper will be **"does cancer incidence differ among ethnicities in Canada?"** This question is important in the field, as it will allow for underlying patterns and trends to surface, allowing for the field of cancer research to accommodate to alleviate the negative patterns. The statistics and analysis gained from the research question/problem will also allow for different fields of academia to contribute to alleviating the negative trends. Social science fields such as sociology can contribute by observing how social structures, processes, and cultural phenomena contribute to the parameters of cancer incidence. Cancer is not only a medical and scientific issue, it is also a humanitarian and social issue. Statistics allows for the bridging of these two realms, allowing for a deeper understanding of the disease and better research to be conducted. This paper will be focusing on the three most common types of cancers present in Canada: lung cancer, colorectal cancer, and Non-Hodgkin lymphoma.

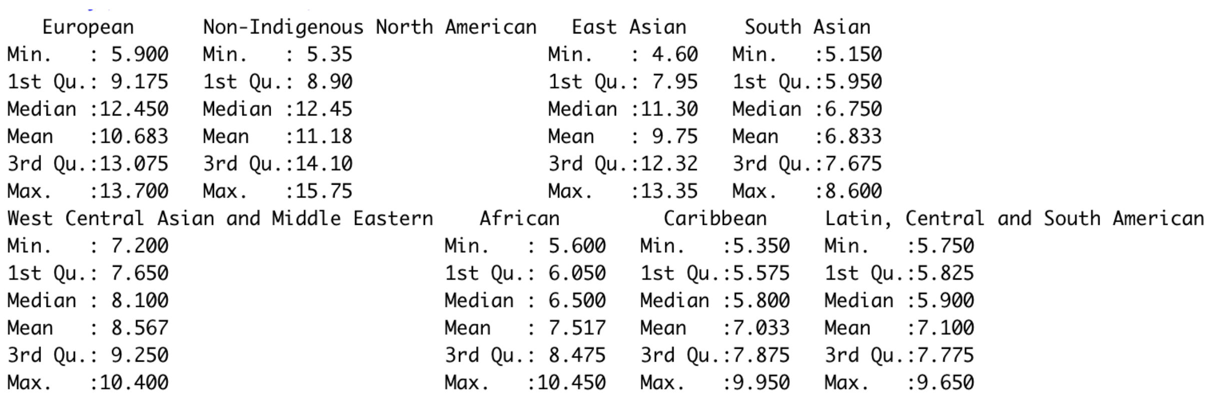
DISCUSSION:

The study's sample is composed of three different cancer types: lung cancer, colorectal cancer, and Non-Hodgkin lymphoma. The samples are drawn from a population of eight different ethnicities. Ethnicity can be often confused with nationality or citizenship but in this sampleset, it refers to the ethnic or cultural origins of a person's ancestry, therefore there can be instances in which a person could identify with one or more ethnicities (Hwee and Evelynne 2021). The ethnicities are categorized as mutually inclusive groups: European, non-Indigenous North American, East Asian, South Asian, West Central Asian and Middle Eastern, African, Caribbean, Latin, Central, and South American (Hwee and Evelynne 2021). The percent sample was obtained through a stratified sampling method which is a form of random sampling. Stratified sampling can be defined as a sample that contains naturally-occurring groups of observation known as stratas and each strata have similar characteristics. Stratified sampling is effective in this sort of sampling as each ethnicity within the population sample can be considered a strata that have one or more similar characteristics. In this case, the characteristics are the three different cancer types that may exist in each ethnicity group/strata. The population study was conducted on May 16th, 2006, where 20% of households were sent questionnaires about education, ethnicity, income, mobility, and employment (Hwee and Evelynne 2021). One drawback of this selection method is that due to overlapping, it may be difficult to classify each individual in the population into a strata. Since the ethnicities are considered the stratas in this dataset, other potential confounding variables such as income, mobility and employment play a role in the data collected but have to be omitted when performing any statistical inference tests. Because this sample is random, people can fall into many categories thus losing the sample's effectiveness. The data is an open source dataset which is made up of quantitatively classified variables that are measured with a ratio scale. Because the samples are classified in the ratio scale, percentages between ethnicities can be effectively compared. The ethnicities recorded are the explanatory variable, while the rates of lung, colorectal, and Non-Hodgkin Lymphoma cancer incidences are the response variables.

The data on cancer incidence was collected through the Canadian Cancer Registry (CCR), a national registry that includes all information on newly diagnosed primary cancers since 1992. From the 20% of households that were obtained, only individuals that reportedly had cancer were used in the sample obtained by the CCR. Therefore a method of non-response bias was introduced in the sample where non-cancer related households/individuals were omitted from the data in order to distribute the cancer cases that were reported (Hwee and Evelynne 2021). The cancer cases were distributed among top 10 cancer types of which the top 3 cancer types that occur in both sexes were used in statistical inference tests. In order to determine any variation in cancer incidence solely between the ethnicities, the mean of Male and Female cancer incidence percentage was further calculated and used for statistical inference tests. According to this new consolidated data, non-Indigenous North Americans had the highest incidence rate of lung cancer at 15.75% while South Asians had the lowest incidence rate of lung cancer at 5.15%. In terms of colorectal cancer, East Asians had the highest incidence rate at 13.35% while South Asians had the lowest incidence rate at 8.6%. For Non-Hodgkin, Western Central Asian/Middle Eastern had the highest incidence rate at 7.2%, while East Asians had the lowest incidence rate at 4.6%. An interesting observation from this consolidated data is the difference in percentages between the different ethnicities. Perhaps as a follow up to the data examined, it would be interesting to see how cultural, societal, and economic factors affect the incidence of cancer for the different ethnicities. How does the general lifestyles of different ethnic groups affect the parameter of cancer incidence? This way of looking at cancer is another reason why cancer could be viewed as not only a scientific and medical issue, but a humanitarian one as well. Perhaps by observing and studying the fundamental differences in the ways of living and circumstances of ethnic groups, a greater understanding of why certain cancers are more prevalent in certain ethnic groups could be unraveled.

The EDA values for each variable of interest are provided below (cancer incidence among different ethnicities). Because the data is made up of quantitative variables, the central tendency and spread must be described. One can view the values used to represent the data by using the R code summary(cancerethnicities) for this particular dataset. This code allows R to calculate the five-number summary which gives the minimum observation (x_{min}), the first quartile(Q_1), the median ($\tilde{x}$), the third quartile (Q_3), and the maximum observation (x_{max}). Furthermore, using the five-number summary presented below, the range (x_{max}-x_{min}) and the interquartile range (Q_3-Q_1) can be determined. Measures of spread are used to describe the variation of the observation, and the range and the interquartile range is used to do so.

Five-number Summary:



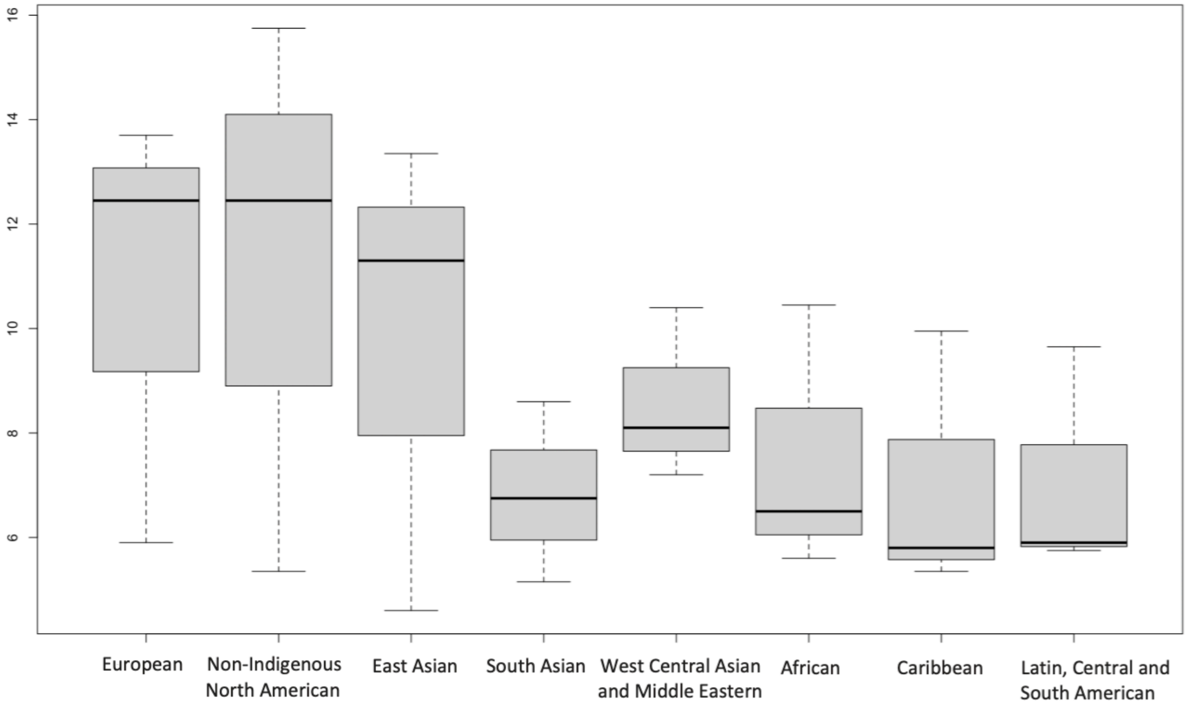
```
   European      Non-Indigenous North American   East Asian     South Asian
Min.   : 5.900   Min.   : 5.35                  Min.   : 4.60  Min.   :5.150
1st Qu.: 9.175   1st Qu.: 8.90                  1st Qu.: 7.95  1st Qu.:5.950
Median :12.450   Median :12.45                  Median :11.30  Median :6.750
Mean   :10.683   Mean   :11.18                  Mean   : 9.75  Mean   :6.833
3rd Qu.:13.075   3rd Qu.:14.10                  3rd Qu.:12.32  3rd Qu.:7.675
Max.   :13.700   Max.   :15.75                  Max.   :13.35  Max.   :8.600
West Central Asian and Middle Eastern    African         Caribbean     Latin, Central and South American
Min.   : 7.200                        Min.   : 5.600   Min.   :5.350  Min.   :5.750
1st Qu.: 7.650                        1st Qu.: 6.050   1st Qu.:5.575  1st Qu.:5.825
Median : 8.100                        Median : 6.500   Median :5.800  Median :5.900
Mean   : 8.567                        Mean   : 7.517   Mean   :7.033  Mean   :7.100
3rd Qu.: 9.250                        3rd Qu.: 8.475   3rd Qu.:7.875  3rd Qu.:7.775
Max.   :10.400                        Max.   :10.450   Max.   :9.950  Max.   :9.650
```

Measures of spread:

- European: range = 7.8000 IQR = 3.900
- Non-Indigenous North American: range = 10.4000 IQR = 5.2000
- East Asian: range = 8.7500 IQR = 4.3700
- South Asian: range = 3.4500 IQR = 1.7250
- West Central Asian and Middle Eastern: range = 3.2000 IQR = 1.6000
- African: range = 4.8500 IQR = 2.4250
- Caribbean: range = 4.6000 IQR = 2.3000
- Latin, Central and South American: range = 3.9000 IQR = 1.9500

A boxplot was chosen as the graphical representation because quantitative variables are being studied and the five-number summary was used. This type of graph is helpful for identifying any potential outliers. Outliers are values that are either extremely large or small in comparison to the rest of the values. One can find outlier values using these two formulas: Q_3 + 1.5(IQR) and Q_1- 1.5(IQR). The R code used to create the boxplot is boxplot(cancerethnicities). Boxplots can be used to visualize the general distribution shape of data. The graph for quantitative data can be characterized in terms of symmetry or skew. A symmetric graph appears the same on both sides, indicating that the mean and median are almost or perfectly identical, while a skewed graph can have a tail pointing to the right or left, indicating that the mean and median will differ. The first three boxplots have a left-skewed shape distribution, the centre box plot is symmetric, and the last four boxplots have a right-skewed shape distribution.

Graphical Representation:



Outlier values based on the boxplot above:

- European: outliers are any values greater than 18.9250 and less than 3.3250
- Non-indigenous North American: outliers are any values greater than 21.9000 and less than 1.1000
- East Asian: outliers are any values greater than 18.8750 and less than 1.3950
- South asian: outliers are any values greater than 10.2625 and less than 3.3625
- West Central and middle east: outliers are any values greater than 11.6500 and less than 5.2500
- African: outliers are any values greater than 12.1125 and less than 2.4125
- Caribbean: outliers are any values greater than 11.3250 and less than 2.1250
- Latin, central, South America: outliers are any values greater than 10.7000 and less than 2.9000

ANALYSIS:

There are eight populations (eight different ethnicities) used to determine the relationship between cancer incidences and ethnicities. Each population consists of three observations (three types of cancer). Since the research question examines the differences in cancer incidences within different ethnicities, the most appropriate statistical analysis to use is one-way Analysis of Variance (ANOVA). ANOVA is a technique that can be used to compare the means of two or more populations and includes a hypothesis testing procedure. ANOVA requires that the samples be independent. This condition is fulfilled as each cancer type is independent from the others. Additionally, the variances of the population must be equal (homoscedastic) in order to conduct ANOVA. Bartlett's test can be used to assess homoscedasticity prior to conducting ANOVA. It is important to note that α=0.05 when interpreting the numerical output in the context of the formal hypothesis statements for this data set.

-Bartlett's test-

Appropriate hypothesis statement for Bartlett's homoscedasticity:

H_0:$\sigma^2_{Lung\ cancer} = \sigma^2_{Colorectal\ Cancer} = \sigma^2_{Non-Hodgkin\ lymphoma\ Cancer}$

H_a: at least one of the k variances (cancer types) differs

Since there are eight populations in the dataset, the dataset needs to be stacked prior to performing Bartlett's test to test for homoscedasticity. The following two functions in R Studio can be used to stack and perform Bartlett's test on this set of data:

```
stack(x)
```

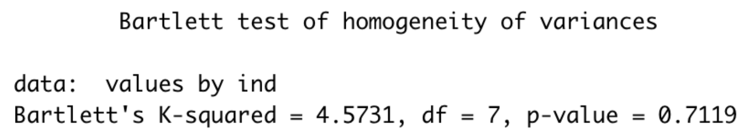
(x is the k-column data set and each column represents one observation)

```
bartlett.test(x~group)
```

(x is the column containing the quantitative variable and group contains the names of the observation in x)

R Studio Output:

```
y= stack(cancerethnicities)
attach(y)
bartlett.test(values~ind)
```



```
        Bartlett test of homogeneity of variances

data:  values by ind
Bartlett's K-squared = 4.5731, df = 7, p-value = 0.7119
```

P value: 0.7119

Since the p-value (0.7119) > 0.05, the condition of equal variance (homoscedasticity) is met. Therefore we fail to reject H_0:$\sigma^2_{Lung\ cancer} = \sigma^2_{Colorectal\ Cancer} = \sigma^2_{Non-Hodgkin\ lymphoma\ Cancer}$ and can proceed to perform ANOVA on this data set.

-ANOVA-

Appropriate hypothesis statement to determine if cancer incidences differ among ethnicities:

H_0: $\boldsymbol{\mu}_{European} = \boldsymbol{\mu}_{Non\ Indigenous\ North\ American} = \boldsymbol{\mu}_{East\ Asian} = \boldsymbol{\mu}_{South\ Asian} = \boldsymbol{\mu}_{West\ Central\ Asian\ and\ Middle\ Eastern} = \boldsymbol{\mu}_{African} = \boldsymbol{\mu}_{Caribbean} = \boldsymbol{\mu}_{Latin,\ Central\ and\ South\ American}$

H_a: at least one $\boldsymbol{\mu}$ differs

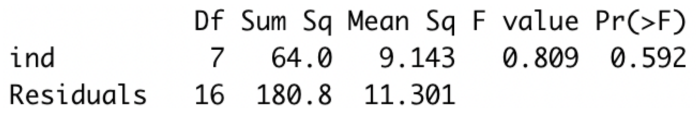
The following function can be used to perform a one-way ANOVA:

summary(aov(x~group))

(x is the column containing the quantitative variable and group contains the names of the observation in x)

R Studio Output:

summary(aov(values~ind))



```
            Df Sum Sq Mean Sq F value Pr(>F)
ind          7   64.0   9.143   0.809  0.592
Residuals   16  180.8  11.301
```

Test statistic: 0.809

P value: 0.592

Since the p-value (0.592) ≥ 0.05, we fail to reject the H_0 and conclude that the average cancer incidence rate amongst the ethnicities are not significantly different.

CONCLUSION:

Using the Bartlett's Test, a p-value of 0.7119 was obtained. Since the value of 0.7119 was greater than or equal to 0.05, the condition of equal variance or homoscedasticity is met and the incidence rates of the three cancer types are statistically similar, therefore the null hypothesis is accepted and the method ANOVA can be performed on the dataset. As said in the analysis, homoscedastic means that the variances of the populations were equal. This is a prerequisite for ANOVA testing. Through the ANOVA test, a p-value of 0.592 is obtained. Since the p-value of 0.592 is greater or equal to 0.05, the null hypothesis is rejected, therefore the average incidence rates do not significantly differ. Even though the average mean percentages of the cancer cases determined through EDA (numerical and graphical summaries of the sample data) differ for each ethnicities in the data, the statistical test of ANOVA identifies that the average cancer percent incidence rate does not make a significant difference between ethnicities. According to the CCR, the European ethnicity group has the highest overall cancer incidence rate of around 850 per 100,000 people while South Asian has the lowest overall cancer incidence rate at around 500 per 100,000 people. Five of the eight ethnicities (Non-Indigenous North American, East Asian, African, Caribbean, Western Central Asian and Middle Eastern) have similar overall incidence rates ranging from 600-650 per 100,000 people, and Latin, Central and South American and South Asian range between 500-550 per 100,000 people. With this information, there are certain implications for our research question, **"does cancer incidence differ among ethnicities in Canada?".** According to ANOVA statistical inference test, there is not much difference in the average incidence rate while the data obtained from CCR shows that the overall Incidence rate of the European ethnicity differs significantly from the rest of the ethnicities. This is somewhat expected as the ANOVA statistical inference test used data on the three most common cancer types among all sexes; lung, colorectal, and Non-Hodgkin Lymphoma cancer. Since all other cancer types were not included in the dataset, the ANOVA statistical inference test identified no mean percent cancer incidence rate difference between all the ethnicities. Perhaps if data on the other cancer types were added, the statistical difference of mean percent cancer incidence rate between European ethnicity and the other ethnicity would be much more apparent. Since sex-specific cancer types such as Breast cancer for Woman and Prostrate Cancer for men were omitted from statistical tests, this significantly changes the mean percent cancer incidence rate for all ethnicities. Perhaps if these sex-specific cancer types were included in the dataset, differences in the ethnicities may be more apparent through the ANOVA statistical inference test. For future tests, 2-sample t-test or Mann-Whitney test may be employed to determine if there is any significant difference in cancer incidence between male and female. Depending on whether the condition of normality is fulfilled through the shapiro test, these statistical tests can be performed and sex-specific cancer types can be readily used in the dataset.

REFERENCES:

1. Hwee J, Evelyne B. 2021. Do Cancer incidence and mortality rates differ among ethnicities in Canada? Canada: Statistics Canada; [accessed February 25th, 2022]. https://www150.statcan.gc.ca/n1/en/pub/82-003-x/2021008/article/00001-eng.pdf?st=UcAgNXX9
2. [NIH] National Cancer Institute. 2021. What is Cancer? Understanding Cancer.[accessed 2022 April 15] https://www.cancer.gov/about-cancer/understanding/what-is-cancer#:~:text=Cancer%20is%20a%20disease%20in,up%20of%20trillions%20of%20cells