Genetic Correspondence and Comparative Spatial Analysis with Age and Gender Specification of ICMR Data on Cancer Incidence in India

Sarbojit Das (231080075)[*], Swapnonil Mondal (231080098)[*], Sameer Verma (220949)[*], Sujash Krishna Basak (231080093)[*], Sayanta Biswas (231080080)[*]

[*]Department of Mathematics & Statistics, Indian Institute of Technology Kanpur, India

April 6, 2024

Abstract

Cancer remains a significant global health challenge, characterized by immense suffering and often limited treatment options. We aim to analyze genetic data related to five common cancer types, each presenting unique challenges and implications for diagnosis and treatment. Our goal is to identify the probable genetic causes underlying these cancers, providing leads for early identification and reducing fatality rates. Focusing on India, we explore the spatial and genetic aspects of cancer, utilizing statistical concepts and analyzing data from specific years. We aim to shed light on cancer prevalence, distribution and genetic underpinnings within the Indian population. Additionally, we investigate gender-specific and age-specific cancer incidence to offer detailed insights into the cancer landscape in India.

1 Introduction

Cancer is a formidable adversary to human health. It persists as a significant global burden, causing immeasurable suffering and presenting substantial challenges to healthcare systems worldwide. Its insidious nature lies in the uncontrolled growth and dissemination of abnormal cells, disrupting the delicate balance of cellular regulation inherent in the human body's natural processes. Despite remarkable advancements in medical science, the complexities of cancer remain a formidable challenge, often accompanied by limited treatment options and devastating outcomes for affected individuals and their families.

This project endeavors to delve into the multifaceted realm of cancer. With a diverse population and a unique set of demographic, environmental and genetic factors, India provides a rich tapestry for exploring the spatial, genetic, gender-specific and age-specific dimensions of cancer prevalence and distribution. Furthermore, our study extends beyond genetic analyses to explore the spatial aspects of cancer incidence in India. Leveraging statistical concepts and data from specific years, we seek to illuminate patterns of cancer prevalence and distribution across different regions of the country. This spatial perspective is crucial for understanding the geographic disparities in cancer burden and for informing targeted interventions and resource allocation efforts. Such insights hold the potential to inform tailored prevention, screening and treatment strategies, thereby improving outcomes and mitigating the impact of cancer on individuals and society at large.

Through this comprehensive exploration of cancer incidence in India, our project aims to contribute to a deeper understanding of the disease landscape. By integrating these aspects, we endeavor to shed light on the complexities of cancer and to ultimately strive towards a future where the burden of cancer is significantly reduced.

2 Dataset

In this project, we deal with multiple datasets that are based on cancer patients and the nature of cancer in both male and female human bodies.

1. Cancer causing genes dataset: The input dataset contains 802 samples corresponding to 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples are categorized into one of the following types of tumors: BRCA, KIRC, COAD, LUAD and PRAD.

Following is an extended description of the tumors associated with this dataset:

- BRCA (Breast Cancer): Breast cancer is one of the most common cancers among women. It originates in the breast tissue and can occur in both men and women.
- KIRC (Renal Cancer): Renal cell carcinoma, or kidney cancer, occurs in the lining of small tubes in the kidney.
- COAD (Colon Cancer): Colorectal cancer includes colon cancer (affecting the large intestine) and rectal cancer (affecting the rectum). Colon cancer is a major cause of cancer-related deaths.
- LUAD (Lung Cancer - *Adenocarcinoma*): Lung *adenocarcinoma* is a subtype of non-small cell lung cancer. It originates in the cells lining the airways and is one of the most common types of lung cancer.
- PRAD (Prostate Cancer): Prostate cancer occurs in the prostate, a small gland in men that produces seminal fluid. It is one of the most common cancers in men.

2. Estimated State-wise cancer incidences: The table provides estimated incidence of cancer cases in India by States and Union Territory-wise for all sites and for both sexes.

3. Gender-wise different sites of cancer: The table presents the estimated cancer incidence, number of cases, crude rate, and cumulative risk by sex and anatomical sites in India for the year 2022.

Following is an extended description of the terminologies associated with this dataset:

- Cum-risk: Cumulative risk of developing cancer in the age range of 0 to 74 years. It represents the likelihood or probability of an individual developing cancer within this specified age range.

$$\boxed{\text{Cumulative Risk} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100}$$

- Crude Rate (CR): Crude rate refers to the total number of cancer cases occurring in a population divided by the total population, expressed as a rate per a specific unit of time (usually per $100,000$ population). It provides a general overview of cancer incidence within a population, without considering factors such as age distribution.

$$\boxed{\text{Crude Rate} = \frac{\text{Number of Cancer Cases}}{\text{Total Population}} \times 100,000}$$

- Age-Adjusted Rate (AAR): Age adjusted rate is a standardized rate that takes into account the age distribution of a population. It allows for fair comparisons of cancer rates between different populations or over time by removing the influence of age as a confounding factor.

$$\boxed{\text{Age-Adjusted Rate} \quad = \sum_{i=1}^{n} \left(\frac{\text{Age-specific rate}_i \times \text{Population}_i}{\text{Total Population}} \right) \times 100,000}$$

- Malig Imn.Prol D (Malignant Immunoproliferative Diseases): This term refers to a group of disorders characterized by the abnormal proliferation of immune cells, leading to the development of malignancies. These diseases involve the uncontrolled growth of cells of the immune system, such as lymphocytes or plasma cells, and can include conditions like lymphomas, leukemias and multiple myeloma.

4. Age-wise cancer incidences: The table provides gender-disaggregated, estimated top five leading sites of cancer (%) in India by age group ($0-14, 15-39, 40-64$ & 65^+ age groups) for the year 2022.

3 Data Pre-processing and Cleaning

We conducted data pre-processing and cleaning tasks on two datasets: "`data.csv`" and "`labels.csv`". These datasets contain information related to cancer research, focusing specially on genetic data and associated labels.

1. Essential Packages: For this project, we utilized several essential packages in `R`, including `tibble`, `tidyverse` and `dplyr`.

2. Pre-processing: This step is very crucial pertaining to the complexity and volume of genetic data involved.

- Data Importing: Our code reads the CSV files "`data.csv`" and "`labels.csv`" into `R` as tibbles using the `read.csv()` function. These datasets contain genetic data and label information related to cancer research.

- Exploratory Data Analysis (EDA): We performed EDA to gain insights into the structure and content of the datasets. This involved examining the structure of both datasets using the `str()` function and displaying the first few entries of each dataset using the `head()` function.

- Categorical Variable Conversion: We converted some columns in the datasets from *character* type to *factor* type using the `as.factor()` function. This was necessary for categorical variables. We also converted the final dataframe into a more structured format, *tibble* to facilitate efficient analysis.

- Data Merging: We merged the two datasets into a single dataframe using the `cbind()` function, combining the label information with the genetic data.

- Data Quality Checks: We checked for null values (`is.null()`), NA values (`is.na()`), and duplicated entries (`duplicated()`) in the merged dataframe. The absence of such issues indicates clean data.

- Summary Statistics: We calculated summary statistics for the genetic data to understand its distribution and characteristics using the `summary()` function.

Table 1: Summary statistics for Genes 1 and 2

gene_1	gene_2
Min. :0.000	Min. :0.000
1st Qu.:2.299	1st Qu.:2.390
Median :3.144	Median :3.127
Mean :3.011	Mean :3.095
3rd Qu.:3.883	3rd Qu.:3.803
Max. :6.237	Max. :6.063

The processed dataframe contains both label information and genetic data. It is saved in the ICMR.Rdata file for further analysis and modelling.

4 Data Insights

The data we have, has 20,534 columns with 800 rows, as mentioned in the description. And distribution of classes is given below:

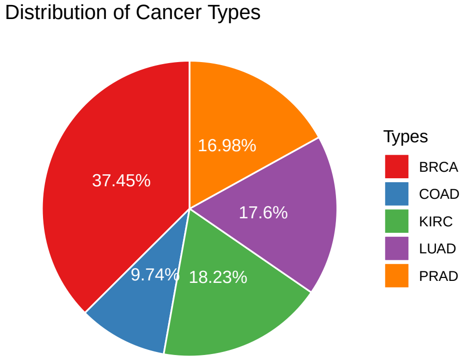


Figure 1: Proportion of Classes

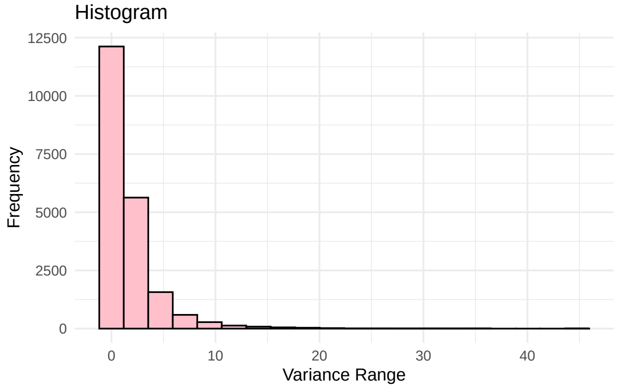


Figure 2: Histogram of Variance Range

From the pie chart, we can see that instances for class BRCA are more than that of others followed by KIRC, LUAD and PRAD and class COAD has the least number of instances.

- Variance Threshold: From the graph, it's easy to see that most of the columns have low variance. Features with high variance have data points that are spread out over a wide range. These features are considered significant because they contain valuable information that can help in discriminating between different classes or categories. Similarly, features with low variance have data points that have little variation. Such features may not carry much discriminative power and can be considered less significant.

- Distribution of average gene expression levels: We calculated the average expression values for each gene across all samples, sorting them in descending order based on these averages. The top 5 genes with the highest average expression values are displayed in a tabular format.

Table 2: Genes with highest mean expression value

Gene	Expression
gene_230	16.43044
gene_5380	16.38196
gene_232	15.96799
gene_18570	15.77775
gene_6857	15.71459

Following this, we generated a histogram and density plot to illustrate the distribution of average gene expression values.

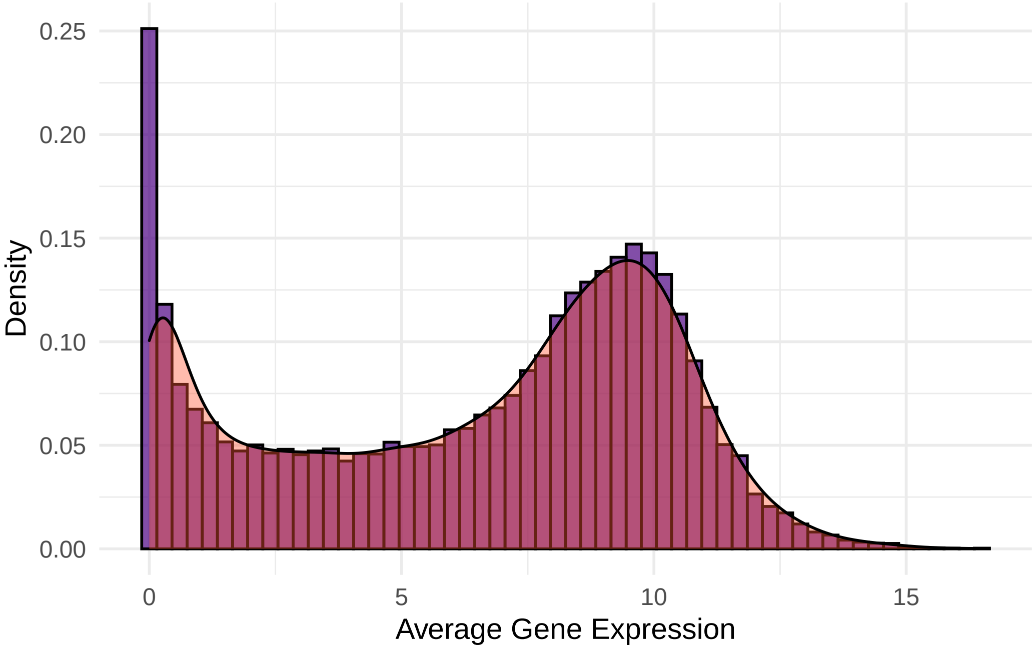


Figure 3: Distribution of mean gene expression levels

Boxplot of cases by gender:

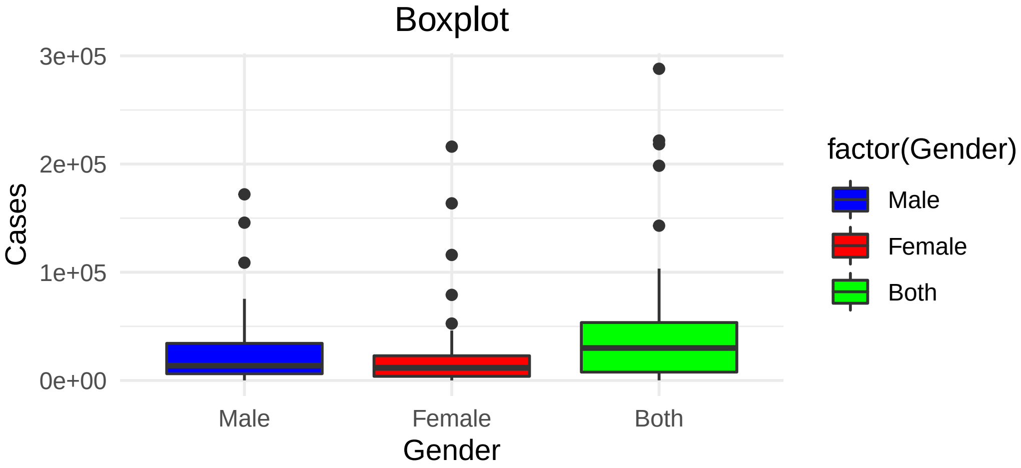


Figure 4: Cases by Gender

5 Visualizations

5.1 Age specific gender-wise cancer incidences

- The density curve for females exhibits a leptokurtic shape, indicating a higher concentration of data points around the mean and heavier tails compared to a normal distribution. In contrast, the density curve for males demonstrates a mesokurtic shape, suggesting a distribution with moderate kurtosis and a balanced spread of data around the mean.
- Our examination of the age-wise distribution of cancer patients across genders revealed a common trend characterized by a negative skewness. This observation implies a tendency towards a longer tail on the left side of the distribution, indicating a higher frequency of cancer occurrences in older age groups compared to younger age groups, irrespective of gender.

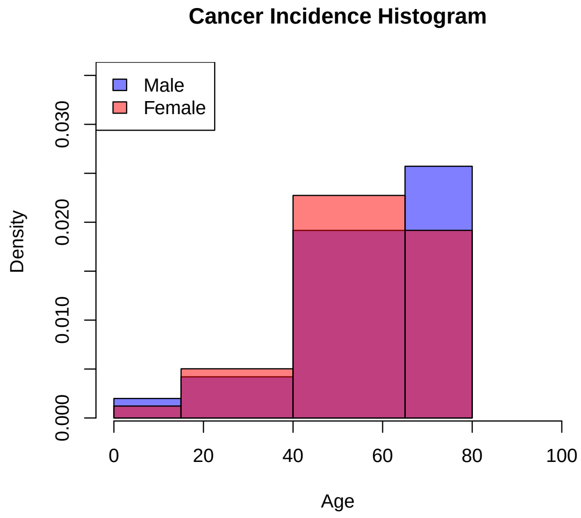


Figure 5: Stacked Histogram

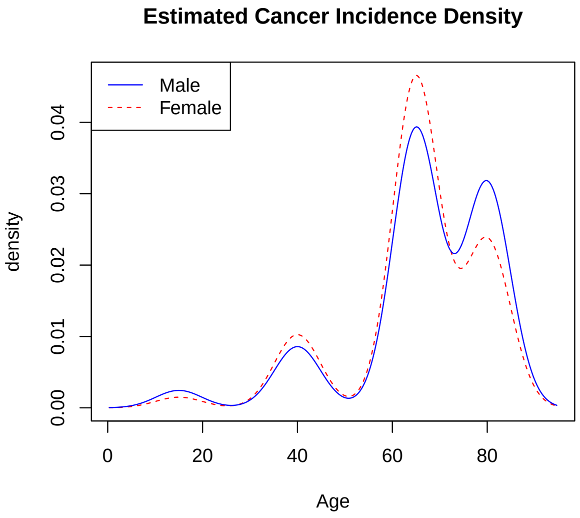


Figure 6: Density Plots

5.2 Gender-wise top ten cancer sites

Leading Cancer Sites Among Females:

- Breast (17.89%): Major risk factors include female gender, age, family history, genetic mutations (e.g., BRCA1 and BRCA2), hormonal factors (e.g., early menarche, late menopause), reproductive history and lifestyle factors.
- Genital System (13.55%): This includes cancers of the cervix, uterus, ovaries, vagina and vulva. Risk factors vary by site but may include HPV infection, sexual activity, smoking, hormonal factors and genetic predisposition.
- Digestive System (9.60%): This includes dietary habits, alcohol consumption, tobacco use and chronic conditions such as obesity and *gastroesophageal reflux disease* (GERD).

- Uterine Cervix (6.55%): HPV infection is the primary risk factor, along with smoking, early sexual activity, multiple sexual partners and immuno-suppression.

- Oral Cavity and Pharynx (4.35%): Risk factors include including tobacco and alcohol use, HPV infection and poor oral hygiene.

- Ovary (3.82%): Risk factors include age, family history, nulliparity, infertility, hormonal factors and possibly endometriosis.

- Respiratory System (2.83%) - Lung and Bronchus (2.31%): Similar risk factors to those in males, primarily smoking and occupational exposures.

- Uterine Corpus (2.31%): Risk factors include hormonal factors (e.g., estrogen exposure), obesity, diabetes and certain genetic syndromes.

- Endocrine System (2.31%): This includes cancers of the thyroid, adrenal glands and other endocrine organs. Risk factors vary by site but may include radiation exposure, family history and certain genetic syndromes.

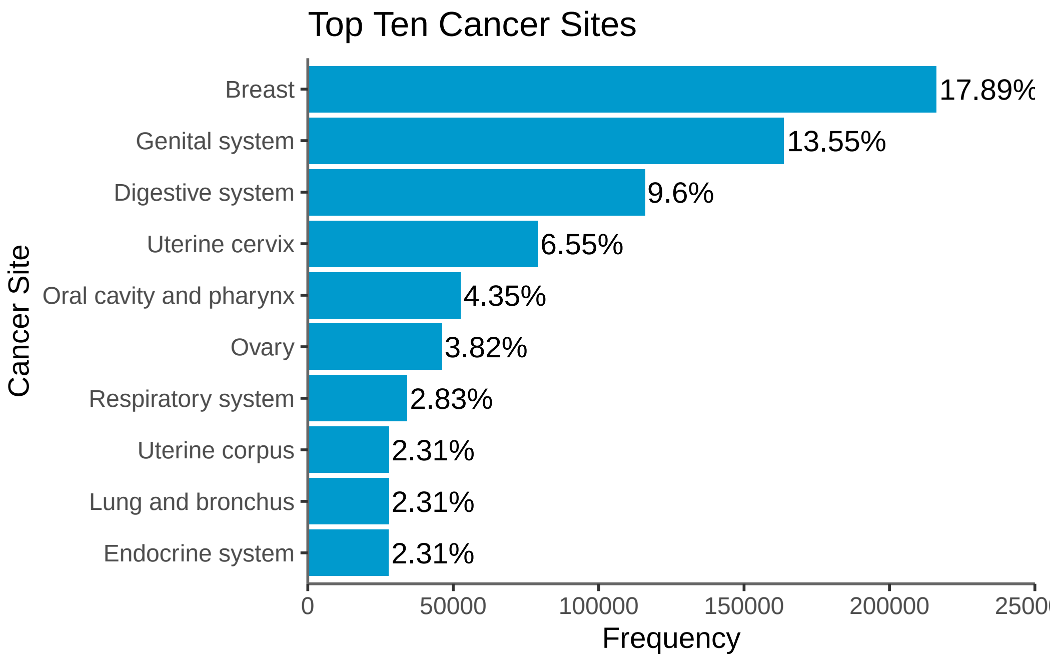


Figure 7: Top Ten Cancer Sites of Female

Leading Cancer Sites Among Males:

- Digestive System (13.06%): This includes cancers of the esophagus, stomach, liver, pancreas and colorectal region. Risk factors may include tobacco and alcohol use, dietary habits (e.g., high intake of processed meats) and chronic conditions such as *gastroesophageal reflux disease* (GERD) and *inflammatory bowel disease* (IBD).

- Oral Cavity and Pharynx (11.07%): Major risk factors include tobacco use (both smoking and smokeless tobacco), heavy alcohol consumption, *human papillomavirus* (HPV) infection and poor oral hygiene.

- Respiratory System (8.26%) - Lung and Bronchus (5.73%): Smoking, including exposure to secondhand smoke, is the primary risk factor for lung cancer. Occupational exposures to carcinogens such as asbestos, radon and diesel exhaust can also contribute.

- Mouth (4.57%): Similar risk factors to oral cavity and pharynx cancers, including tobacco and alcohol use, HPV infection and poor oral hygiene.

- Genital System (4.15%) - Prostate (3.32%): Prostate cancer is influenced by age, family history, and possibly dietary factors. Genetic predisposition and hormonal factors, particularly testosterone, play significant roles.

- Tongue (3.18%) - Other Oral Cavity (3.09%): Risk factors are similar to those for oral cavity and pharynx cancers, including tobacco and alcohol use, HPV infection and poor oral hygiene.

- Urinary System (2.65%): This includes cancers of the bladder, kidney and other urinary organs. Risk factors include smoking, occupational exposures, certain medications and genetic factors.

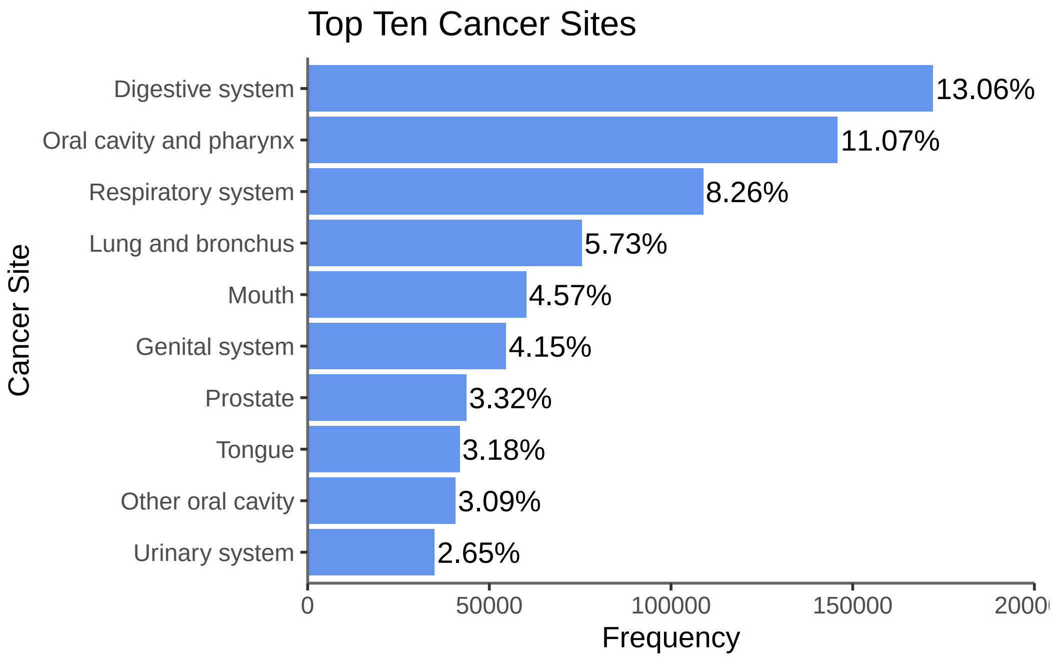


Figure 8: Top Ten Cancer Sites of Male

6 Spatial Analysis

- Northern and eastern parts of India show lower cancer incidence compared to regions like Uttar Pradesh, Maharashtra, Bihar, West Bengal, Hyderabad and Tamil Nadu. Hilly areas and Andaman Nicobar islands exhibit lower cancer rates compared to the plains. Western and southeastern parts of India demonstrate moderate cancer incidence.
- Uttar Pradesh stands out as having the highest cancer incidence among Indian states. Maharashtra, Bihar, West Bengal, Hyderabad and Tamil Nadu also exhibit high numbers of cancer cases.
- Factors such as pollution, industrial activities and exposure to carcinogens may contribute to higher cancer rates in certain regions.
- Clean and less polluted environments in hilly areas may contribute to lower cancer rates. Differences in lifestyle patterns, including dietary habits and physical activity levels, may also play a role in the observed regional disparities.

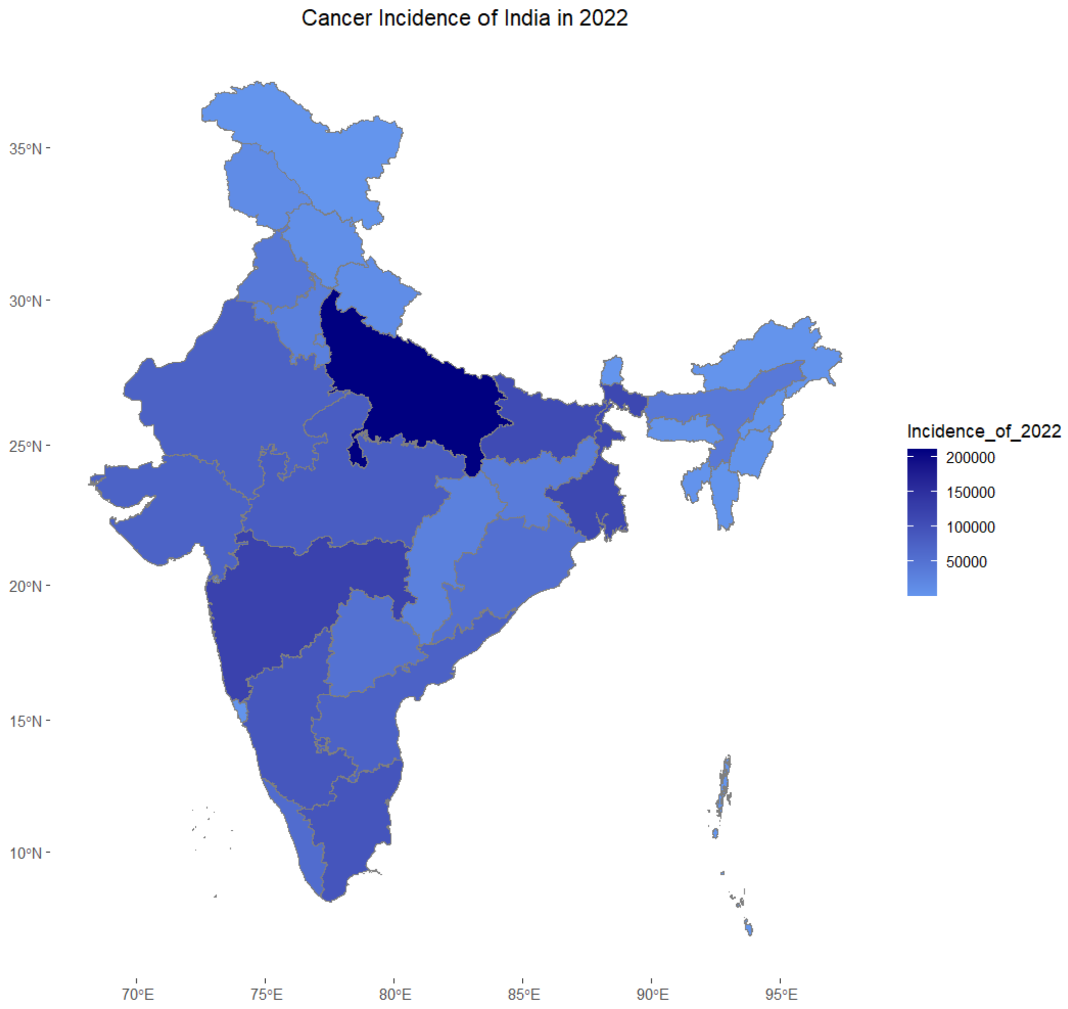


Figure 9: Cancer Incidence of India in the year 2022

We performed a comparative 10 year gap analysis with the cancer incidences in years, 2011 and 2021 respectively.

- We observed that Karnataka had improved over in cancer eradication whereas Hyderabad plunged to an increment in the number of cancer cases. The Telangana state, formed in 2 June, 2014, had less amount of cancer cases in 2021. Ladakh was formed as a Union Territory in 31 October, 2019. It showed few cancer patients in 2021, possibly due to its smaller population ratio.
- West Bengal displayed very slight changes in the cancer incidence ratio (i.e., $\frac{\text{Total no. of cancer patients in that state}}{\text{Total no. of cancer patients in India}}$) over the 10-year period.
- In a quick snapshot, there is a massive increment in the total number of cancer patients nationwide from approximately $160,000$ in 2011 to around $200,000$ in 2021. This substantial increase underscores the importance of heightened attention from both the public and government sectors towards cancer eradication efforts and mass awareness campaigns.

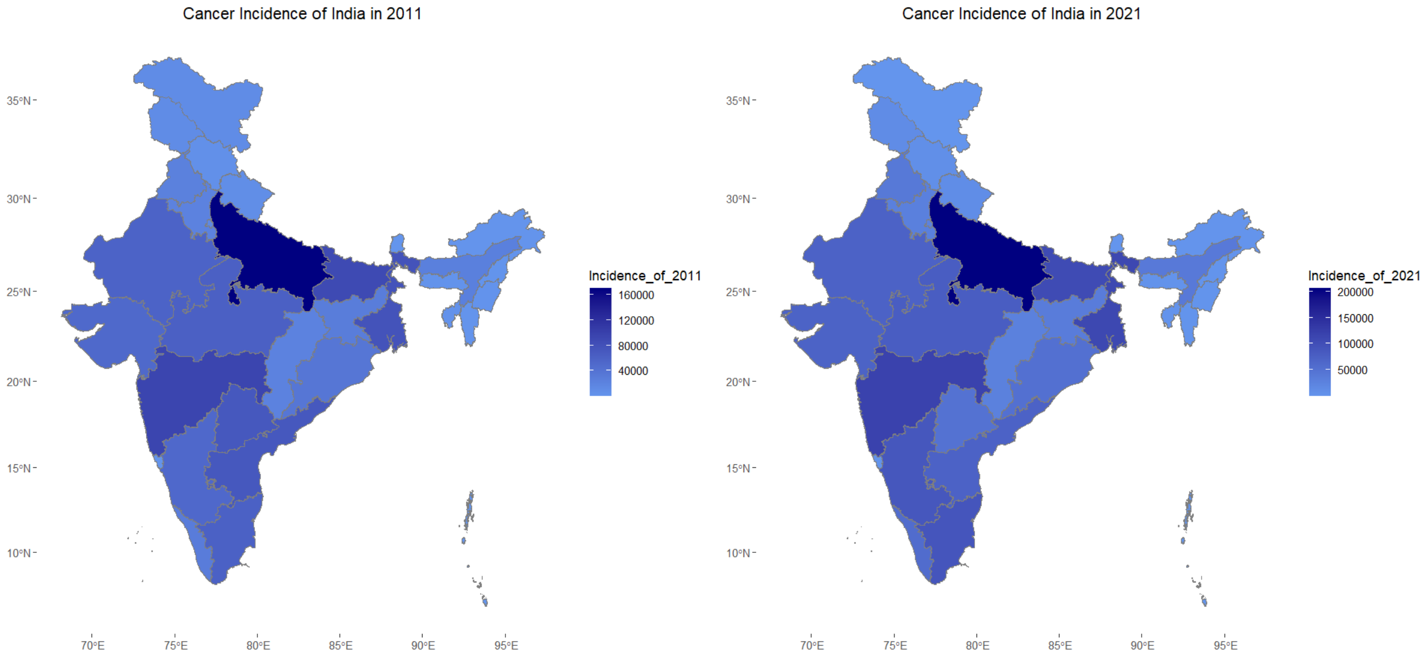
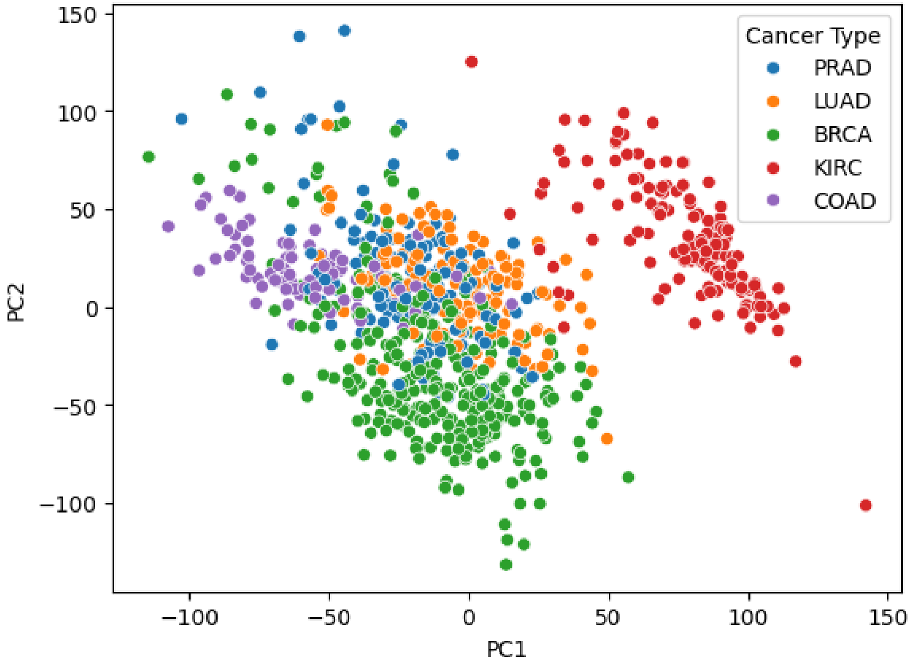


Figure 10: A comparative 10 Year Gap Cancer Incidence in India

7 Principal Component Analysis

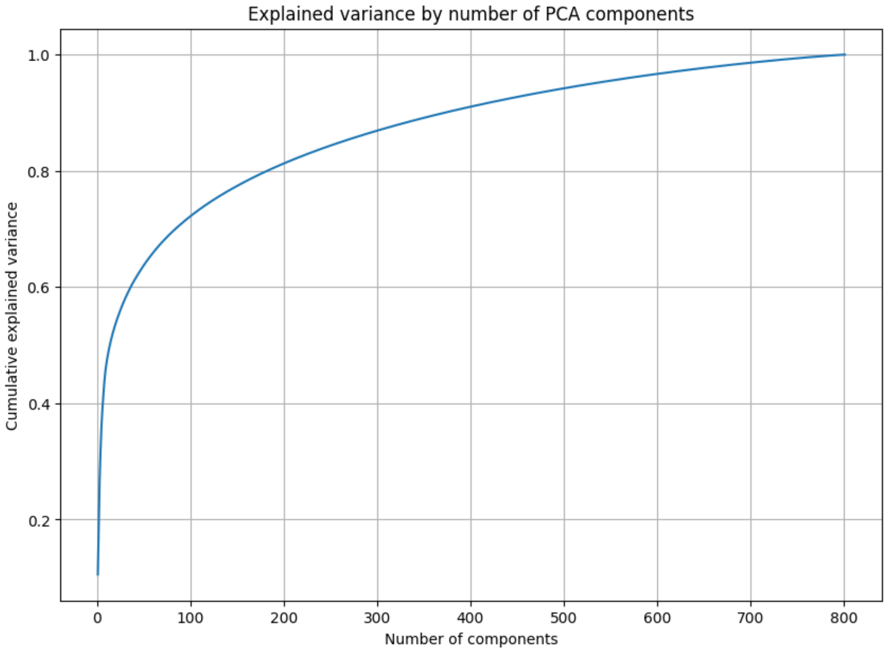
Having high dimensionality or feature space can lead model to perform poorly known as the "*Curse of Dimensionality*". To deal with this, we reduce the feature space by performing the dimension reduction technique, Principal Component Analysis (PCA) and we consider 95% variance explainability.

Each sample has expression values for around 20K genes. However, it may not be necessary to include all 20K gene expression values to analyze each cancer type. Therefore, we need to identify a smaller set of attributes which would then be used to fit multiclass classification models. So, the first task targets dimensionality reduction using PCA.



- PCA effectively reduces the high-dimensional data to 2-dimensions and provides insights into the distribution of different cancer types. The use of color to represent different cancer types allows for easy differentiation and understanding of the data.
- This plot shows that there are distinct clusters formed by different cancer types, indicating that the gene expression patterns vary across different cancer types.
- Although some genes overlap between the clusters, the plot does not show any clear separation between all cancer types.

Cumulative Explained Variance Plot:



- This plot typically shows the cumulative sum of explained variances explained on the Y-axis and the number of principal components on the X-axis.

- As more principal components are added, the cumulative proportion of variance increases.

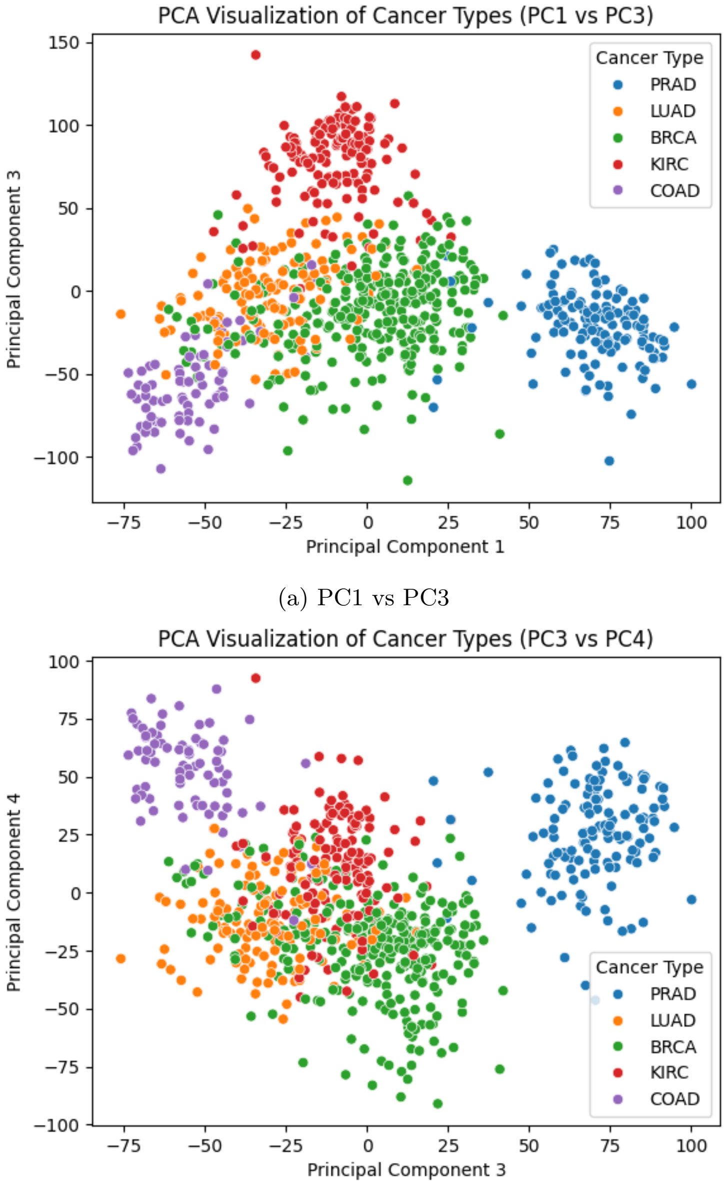
- It helps in identifying the point at which adding more principal components does not significantly increase the explained variance, which can be a good guide for determining the appropriate number of principal components to retain.

- The '*elbow*' or point where the curve starts to level off is often used to decide how many components one wants to keep.

- From this curve, we see that it starts to flatten around 150 - 200 PCs, suggesting a potential elbow point in this range.

Remarks:

- Class PRAD is notably more separable than other classes, indicating effective feature differentiation.
- PC1 and PC3 hold valuable information for distinguishing KIRC and PRAD from other cancer types.
- Classes PRAD and COAD exhibit enhanced separability, emphasizing the significance of PC3 and PC4.
- PC3 and PC4 capture distinctive features for characterizing PRAD, making it more distinguishable.



(a) PC1 vs PC3

(b) PC3 vs PC4

Figure 11: Scatter Plot of Data Distribution Based on Principal Components

8 Parametric Tests

8.1 Multiple F-tests

We want to determine whether means of gene information encoded by a particular gene (*response*) differ statistically significantly among the independent cancer groups (*covariates*). In this case, with reference to the `ICMR` dataset, the independent categorical variable is the '`Class`' column which represents the 5 different cancer types. The dependent variable is the gene expression levels of a specific gene. We analyze each gene individually in this setup, making it multiple F-tests.

We have,

- 5 groups (categories of cancer types)
- n_i observations in the i^{th} group (where $i = 1, 2, \ldots, 5$)
- N total observations ($N = n_1 + n_2 + \ldots + n_5$)
- X_{ij} is the observation in the i^{th} group and the j^{th} gene expression level ($j = 1, 2, \ldots, 20532$)
- $\overline{X_n} \overset{asym}{\sim} \mathcal{N}(\mu,\ \frac{\sigma^2}{n})$ i.e., $\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,\ 1)$ as $\boxed{n \to \infty}$ (by Central Limit Theorem) where $\overline{X_n}$ is a random variable with mean μ and variance $\frac{\sigma^2}{n}$ that denotes mean gene expression for a particular gene.

The null hypothesis (H_0) for each F-test is that there are no differences in mean gene expression levels among the different cancer types.

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_5$$

where: μ_i is the population mean of the i^{th} group.

The alternative hypothesis (H_1) is that at least one pair of mean gene expression levels is different.

$$\mathrm{H}_1 : \text{At least one } \mu_i \text{ is different from the others.}$$

To test these hypotheses, we calculate the F-statistic: $\mathrm{F} = \frac{\text{Between-group variance}}{\text{Within-group variance}}$

Rejection Criteria: If the p-value associated with the F-statistic is below a certain threshold (typically 0.05), we reject the null hypothesis.

Table 3: Top 3 high variance genes

Gene	F_statistic	p_value	variance
gene_9176	3463.550	0	44.76385
gene_9175	4194.489	0	36.36194
gene_15898	1905.191	0	34.50391

Table 4: Bottom 3 high variance genes

Gene	F_statistic	p_value	variance
gene_12668	3.163906	0.0137332	0.0014541
gene_12670	3.250183	0.0118692	0.0013394
gene_4834	2.446886	0.0450807	0.0005937

8.2 One vs. All t-test

In a one-vs-all t-test scenario, we compare the means of one cancer group (the "*one*" group) against the means of all other groups excluding that cancer group (the "*all*" group). It's commonly used in situations where one wants to compare a specific group against the rest of the data.

Table 5: Assumptions

Independence	Gene expressions within the "*one*" group and "*all*" group are independent of each other.
Normality	Each group's data follow approximately the normal distribution (by Central Limit Theorem).
Random Sampling	The collective gene expressions data are obtained through a random sampling process from the population of cancer patients. This ensures that the sample is representative of the population.
Unequal Variances	The compared variances of the two groups are unequal.

- Null Hypothesis (H_0): The null hypothesis assumes that there is no difference between the means of the "*one*" group and the means of the "*all*" group.

$$H_0 : \mu_{\text{one}} = \mu_{\text{all}}$$

- Alternative Hypothesis (H_1): The alternative hypothesis states that the means of the "*one*" group and the "*all*" group are different.

$$H_1 : \mu_{\text{one}} \neq \mu_{\text{all}}$$

- Test Statistic: $t_{\text{cal}} \overset{H_0}{=} \dfrac{\bar{x_1} - \bar{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ where:
  - $\bar{x_1}$ and $\bar{x_2}$ are the means of the "*one*" group and the "*all*" group respectively.
  - s_1 and s_2 are the standard deviations of the "*one*" group and the "*all*" group respectively.
  - n_1 and n_2 are the sample sizes of the "*one*" group and the "*all*" group respectively.
- Rejection Criteria: If the p-value is smaller than a predetermined significance level (commonly 0.05), we reject the null hypothesis. Else, we fail to reject H_0 and further data analysis is required.

- Results: We performed this test for the top 3 high variance genes.

```
Performing 'one vs. all' t-tests for gene_9176 :
PRAD vs. All for gene_9176 : p-value = 0
LUAD vs. All for gene_9176 : p-value = 2.3192579e-50
BRCA vs. All for gene_9176 : p-value = 3.9070211e-28
KIRC vs. All for gene_9176 : p-value = 2.5060458e-42
COAD vs. All for gene_9176 : p-value = 0.00025240039
Performing 'one vs. all' t-tests for gene_9175 :
PRAD vs. All for gene_9175 : p-value = 2.0867204e-256
LUAD vs. All for gene_9175 : p-value = 4.2859671e-35
BRCA vs. All for gene_9175 : p-value = 8.2966265e-23
KIRC vs. All for gene_9175 : p-value = 4.8127605e-50
COAD vs. All for gene_9175 : p-value = 8.2634542e-27
Performing 'one vs. all' t-tests for gene_15898 :
PRAD vs. All for gene_15898 : p-value = 6.2853312e-37
LUAD vs. All for gene_15898 : p-value = 2.4837977e-99
BRCA vs. All for gene_15898 : p-value = 3.5424646e-33
KIRC vs. All for gene_15898 : p-value = 8.7197042e-28
COAD vs. All for gene_15898 : p-value = 7.6822615e-16
```

9 Multinomial Logistic Regression

- Goal: We want to fit Logistic Regression for predicting the probability of the occurrence of 5 different types of cancer for an individual.
- Motivation: Cancer types across different sites are of penultimate importance, particularly in studying the genetic effects on these cancer types. One common question that arises is that to which extent do genetics influences these cancers. Our dataset primarily focuses on 5 types of cancer: BRCA, COAD, KIRC, LUAD and PRAD. Through Principle Component Analysis (PCA), we have determined that the first 500 genes capture the maximum variation. Now, we aim to investigate whether these genes exhibit any significant effects on these cancer types.

So, here we consider the different types of cancer as distinct labels, each associated with corresponding gene expression values. Our interest lies in fitting a Multinomial Logistic Regression Model. Let us examine the following table, which defines the probabilities of occurrence for each cancer type.

Table 6: Cancer Types with Corresponding Probabilities for j^{th} individual

Sl. No.	1	2	3	4	5	
Cancer Types	BRCA	COAD	KIRC	LUAD	PRAD	Total
Probability	π_{1j}	π_{2j}	π_{3j}	π_{4j}	π_{5j}	1

Now, from this we observe that $\sum_{i=1}^{5} \pi_{ij} = 1$ so, $\pi_{5j} = 1 - \sum_{i=1}^{4} \pi_{ij}$, considering 5^{th} category as the Pivotal Category. Thus, it is good enough to estimate π_i ; $i = 1, 2, 3, 4$ and we can automatically get back the estimated value of π_5 which in turn will also reduce our manual labour.

Model: Let Y_j be the random variable of the occurrence of a type of cancer of j^{th} person,

$$Y_j = \begin{cases} 1 & \text{if BRCA occurred} \\ 2 & \text{if COAD occurred} \\ 3 & \text{if KIRC occurred} \\ 4 & \text{if LUAD occurred} \\ 5 & \text{if PRAD occurred} \end{cases}$$

which means

$$Y_j = \begin{cases} 1 & \text{with probability } \pi_{1j} \\ 2 & \text{with probability } \pi_{2j} \\ 3 & \text{with probability } \pi_{3j} \\ 4 & \text{with probability } \pi_{4j} \\ 5 & \text{with probability } \pi_{5j} = 1 - \sum_{i=1}^{4} \pi_{ij} \end{cases}$$

Therefore, $P(Y_j = i) = \pi_{ij} = \frac{e^{X_j^T \beta_i}}{1+e^{X_j^T \beta_i}}$; $i = 1, 2, 3, 4$

Now, by Maximum Likelihood Estimation we try to maximize the likelihood function,

$$L(\beta) = L(\beta_1, \beta_2, \beta_3, \beta_4) = \prod_{j=1}^{n} \left[\left\{ \prod_{i=1}^{4} \left(\frac{e^{X_j^T \beta_i}}{1 + e^{X_j^T \beta_i}} \right)^{\mathbb{I}(Y_j = i)} \right\} \left(\frac{1}{1 + e^{X_j^T \beta_i}} \right)^{\mathbb{I}(Y_j = 5)} \right]$$

in order to fit the model. We can maximize it by various numerical ways.

Let us consider only Gene -1 first.

Table 7: Estimates and p-values of the regression coefficients

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	2.345370	0.343819	6.822	9.01×10^{-12}
(Intercept):2	-0.547328	0.494891	-1.106	0.269
(Intercept):3	2.398823	0.365845	6.557	5.49×10^{-11}
(Intercept):4	0.296070	0.411232	0.720	0.472
$gene_1 : 1$	-0.492740	0.099693	-4.943	7.71×10^{-07}
$gene_1 : 2$	-0.002505	0.137841	-0.018	0.985
$gene_1 : 3$	-0.787602	0.112672	-6.990	2.75×10^{-12}
$gene_1 : 4$	-0.076462	0.115619	-0.661	0.508

Here from the p-values, we can see that Gene -1 has significant effects in causing BRCA and KIRC, at level 0.05.

Now, let us fit the model and obtain the predicted probabilities of different types of cancer.

Let us consider two individuals with gene expression values of Gene -1 as 3.4678533 and 2.9411814. Then the fitted probabilites are as follows:

Table 8: Fitted Probabilities of the Specified Two Individiuals

Individual	BRCA	COAD	KIRC	LUAD	PRAD
1	0.3626322	0.11002989	0.1375935	0.1978839	0.1918605
2	0.3962039	0.09286059	0.1755878	0.1736390	0.1617087

From this, we can see that for the first individual, according to the gene expression value of Gene -1, there is greater probability of occurring BRCA. For the second person, the scenario is more or less the same.

Again, let us check it for Gene -1 and Gene -2 collectively.

Now, let us fit the model and obtain the predicted probabilities of different types of cancer.

Let us consider two individuals with gene expression values of Gene -1 and Gene -2 as follows: 3.4678533, 2.9411814 and 2.9411814, 2.6632763 respectively. Then the fitted probabilites are:

Table 9: Fitted Probabilities of the Specified Two Individiual

Individual	BRCA	COAD	KIRC	LUAD	PRAD
1	0.3461035	0.1183316	0.1068651	0.2312489	0.19745091
2	0.4627836	0.1008106	0.2074741	0.1551230	0.07380868

From this, we can see that for the first individual, according to the gene expression value of Gene -1, there is greater probability of occurring BRCA. For the second person, the scenario is more or less the same.

10 Non-Parametric Tests

10.1 Shapiro-Wilk Normality Test

In our analysis, we have used this test in two different cases. At first, we had to check whether the gene expression values in the dataset, ICMR.Rdata are normally distributed or not. Secondly, we used it to check whether the ages of male and female cancer patients are normally distributed. This is performed using the `R` function called `shapiro.test()`. For further details see **Appendix**.

Testing for Gene Expression Values

Here expressions of $20,532$ different genes are recorded. Among them only $1,767$ gene expressions can be considered to be normally distributed at 5% level of significance. Each of the $20,532$ genes are tested for normality individually. Each gene has 800 different expression values.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_n} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} expression value for the i^{th} gene. Here $i = 1, 2, 3, ..., 20532$ and $j = 1, 2, ...800$.
- The distribution functions F_i are assumed to be absolutely continuous $\forall i$.
- The test is a level $\alpha = 0.05$ both sided test.

The i^{th} null hypothesis H_{i0} is that the i^{th} gene expression comes from Φ_i which is the normal distribution function. The corresponding alternative hypotheses are opposite.

$$H_{i0} : F_i = \Phi_i \text{ against } H_{ia} : F_i \neq \Phi_i \ \forall i$$

The i^{th} Wilk-Shapiro test statistic is,

$$W_i = \frac{(\sum_{j=1}^{800} a_{i_j} X_{i_{(j)}})^2}{(\sum_{j=1}^{800} (X_{i_j} - \bar{X}_i)^2)} \text{ where } a_{i_j} = (a_{i_1}, a_{i_2}, ...a_{i_n}) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

- $X_{i_{(j)}}$ and $\bar{X}_i$ are the j^{th} order statistics and the i^{th} sample mean respectively.
- $m_i = (m_{i_1}, m_{i_2}, ...m_{i_{800}})^{1/2}$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution.
- V_i is the covariance matrix of those order statistics of the i^{th} sample.
- Rejection Criteria: If the p-value is smaller than a pre-determined significance level α, we reject H_0.

Findings: After running Shapiro-Wilk Test on each of the 20,532 genes' expression values, we assembled all the informations like name of the genes, W-Statistic values, p-values and accept-reject decisions into a single dataframe named shapiro.csv. The output is displayed below through tables.

Table 10: 5 Random draws of rows

Gene	W	p-value	Rejected
gene_12090	0.9534434	9.950693e-15	TRUE
gene_11126	0.9977158	3.538067e-01	FALSE
gene_1929	0.9921550	3.997301e-04	TRUE
gene_14104	0.9828147	8.043263e-08	TRUE
gene_11549	0.9887918	1.271188e-05	TRUE

The table below shows the genes which have the highest p-values. That means these genes' expression values are very likely to be normally distributed.

Table 11: Top 4 normally distributed gene expressions

Gene	W	p-value	Rejected
gene_18776	0.99939029519	0.9977	FALSE
gene_13337	0.99929787407	0.9937	FALSE
gene_745	0.99921778732	0.9872	FALSE
gene_4471	0.99921284749	0.9867	FALSE

Testing for Normality of Cancer Patients' Age

Here age data of $1,461,427$ cancer patients are recorded. Among them $749,251$ patients are female and $712,176$ patients are male. We have tested whether this data is normally distributed.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_{n_i}} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} person of the i^{th} gender. Here $i \in \{f, m\}$ and $j = 1, 2, ...n_i$ where $n_m = 712176$, $n_f = 749251$
- The distribution functions F_i are assumed to be absolutely continuous $\forall i$.
- The test is a level $\alpha = 0.05$ both sided test.

The i^{th} null hypothesis H_{i0} is that the i^{th} age sample comes from Φ_i which is the normal distribution function. The corresponding alternative hypotheses are opposite.

$$H_{i0} : F_i = \Phi_i \text{ against } H_{ia} : F_i \neq \Phi_i \ \forall i$$

The i^{th} Wilk-Shapiro test statistic is,

$$W_i = \frac{(\sum_{j=1}^{n_i} a_{i_j} X_{i_{(j)}})^2}{(\sum_{j=1}^{n_i} (X_{i_j} - \bar{X}_i)^2)} \text{ where } a_{i_j} = (a_{i_1}, a_{i_2}, ...a_{i_{n_i}}) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

- $X_{i_{(j)}}$ and $\bar{X}_i$ are the j^{th} order statistics and the i^{th} sample mean respectively.
- $m_i = (m_{i_1}, m_{i_2}, ...m_{i_{n_i}})^{1/2}$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution.
- V_i is the covariance matrix of those order statistics of the i^{th} sample.
- Rejection Criteria: If the p-value is less than $\alpha = 0.5$, we reject H_0.

Findings: Here it turns out that the age distribution of female cancer patients are very less likely to have normal distribution. As the p-value of the test is $p_f < 0.05$. So, H_{f0} is rejected at 5% level of significance.

```
    Shapiro-Wilk normality test

data:  f
W = 0.95473, p-value = 1.992e-11
```

Similarly, as $p_m < 0.05$ the null hypothesis H_{m0} is rejected at 5% level of significance. Therefore, it is safe to say age of male cancer patients are not normally distributed.

```
    Shapiro-Wilk normality test

data:  m
W = 0.95473, p-value = 1.992e-11
```

10.2 Kolmogorov-Smirnov Test

10.2.1 One Sample Test

Previously, we performed Shapiro-Wilk test to check the normality of the age of cancer patients. WS test works better for checking whether the data is normal or not. But, to check about the other distribution we are using very popular Kolmogorov-Smirnov test. For theory, see **Appendix**.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_{n_i}} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} person of the i^{th} gender. Here $i \in \{f, m\}$ and $j = 1, 2, ...n_i$ where $n_m = 712176$, $n_f = 749251$.
- The distribution functions F_i are assumed to be absolutely continuous $\forall i$.
- The test is a level $\alpha = 0.05$ both sided test.

The i^{th} null hypothesis H_{i0} is that the i^{th} age sample comes from Γ_i which is the gamma distribution function. The corresponding alternative hypotheses are opposite.

$$H_{i0} : F_i = \Gamma_i \text{ against } H_{ia} : F_i \neq \Gamma_i \ \forall i$$

The Kolmogorov-Smirnov test statistics are,

$$D_{n_m} = \max_x |S_{n_m}(x) - \Gamma_m(x)| \quad \text{and} \quad D_{n_f} = \max_x |S_{n_f}(x) - \Gamma_f(x)|$$

- $S_{n_m}(x)$ and $S_{n_f}(f)$ are respectively the empirical distribution function of male and female age sample.
- Rejection Criteria: If the p-value is less than $\alpha = 0.5$, we reject H_0.

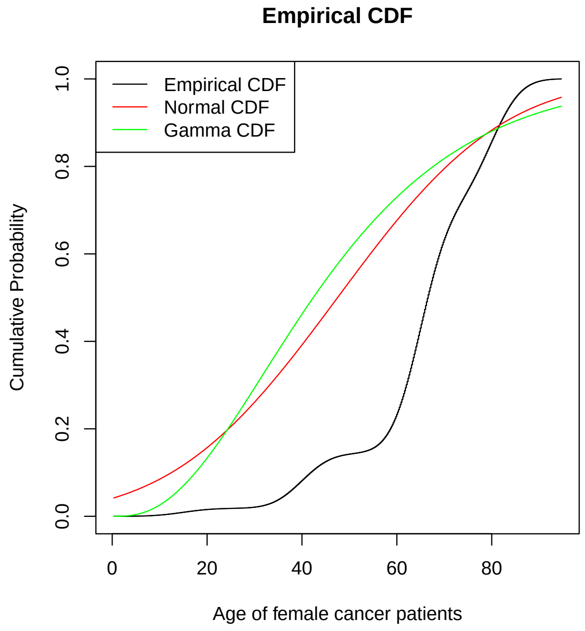
Findings: Here it turns out that the age distribution of female cancer patients are very less likely to have gamma distribution. As the p-value of the test is $p_f < 0.05$. So, H_{f0} is rejected at 5% level of significance.

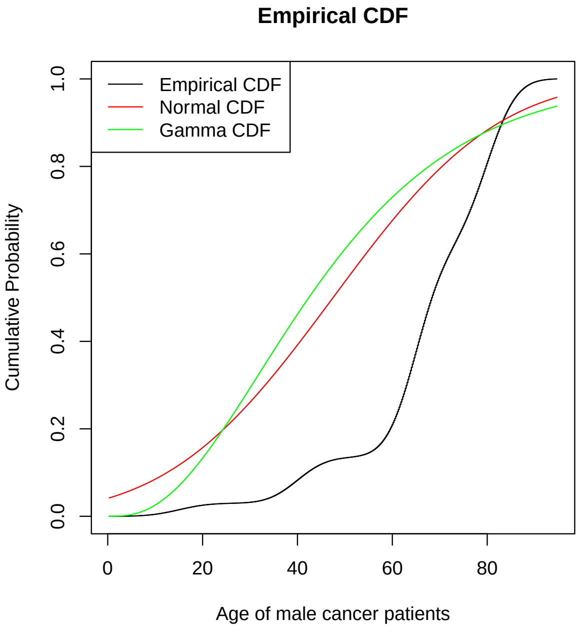
```
    One-sample Kolmogorov-Smirnov test

data:  f
D = 0.99482, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Similarly, as $p_m < 0.05$ the null hypothesis H_{m0} is rejected at 5% level of significance. Therefore, it is safe to say age of male cancer patients are not gamma distributed.

```
    One-sample Kolmogorov-Smirnov test

data:  m
D = 0.99482, p-value < 2.2e-16
alternative hypothesis: two-sided
```

We have summarised our findings from the above in the following plots given below:





10.2.2 Two Sample Test

Now we will test whether the age of male cancer patients and female cancer patients come from the same population or not by two sample KS Test. For theory, see **Appendix**.

We have,

- $X_{i_1}, X_{i_2}, ..., X_{i_{n_i}} \overset{iid}{\sim} F_i(x)$ where X_{i_j} is the j^{th} person of the i^{th} gender. Here $i \in \{f, m\}$ and $j = 1, 2, ...n_i$ where $n_m = 712176$, $n_f = 749251$.
- The distribution functions F_m and F_f are assumed to be absolutely continuous.
- The test is a level $\alpha = 0.05$ both sided test.

The null hypothesis H_0 is that the two distribution functions F_m and F_f are same. While, the alternative hypothesis is they are not equal.

$$H_0 : F_f = F_m \quad \text{against} \quad H_a : F_f \neq F_m$$

The Kolmogorov-Smirnov test statistic can be defined as,

$$D_{m,f} = \max_x |S_{n_f}(x) - S_{n_m}(x)|$$

- $S_{n_m}(x)$ and $S_{n_f}(f)$ are respectively the empirical distribution function of male and female age sample.
- Rejection Criteria: If the p-value is less than $\alpha = 0.5$, we reject H_0.

Findings: Here it is evident that the p-value is very high. At 5% (or even higher) level of significance H_0 cannot be rejected as $p > 0.05$. It is safe to conclude that the gender-wise age of cancer patients have same distribution.

```
    Two-sample Kolmogorov-Smirnov test

data:  m and f
D = 0, p-value = 1
alternative hypothesis: two-sided
```

- In the analyzed age group, our findings reveal a notable trend in cancer prevalence, with breast cancer and cervix cancer exhibiting higher incidence rates among females compared to males. This observation suggests that these types of cancer disproportionately affect females within this age range .
- Furthermore, our analysis indicates a notable increase in the number of patients diagnosed with these cancers, underscoring their dominant presence in the realm of cancer analysis.

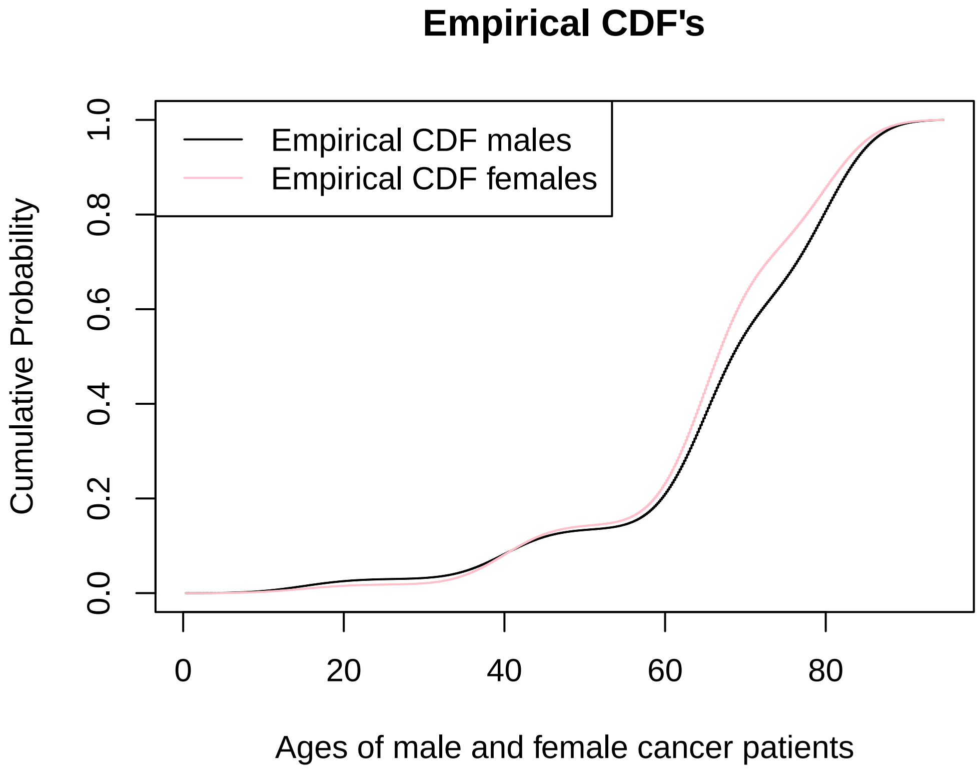


Figure 12: Plot of Empirical CDF's

11 Conclusion

In brief, our analysis of cancer incidence in India reveals significant findings. Uttar Pradesh emerged as a high-risk region and the noticeable increase in cancer cases across India in recent times needs further investigation and intervention. Interestingly, our examination of the male-female age-wise distribution of cancer patients indicated a similarity with a negative skewness, suggesting a common trend in cancer occurrence across genders and age groups. Through exploratory data analysis (EDA), we gained insights into the structure and content of the datasets, ensuring data quality through categorical variable conversion, data merging and data quality checks.

Furthermore, our exploration of different cancer types demonstrated non-uniform distribution, with varying percentages of occurrence across the different types. We observed that Breast Cancer (BRCA) contained the most data points, almost 38% of the entire dataset.

To delve deeper into the factors influencing cancer types, we employed the multinomial logistic regression model, inputing gene expression values of various genes —particularly those capturing maximum variability identified through Principal Component Analysis (PCA). We performed parametric tests to assess differences in mean gene expression levels among cancer types. Rejection of the null hypothesis for certain genes in these tests signified the likeliness of significant differences for which further data analysis is required. Additionally, the one-vs-all t-test allowed us to compare specific cancer groups against the rest, contributing to a better understanding of inter-grouped differences.

12 Acknowledgement

We express our sincere gratitude to Prof. Subhajit Dutta for his continuous supervision and invaluable support throughout the completion of this project. His guidance, insights and encouragement have been fundamental in shaping our trajectory. We had the privilege of applying the knowledge and skills gained under his course, `MTH209: Data Science Lab - II` to tackle a real-world problem. This opportunity allowed us to delve deep into the subject matter and hone our abilities through practical applications.

We also extend our heartfelt thanks to our Teaching Assistants, Mr. Arghya Mukherjee and Ms. Annesha Deb for their assistance in guiding us through the process of selecting the project topic and steering us in the right direction. Amidst the initial confusion about choosing a topic, their wisdom helped us align our mutual interests to meet our goals. We are grateful for the unwavering support and encouragement from both of them which played a monumental role in the successful completion of this project.

13 Appendix

13.1 Testing of Significance of the Regression coefficients in the Multinomial Logistic Regression

Model: Here our model is $E(Y_j) = \frac{e^{X_j^T \beta_i}}{1+e^{X_j^T \beta_i}}$, $j = 1, 2, ..., n$

Assumptions: The error term ϵ_j is such that $E(\epsilon_j) = 0$ and $Var(\epsilon_j) = \sigma_j^2$ where $Var(Y_j) = \sigma_j^2$ $j = 1, 2, ..., n$

Hypothesis : $H_{0i} : \beta_i = 0$ vs $H_{1i} : \beta_i \neq 0$, $i = 1, 2, ..., k$

Test statistic: Here our test statistic is

$$T_i = \sqrt{k}\frac{\hat{\beta}_i}{\text{standard error}} \underset{H_0}{\overset{a}{\sim}} N(0, 1)$$

Test Rule: We reject H_{0i} at level α iff $p - value = P(|T_i| \geq observed(T_i)) \leq \alpha$; $i = 1, 2, ..., k$

13.2 Shapiro-Wilk Normality Test

Shapiro and Wilk (1965) test was originally restricted for sample size of less than 50. This test was the first test that was able to detect departures from normality due to either skewness and kurtosis, or both. It has become the preferred test because of its good power properties.

- Model: $X_1, X_2, ..., X_n \overset{iid}{\sim} F(x)$
- Assumption: F is absolutely continuous.
- Hypothesis: The null hypothesis H_0 assumes the random samples comes from a normal population with distribution function Φ, while the alternative says that it does not come from Φ.

$$H_0 : F = \Phi \text{ against } H_a : F \neq \Phi$$

- Test Statistics: The required test statistic $W \in [0, 1]$

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{(\sum_{i=1}^{n} (X_i - \bar{X})^2)} \text{ where } a_i = (a_1, a_2, ...a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

  - $X_{(i)}$ and $\bar{X}$ are the i^{th} order statistics and the sample mean respectively.
  - $m = (m_1, m_2, ...m_n)^{1/2}$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution.
  - V is the covariance matrix of those order statistics.

- Rejection Criteria: If the p-value is smaller than a pre-determined significance level α, we reject H_0.

13.3 Empirical Distribution Function (EDF)

For a random sample from the distribution F_X, the empirical distribution function or edf, denoted by $S_n(x)$, is simply the proportion of sample values less than or equal to the specified value x, that is,

$$S_n(x) = \frac{\text{number of sample values} \leq x}{n}$$

Suppose that the n sample observations are distinct and arranged in increasing order so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest, . . . , and $X_{(n)}$ is the largest. Then, a formal definition of the edf $S_n(x)$ is

$$S_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \frac{i}{n} & \text{if } X_{(i)} \leq x < X_{(i+1)} \text{ for } \; i = 1, 2, \dots, n \\ 0 & \text{if } x \geq X_{(n)} \end{cases}$$

13.4 The Kolmogorov-Smirnov One Sample Test

Model : $X_1, X_2, ..., X_n \overset{iid}{\sim} F(x)$
Assumption : Here we assume the F is absolutely continuous distribution function.
Hypothesis : Here we are interested in testing whether the random sample come from a known distribution with distribution function $F_0(x)$ (say). That is,

$$H_0 : F = F_0 \text{ against } H_a : F \neq F_0$$

where $F = F_0 \equiv F(x) = F_0(x) \; \forall \; x$
and $F = F_0 \equiv F(x) \neq F_0(x)$ with strict inequality for atleast one x
Here we are mainly interested in checking Normality. So, $F_0(x) = \Phi(x)$
So, the hypothesis becomes,

$$H_0 : F = \Phi \text{ against } H_a : F \neq \Phi$$

where Φ is the cumulative distribution function of standard normal distribution.
Test Statistic : For testing H_0 against H_a we define the following statistic,

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F_n(x)} - \Phi(x) \right|$$

Notice that $\widehat{F_n(x)}$ approximate the true distribution function Φ . $\widehat{F_n(x)}$ by definition is a step function i.e. the absolute difference measured by D_n provide us the departure of the true situation from the null hypothesis towards the corresponding alternative.
The distribution of the test statistics is free from the population distribution,
Distribution Free :
Exact Distribution Free (EDF) of D_n : Here we use the ordered statistics,

$$X_{(1)}, X_{(2)}, ..., X_{(n)}, \text{ Further we denote } X_{(0)} = -\infty \text{ and } X_{(n+1)} = \infty$$

$$\begin{aligned}
D_n &= \sup_{x\in\mathbb{R}} \left|\widehat{F_n(x)} - \Phi(x)\right| \\
&= \max_{i=0,1,2,...,n} \left\{ \sup_{X_{(i)}\leq x\leq X_{(i+1)}} \left|\widehat{F_n(x)} - \Phi(x)\right| \right\} \\
&= \max\left\{ o, \max_{i=1,2,...,n} \left\{ \sup_{X_{(i)}\leq x\leq X_{(i+1)}} \left|\widehat{F_n(x)} - \Phi(x)\right| \right\} \right\} \\
&= \max\left\{ o, \max_{i=1,2,...,n} \left\{ \sup_{X_{(i)}\leq x\leq X_{(i+1)}} \left|\frac{i}{n} - \Phi(x)\right| \right\} \right\} \qquad \left[\because \widehat{F_n(x)} = \frac{i}{n}\right] \\
&= \max\left\{ o, \max_{i=1,2,...,n} \left\{ \left|\frac{i}{n} - \Phi(x_{(i)})\right| \right\} \right\}
\end{aligned}$$

Notice that,

$$X_1, X_2, ..., X_n \underset{H_0}{\overset{iid}{\sim}} \Phi$$

$$\implies U_i = \Phi(X_i) \underset{H_0}{\overset{iid}{\sim}} U(0,1)$$

Since Φ is absolutely continuous, we can conclude that,

$$\implies U_{(i)} = \Phi\left(X_{(i)}\right) \underset{H_0}{\overset{iid}{\sim}} U(0,1)$$

That is D_n (under H_0) depends on the ordered statistics $\left(U_{(1)}, U_{(2)}, ..., U_{(n)}\right)$ from U(0,1). Hence the test is based on D_n is EDF (Exact Distribution Free).

Test : Notice that D_n depends on the empirical distribution function $\widehat{F_n(x)}$ which represents the true distribution function. Thus the directional difference measured by D_n actually indicate the departure of the true situation form the null hypothesis towards the alternative H_a i.e. under $H_a : F \neq F_0$ becomes larger than that under H_0. On the other hand a small value of D_n indicates the acceptance of H_0. Thus a right tail test based on D_n will be appropriate for testing $H_0 : F = \Phi$ against $H_a : F \neq \Phi$.

We reject $H_0 : F = \Phi$ against $H_a : F \neq \Phi$ at level α if,

$$D_n > d_\alpha$$

In terms of p-value we can say, we reject $H_0 : F = \Phi$ against $H_a : F \neq \Phi$ at level α if,

$$p - value = P_{H_0}\left(D_n \geq observed(D_n)\right) \leq \alpha$$

13.5 The Kolmogorov-Smirnov Two Sample Test

Two sample KS test is a non-parametric test which is used to compare two different samples are coming from the same population or not.

The order statistics corresponding to two random samples of size m and n from continuous populations F_X and F_Y, are

$$X_{(1)}, X_{(2)}, ..., X_{(m)} \text{ and } \; Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$$

Their respective empirical distribution functions, denoted by $S_m(x)$ and $S_n(x)$, are defined as:

$$S_m(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \frac{k}{m} & \text{if } X_{(k)} \leq x < X_{(k+1)} \text{ for } \; k = 1, 2, \ldots, m-1 \\ 0 & \text{if } x \geq X_{(m)} \end{cases}$$

and

$$S_n(x) = \begin{cases} 0 & \text{if } x < Y_{(1)} \\ \frac{k}{n} & \text{if } Y_{(k)} \leq x < Y_{(k+1)} \text{ for } \; k = 1, 2, \ldots, n-1 \\ 0 & \text{if } x \geq Y_{(n)} \end{cases}$$

In a combined ordered arrangement of $m + n$ sample observations, $S_m(x)$ and $S_n(x)$ are the respective proportions of X and Y observations which do not exceed the specified value x.

If the null hypothesis

$$H_0 : F_Y(x) = F_X(x) \;\; \forall x$$

is true, the population distributions are identical and we have two samples from the same population. The empirical distribution functions for the X and Y samples are reasonable estimates of their respective population cdf. Therefore, allowing for sampling variation, there should be reasonable agreement between the two empirical distributions if indeed H_0 is true; otherwise the data suggest that H_0 is not true and therefore should be rejected. In other words, how close do the two empirical cdf's have to be so that they could be viewed as not significantly different, taking account of the sampling variability. This approach necessarily requires a definition of closeness. The two-sided Kolmogorov-Smirnov two-sample test criterion, denoted by $D_{m,n}$, is based on the maximum absolute difference between the two empirical distributions.

$$D_{m,n} = \max_x |S_m(x) - S_n(x)|$$

Since here only the magnitudes, and not the directions, of the deviations are considered, $D_{m,n}$ is appropriate for a general two-sided alternative

$$H_A : F_Y(x) \neq F_X(x) \text{ for some } x.$$

and the rejection region is in the upper tail, defined by

$$D_{m,n} \geq c_\alpha$$

where

$$P(D_{m,n} \geq c_\alpha \,|\, H_0) \leq \alpha$$

Because of the Glivenko-Cantelli theorem, the test is consistent for this alternative. The p-value is

$$p = P(D_{m,n} \geq D_{observed} \,|\, H_0)$$

14 References

- Fundamentals of Statistics (*Volume One*); A.M. Gun, M.K. Gupta, B. Dasgupta; The World Press Pvt. Ltd., 2019.

- The Cancer Breakthrough: A Nutritional Handbook for Doctors and Patients (*Paperback*) – 29 July, 2007.

- Razali, N. and Wah, Y. (2011) Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2, 21-33.

- Wilk, W. (2015) Shapiro Wilk And Related Tests For Normality, Massachusetts Insitute of Technology.

- Gibbons, J.D. and Chakraborti, S. (2014), Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded.