Gender-wise and location-wise Diversity in tobacco related cancers in Delhi

Supervised by

Dr. Vivek Verma

Assistant Professor

Department of Statistics

Assam University, Silchar

Presented by

Nirman Nath

Msc. 4th Sem

Department of Statistics

Assam University , Silchar

Introduction

Cancer

One of the leading cause of disease worldwide(10 million in 2020 ,WHO reports).

Tobacco-related cancer (TRC) accounts for and among smokers expected to reach 2 billion worldwide by 2030.[National Library of medicine]

Tobacco use remains the leading preventable cause of death (1 in 5 deaths)due to cancer in the US (American Cancer Society)

About 50% of men and 9% of women are smokers in developing countries, compared with 35% of men and 22% of women in high-resource countries.(GlOBOCAN ,2012)

Government made policies to control tobacco smoking and awareness campaigns.

Literature review

Author	Year	Findings
Asthana S. et al.,	2021	Proportion of TRC in relation to all cancers was high in different registries of India including the Northeast region.
Verma V., *et al.*,	2020	Diversity estimate under Bayesian paradigm unveiled better results compared to classical in substance use behaviour in child in Delhi.
Joannie Lortet-Tieulent,et al.,	2015	Changing variation in the incidence rates of tobacco-related Cancers in genders by European region
Arnold M ,et al.,	2010	The study unveils the cancer risk diversity in non-western migrants in and between different European countries.

Objective of the study

To study the distribution of cancer due to tobacco across genders and organ sites for consecutive 3 years viz 2013,2014 and2015 in capital city of India

To study the statistical diversity of cancer cases between genders and organ sites by using Shannon Index under both Classical and Bayesian paradigms

` Methodology

Data:

- Source: Delhi Cancer Registry
- Maintained by : All India Institute of Medical Science(AIIMS, New Delhi).
- Years: 2013,2014 & 2015

Study Variable:

Number of males , number of females , total number of individuals , different cancer sites are the study variables taken into consideration.

Study Population

- In Delhi, number of cancer cases registered in 2015 was 21538. Out of these, males were 11435 (53.1%) and females 10103 (46.9%).(Delhi cancer registry,2015)
- In 2016, men from southern, northern and eastern regions had the highest crude incidence rates of lung cancer, while mouth and oesophageal cancers ranked first in the western, central and northeast regions of the country followed by lung cancer.[Kulothungan V., *et al*.,2022]
- Cancer cases in India projected to rise from 14.6 lakh in 2022 to 15.7 lakh in 2025[TOI]

Methods: Measures of diversity

- Shannon Index and Simpson Index: Measures of diversity or species richness
- Provide insights into the evenness and distribution of species within a given community.
- Shannon Index(known as Shannon-Weaver Index or Shannon Entropy): Quantifies the diversity of species in a community by considering both the number of species present and their relative abundances(proportions or percentages).
- In the Shannon index, p is the proportion (n/N) of individuals of one particular species found (n) divided by the total number of individuals found (N), Σ is the sum of the calculations, and s is the number of species.
- The three measures which can be checked are evenness , richness and diversity.
- Evenness represents the degree to which individuals are split among species with low values indicating that one or a few species dominate, and high values indicating that relatively equal numbers of individuals belong to each species.

- Richness is simply the number of species in the community while the species diversity is the number of different species combined with the relative abundances of the individuals within each of those species in community.

- For complex community level diversity analysis it is still unclear which diversity measure is most effective but the basic idea for a diversity measure is that how abundances or proportions of certain species can be analysed for dominance of one or more species. These dominance cannot be looked through in the species accurately without the help of diversity measure.

- Researchers and ecologists use species diversity as one of the metrics to assess the health and diversity of ecosystems. It helps in understanding the relationships between different species and their roles in maintaining the overall ecological balance.

- This study delves into the methodology, where a mathematical formulation of SDI (Shannon Diversity Index) under both Classical and Bayesian paradigm is obtained and the comparisons of the results.

Shannon Diversity Index Under Classical Paradigm

The Shannon Diversity Index can be denoted and defined as follows:

$$H(p) = -\sum_{i=1}^{s} p_i \log(p_i).$$

where p_i= proportion of the species.

s= no of species to be compared

After derivation, the estimated diversity index is as follows:

$$E(H(p)) = -\sum_{i=1}^{s} p_i\left[\log(np_i) + \frac{(1 - p_i)}{2n} - \log n\right] \quad (1)$$

$$V(H(p)) = \frac{1}{n}\sum_{i=1}^{s} p_i(1 - p_i)([1 + \log(n\, p_i)]^2 + (\log n)^2) \quad (2)$$

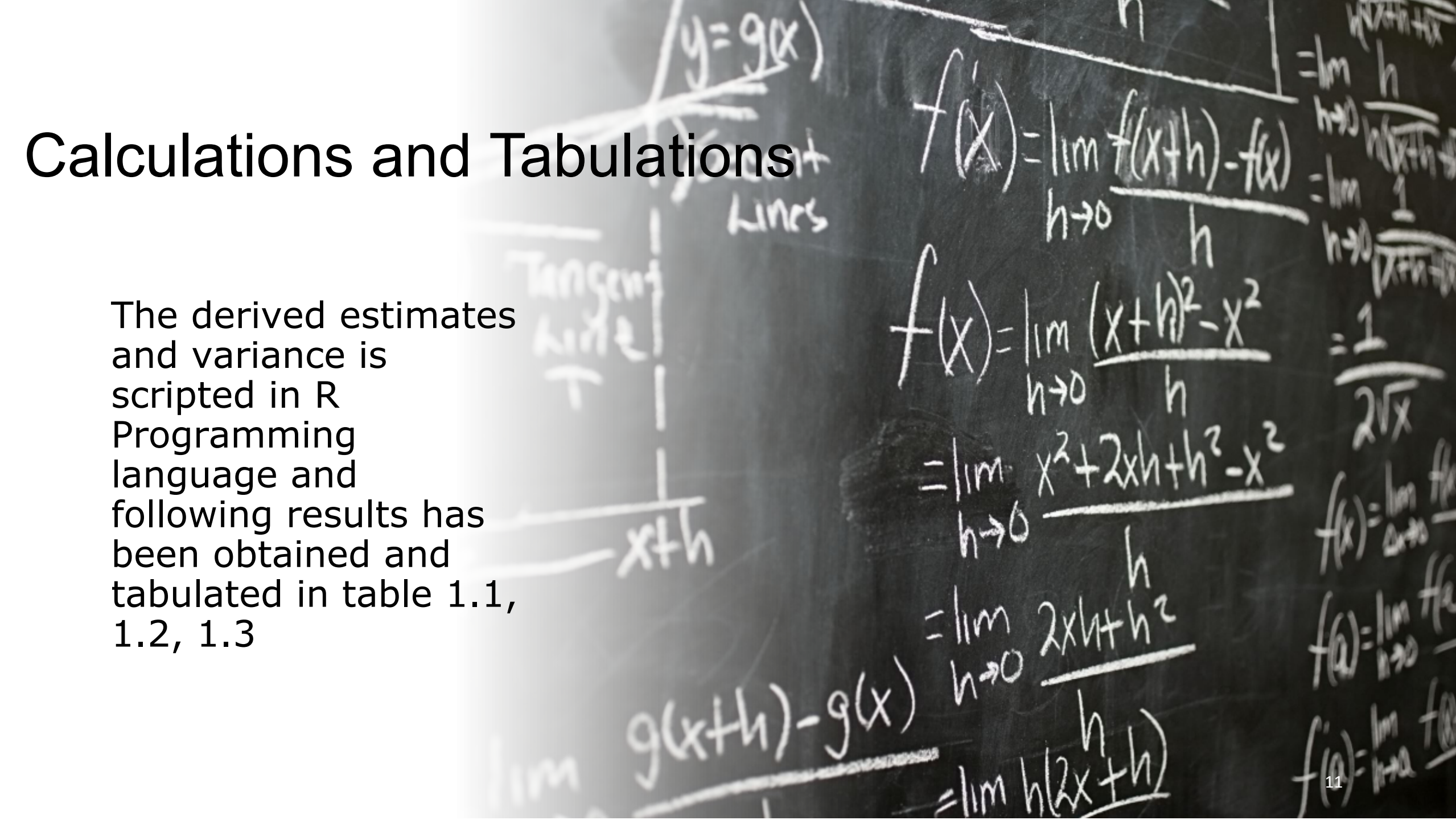
where p_i=proportion of ith species

s= total no of species

n=sum of the observations of s species

Calculations and Tabulations

The derived estimates and variance is scripted in R Programming language and following results has been obtained and tabulated in table 1.1, 1.2, 1.3



Table[1.1] for 2013

Site	Male	Female	Total	H(p)	V(H(p))[in 10^{-2}]
Lip	39	12	51	0.545	0.063
Tongue	719	256	975	0.575	1.557
Mouth	791	23	1014	0.526	1.457
Oropharynx	149	31	180	0.459	0.208
Hypopharynx	173	41	214	0.488	0.268
Pharynx	23	8	31	0.570	0.039
Oesophagus	495	258	753	0.642	1.444
Larynx	515	51	566	0.302	0.513
Lung	1073	332	1405	0.546	2.007
Urinary Bladder	398	81	479	0.454	0.590
H(p)	1.97	1.86	1.96		
V(H(p))	0.02	0.057	0.019		

Table[1.2] for 2014

Sites	Male	Female	Total	H(p)	
Lip	47	9	56	0.438	0.063
Tongue	628	180	808	0.530	1.363
Mouth	824	179	1003	0.468	1.462
Oropharynx	118	16	134	0.364	0.140
Hypopharynx	146	27	173	0.432	0.215
Pharynx	54	15	69	0.521	0.097
Oesophagus	406	243	649	0.660	1.550
Larynx	460	56	516	0.343	0.569
Lung	1122	337	1459	0.540	2.319
Urinary Bladder	449	100	549	0.474	0.816
H(p)	1.97	1.86	1.86		
V(H(p))	0.024	0.061	0.020		

Table[1.3] for 2015

Sites	Male	Female	Total	H(p)	
Lip	43	13	56	0.538	0.078
Tongue	718	194	912	0.517	1.436
Mouth	842	160	1002	0.439	1.317
Oropharynx	155	19	174	0.344	0.168
Hypopharynx	139	19	158	0.366	0.160
Pharynx	36	10	46	0.519	0.060
Oesophagus	396	249	645	0.666	1.534
Larynx	449	57	506	0.351	0.540
Lung	1207	340	1547	0.526	2.295
Urinary Bladder	506	110	616	0.468	0.872
H(p)	1.95	1.85	1.95		
V(H(p))	0.023	0.061	0.019		

SDI under Bayesian Paradigm

$X_1, X_2 \ldots\ldots, X_s$: *Cases per sites of cancer due to tobacco consumption and* are independently distributed as

$X_i | p_i \sim$ *Binomial (n, pi)*

Prior distribution is

$p_i \sim$ Beta (α , β) ;where α and β are constants

The posterior density is given by :

$q(pi|Xi) \sim Beta(Xi + \alpha , n - Xi + \beta)$; (i=1,2…..10)

pi= proportions of cancer cases at ith site

Posterior is of the form of beta binomial distribution and hence

Posterior mean of pi|xi

- Mean= $\frac{Xi + a}{a + \beta + n}$

Posterior variance is given by :

- Variance= $\frac{(Xi + a)(n - x + \beta)}{(a + \beta + n)^2 (a + \beta + n + 1)}$

Simulating $N = 1000$ values from the posterior distribution as

$\{ p_1^{(t)}, p_2^{(t)}, p_3^{(t)}, \ldots, p_s^{(t)} ; t = 1, 2, \ldots, N\}$,

then computed $H^{(1)}(p), H^{(2)}(p), \ldots, H^{(N)}(p)$

where

$$H^{(t)}(p) = -\sum_{i=1}^{s} p_i \log(p_i).$$

The tables[2.1,2.2,2.3] for the year 2013 gender-wise Bayesian has been obtained and shown here :

Table 2.1: Male

Sites	Classical	Bayesian	[α,β] (1,1)	(1/2,1/2)	(3,3)	(2,8)	(8,2)
Lip	0.008914		0.0091	0.0090	0.0095	0.0093	0.0107
Tongue	0.164343		0.1644	0.1644	0.1648	0.1644	0.1657
Mouth	0.1808		0.1809	0.1808	0.1812	0.1808	0.1822
Oropharynx	0.034057		0.0342	0.0341	0.0346	0.0344	0.0358
Hypopharynx	0.039543		0.0397	0.0396	0.0401	0.0399	0.0412
Pharynx	0.005257		0.0054	0.0053	0.0059	0.0057	0.0070
Oesophagus	0.113143		0.1133	0.1132	0.1136	0.1133	0.1147
Larynx	0.117714		0.1178	0.1178	0.1182	0.1179	0.1192
Lung	0.245257		0.2453	0.2453	0.2456	0.2451	0.2465
Urinary bladder	0.090971		0.0911	0.0910	0.0915	0.0912	0.0925
H	1.97	H(B)	1.9819	1.9809	1.9888	1.9843	2.008277
V(H)	0.024045	V(H(B)) [in 10^{-4}]	2.772	2.755	2.610	2.931	2.794

Table 2.2 : Female

Sites	Classical	Bayesian	[α,β] (1,1)	(1/2,1/2)	(3,3)	(2,8)	(8,2)
Lip	0.0092		0.0100	0.0096	0.0115	0.0107	0.0153
Tongue	0.1979		0.1984	0.1982	0.1993	0.1980	0.2026
Mouth	0.1724		0.1729	0.1727	0.1739	0.1726	0.1772
Oropharynx	0.0239		0.0247	0.0243	0.0261	0.0253	0.0299
Hypopharynx	0.0317		0.0324	0.0320	0.0338	0.0330	0.0376
Pharynx	0.0061		0.0069	0.0065	0.0084	0.0076	0.0122
Oesophagus	0.1995		0.2000	0.1997	0.2009	0.1995	0.2041
Larynx	0.0394		0.0401	0.0397	0.0415	0.0406	0.0452
Lung	0.2567		0.2571	0.2569	0.2578	0.2563	0.2609
Urinary bladder	0.0626		0.0633	0.0629	0.0646	0.0636	0.0683
H	1.86	H(B)	1.879541	1.872323	1.905428	1.887445	1.971939
V(H)	0.0574	V(H(B)) [in 10^{-4}]	8.4197	8.9546	8.7953	9.6811	8.7152

Table 2.3: Total

Sites	Classical	Bayesian	[α,β] (1,1)	(1/2,1/2)	(3,3)	(2,8)	(8,2)
Lip	0.0089		0.0091	0.0090	0.0095	0.0093	0.0103
Tongue	0.1720		0.1721	0.1720	0.1723	0.1720	0.1731
Mouth	0.1788		0.1790	0.1789	0.1792	0.1789	0.1799
Oropharynx	0.0317		0.0319	0.0318	0.0322	0.0320	0.0331
Hypopharynx	0.0377		0.0379	0.0378	0.0382	0.0380	0.0390
Pharynx	0.0054		0.0056	0.0055	0.0059	0.0058	0.0068
Oesophagus	0.1328		0.1329	0.1329	0.1332	0.1329	0.1340
Larynx	0.0998		0.1000	0.0999	0.1002	0.1000	0.1010
Lung	0.24788		0.2479	0.2479	0.2481	0.2477	0.2488
Urinary Bladder	0.0845		0.0846	0.0845	0.0849	0.0847	0.0857
H	1.96	H(B)	1.969679	1.967427	1.975167	1.971705	1.990469
V(H)	0.019752	V(H(B))	$2.2371*10^{-4}$	$2.3239*10^{-4}$	$2.3789*10^{-4}$	$2.0995*10^{-4}$	$2.2841*10^{-4}$

Similarly we have obtained the tables for 2014 and 2015

Also
The density plots of Bayesian proportions for the given constants[$\alpha=3,\beta=3$] for different sites of cancers and for genders male ,female and total are also obtained using R and are shown in next slides.

Figure 1:Genderwise cancer Distribution of patients having Lip as the affected organ site

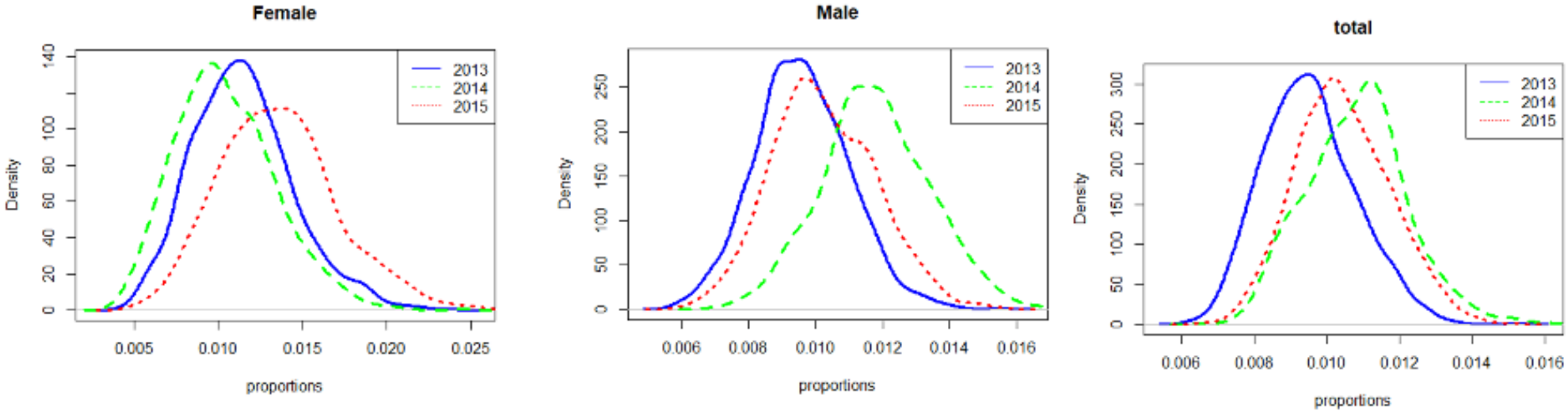
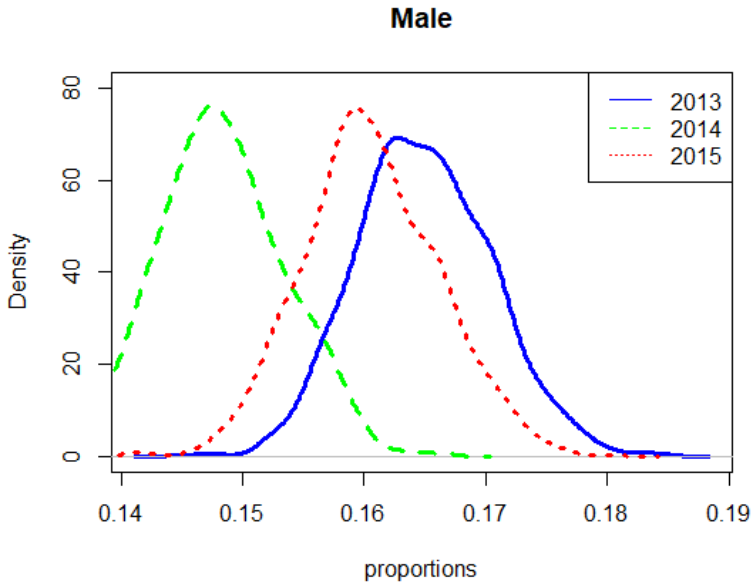
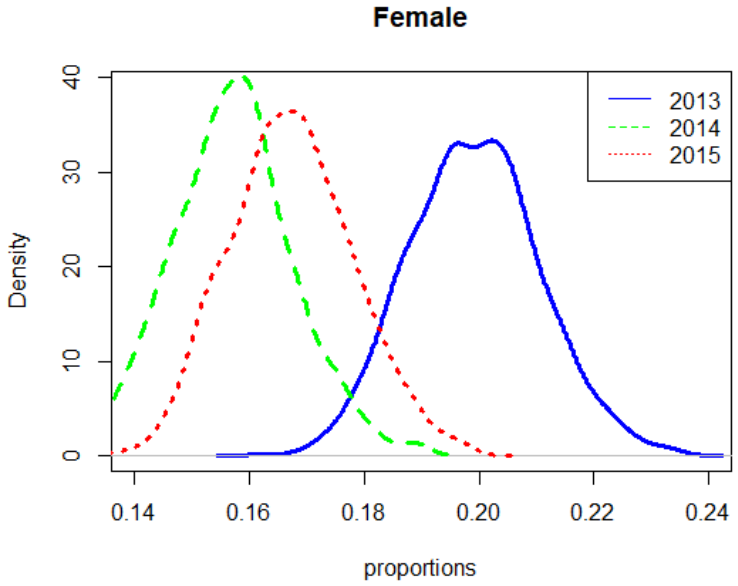


Figure 2:Genderwise cancer Distribution of patients having Tongue as the affected organ site





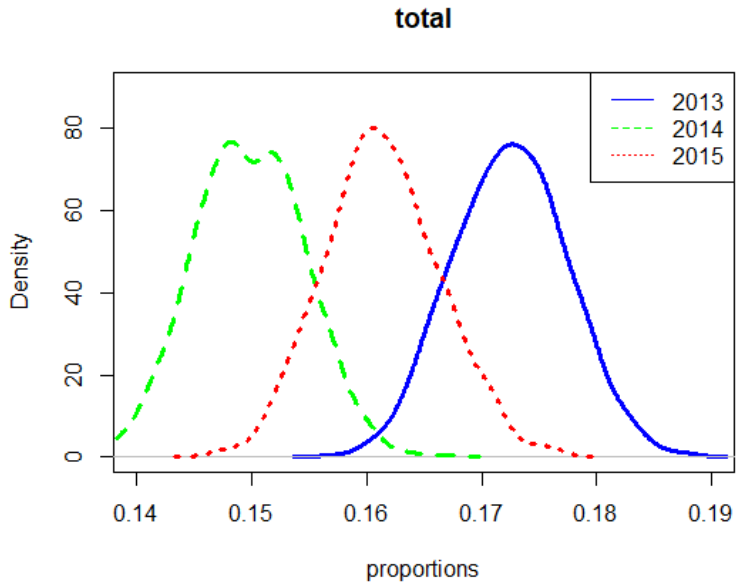
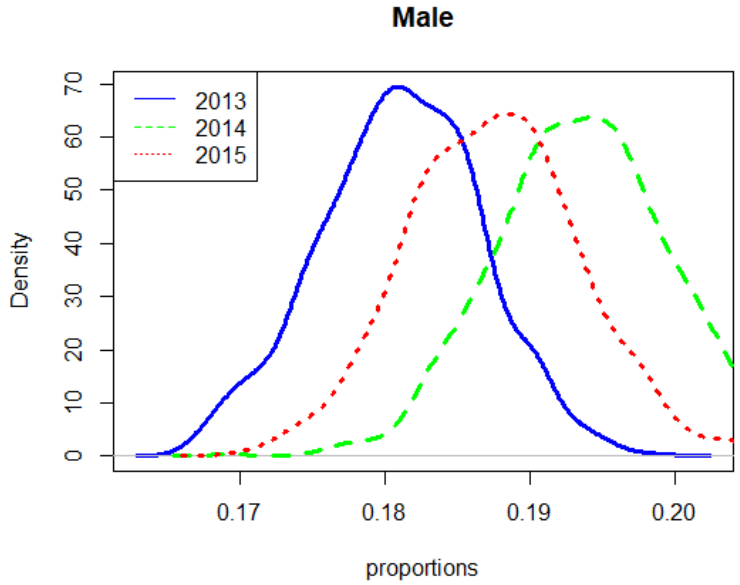
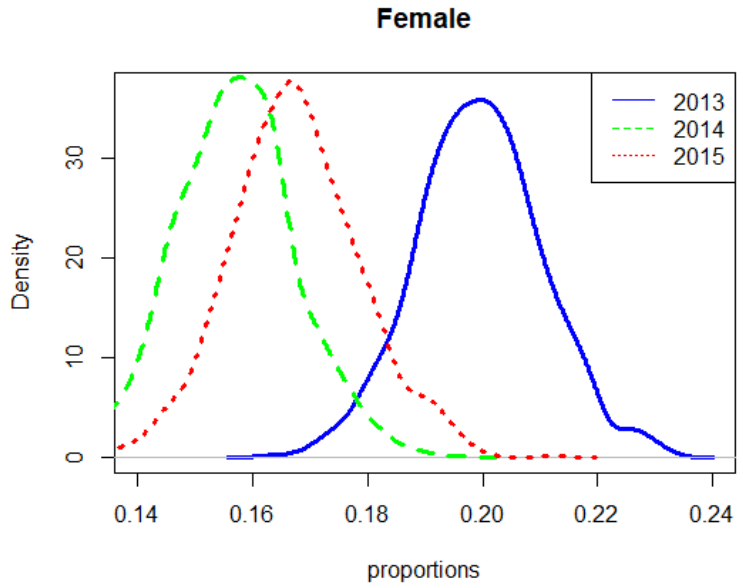


Figure 3:Genderwise cancer Distribution of patients having Mouth as the affected organ site





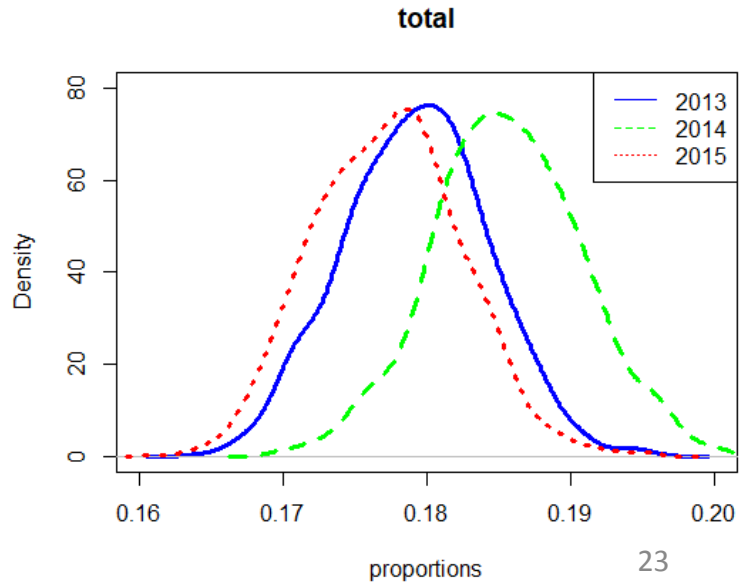
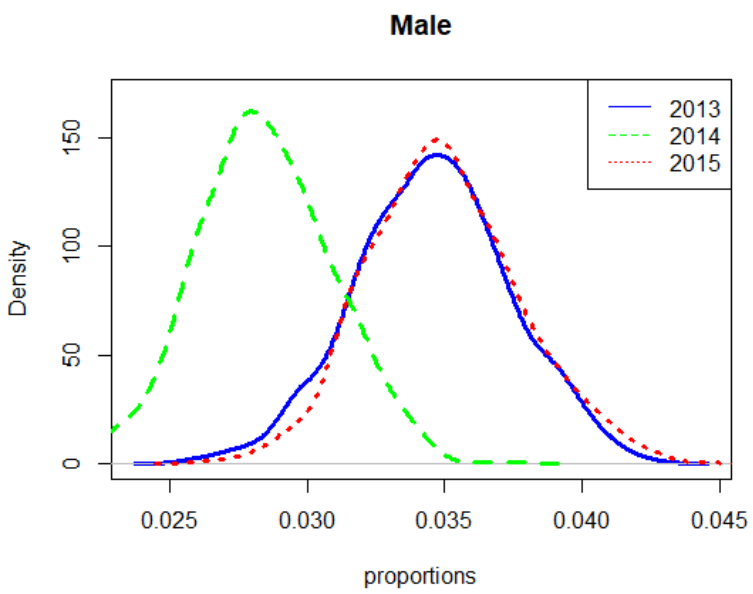
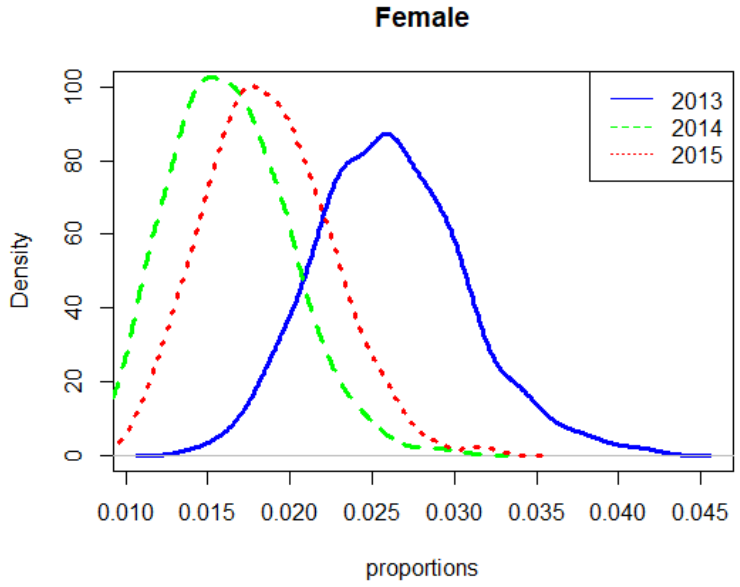


Figure 4:Genderwise cancer Distribution of patients having Oropharynx as the affected organ site





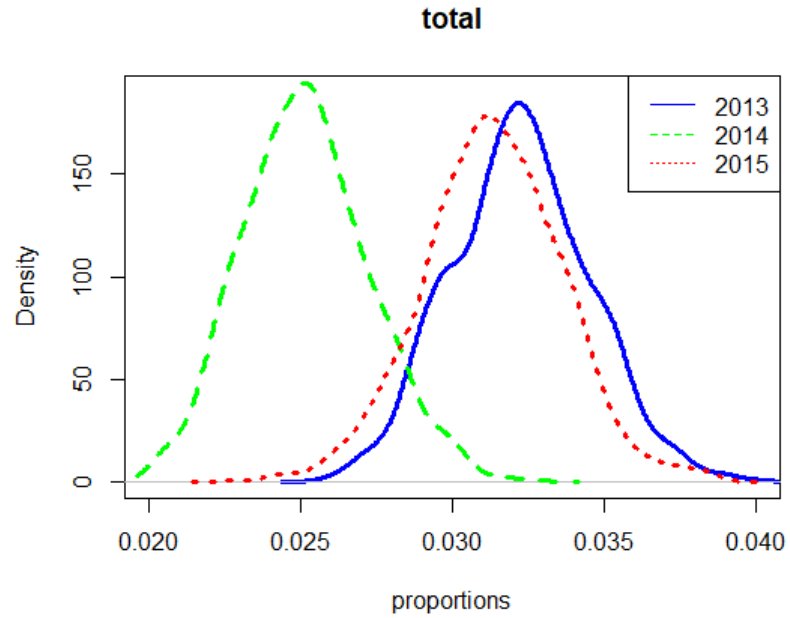
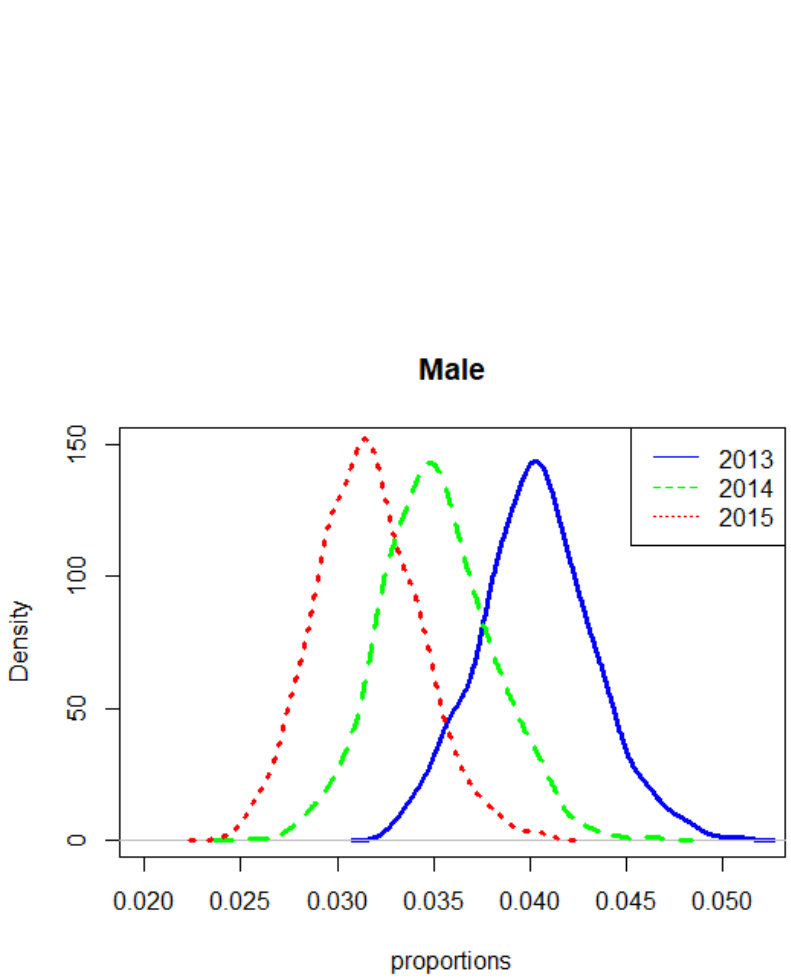
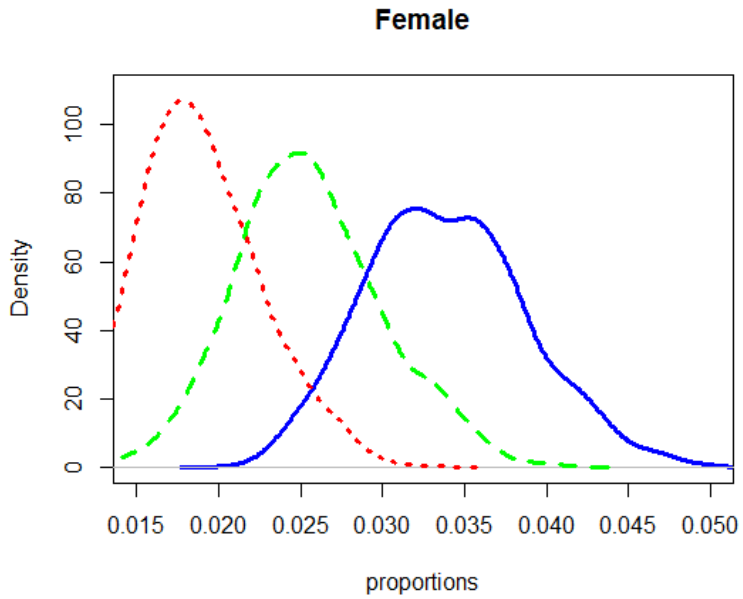


Figure 5:Genderwise cancer Distribution of patients having Hypopharynx as the affected organ site





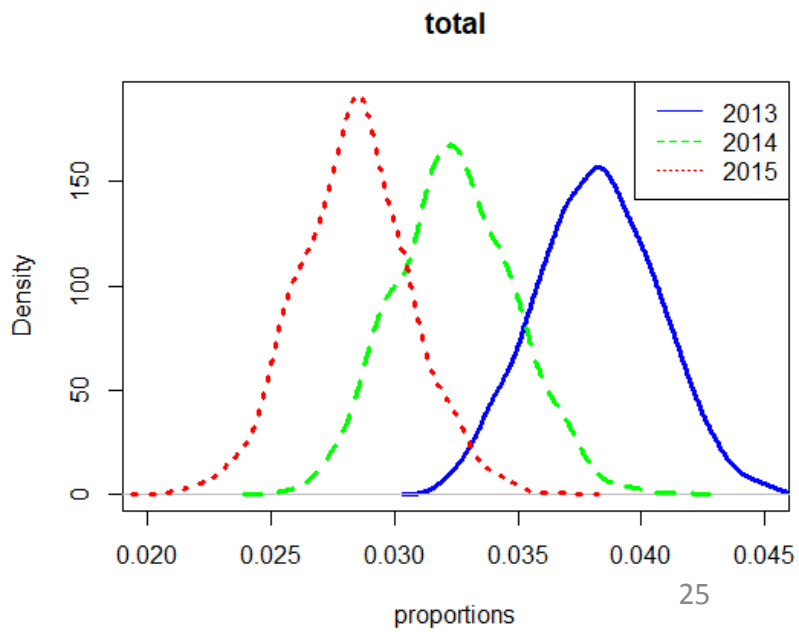
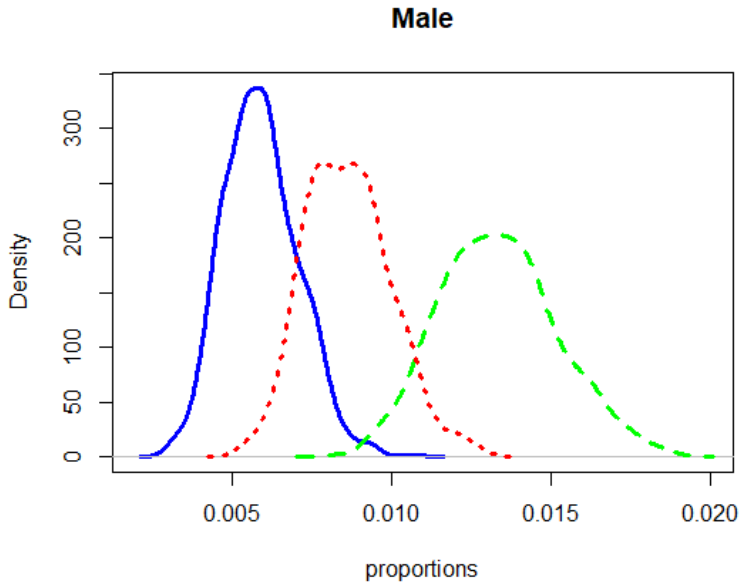
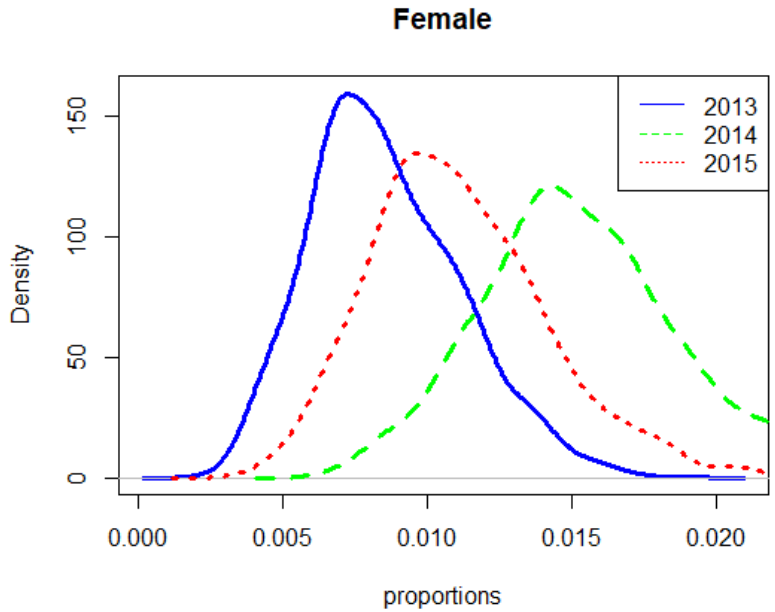


Figure 6 :Genderwise cancer Distribution of patients having Pharynx as the affected organ site





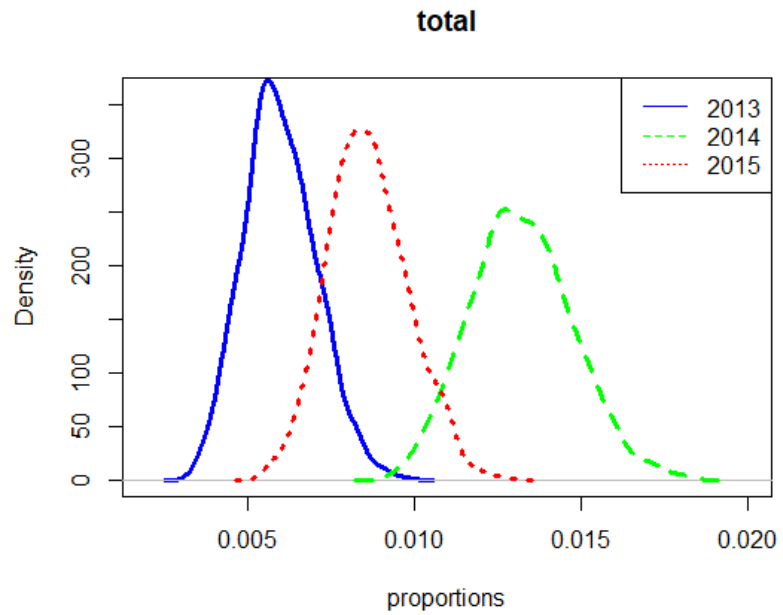
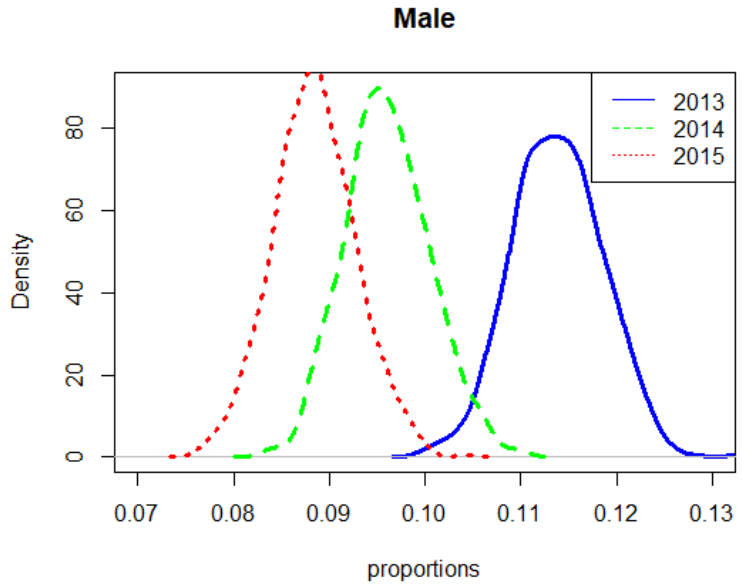
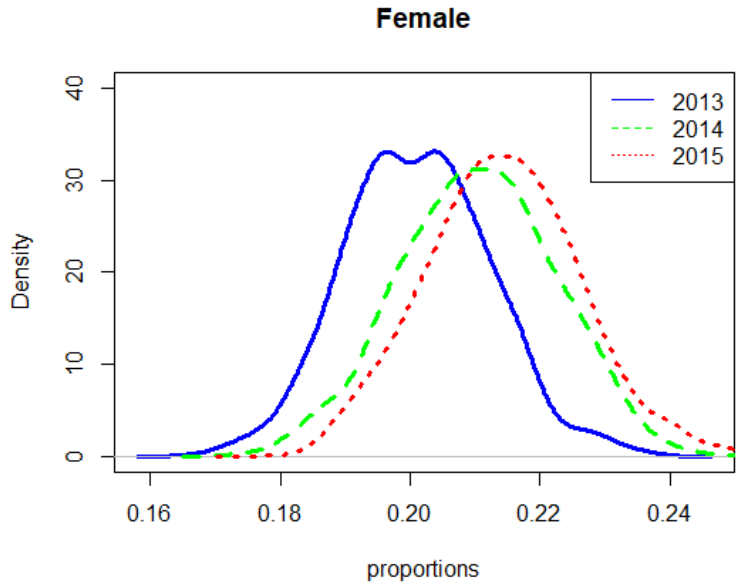


Figure 7 : Genderwise cancer Distribution of patients having Oesophagus as the affected organ site





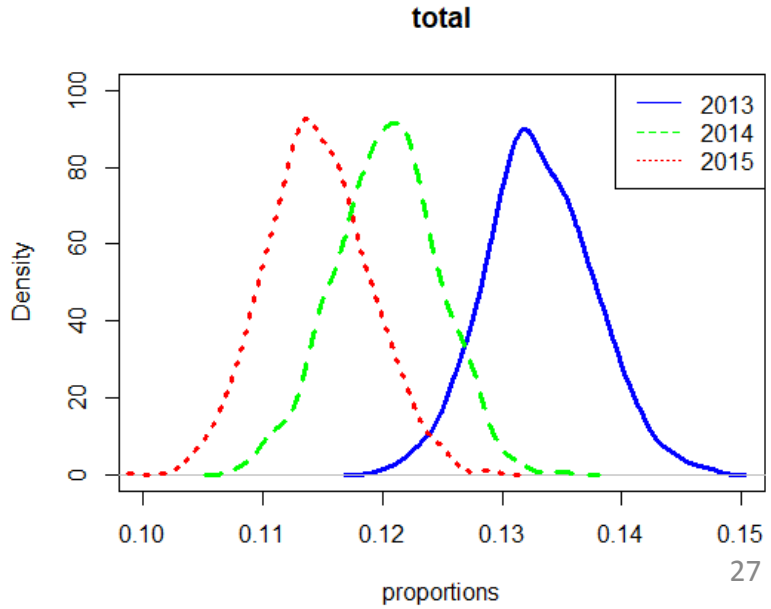
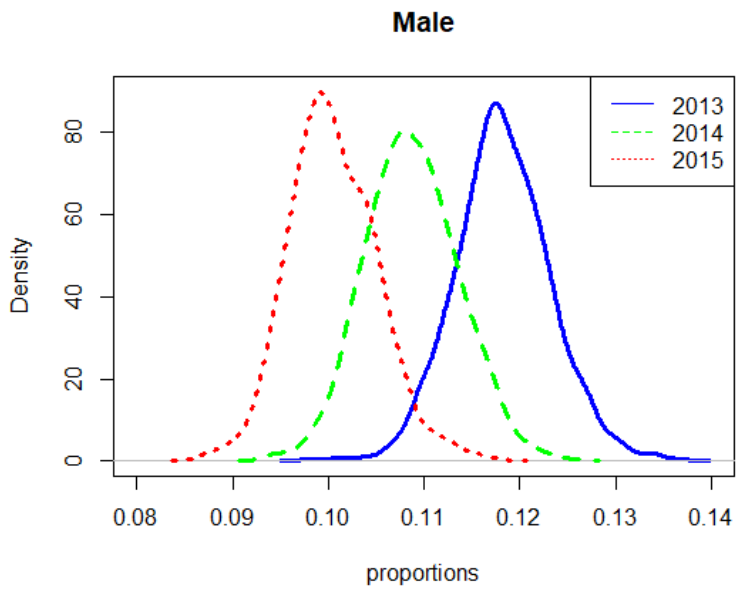
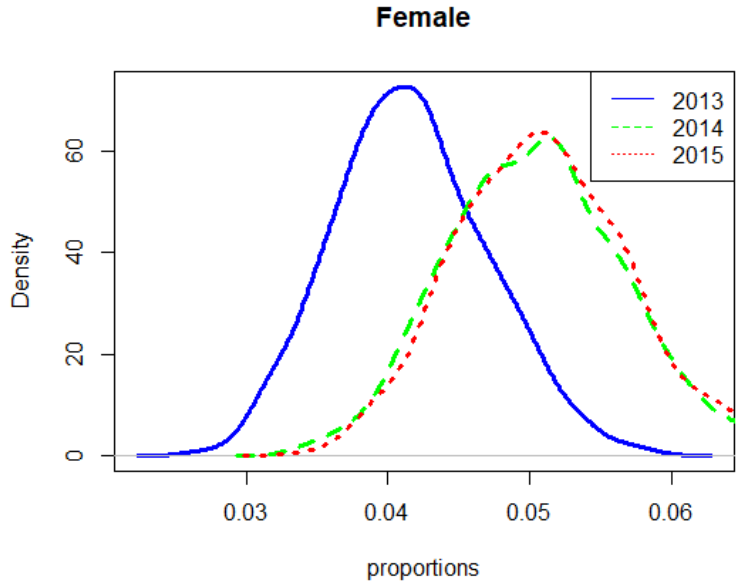


Figure 8 :Genderwise cancer Distribution of patients having Larynx as the affected organ site





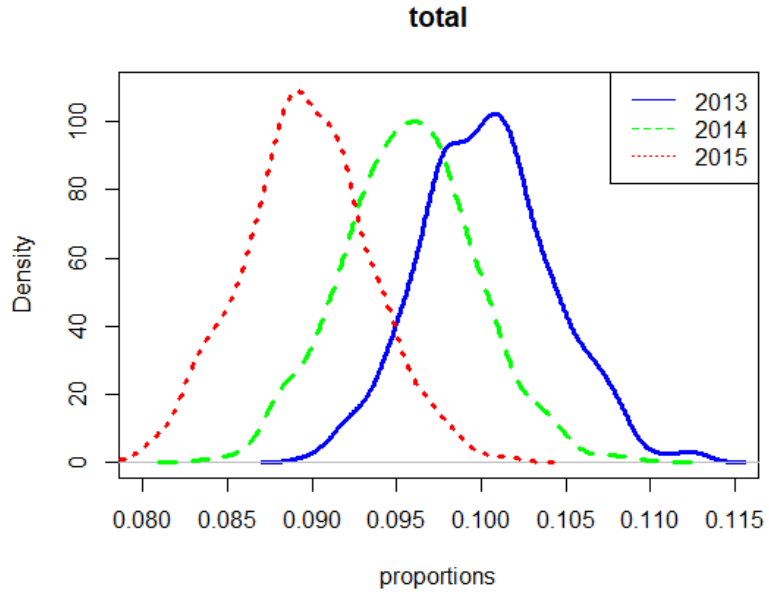


Figure 9 :Genderwise cancer Distribution of patients having Lungs as the affected organ site

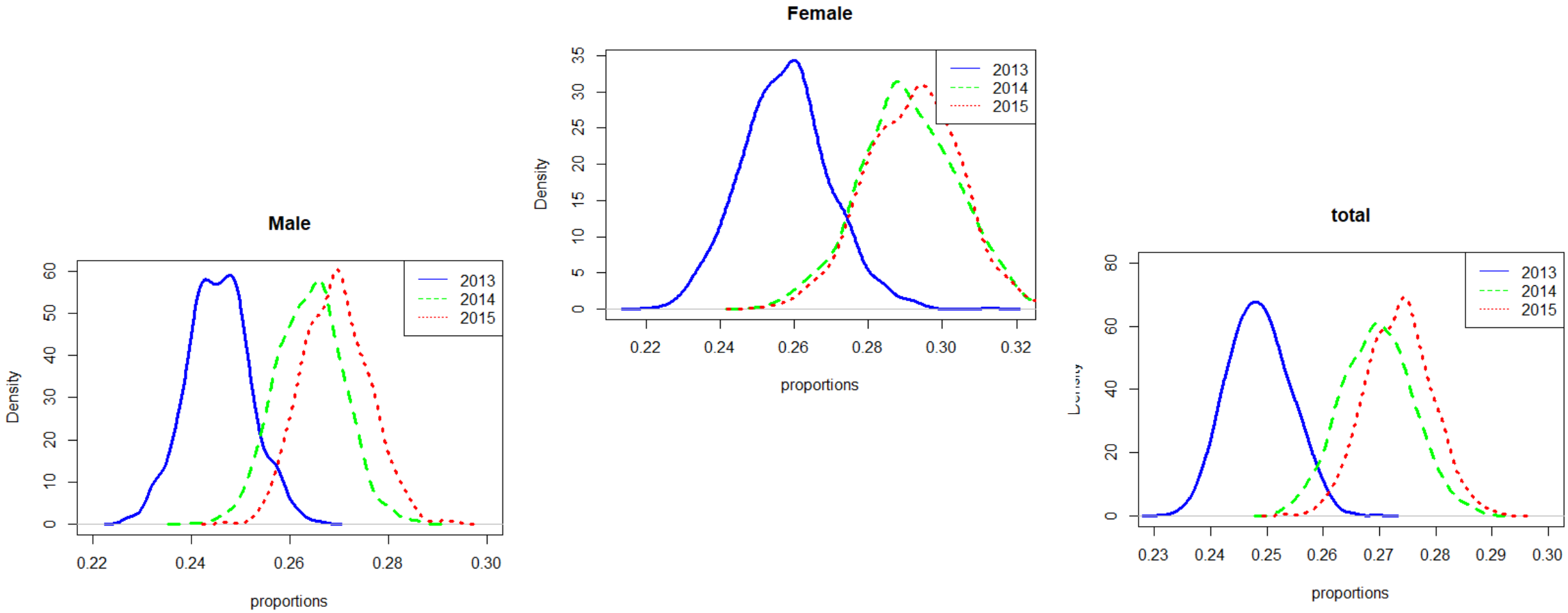
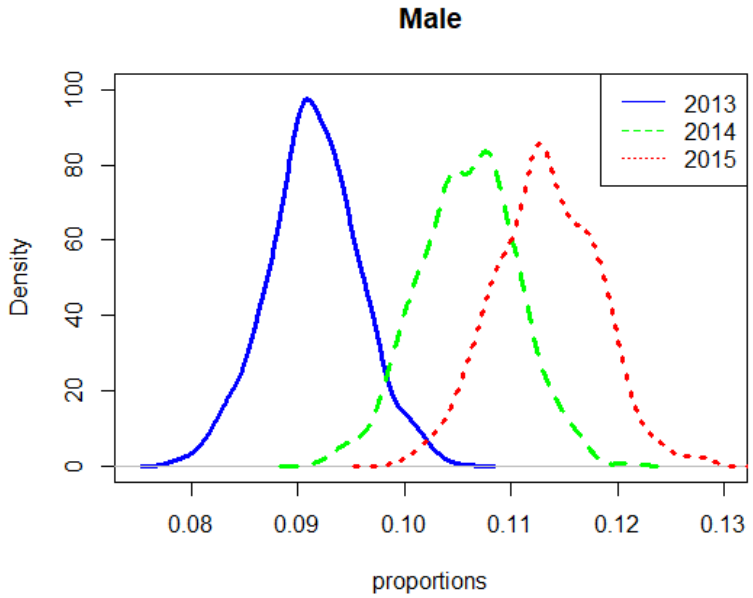
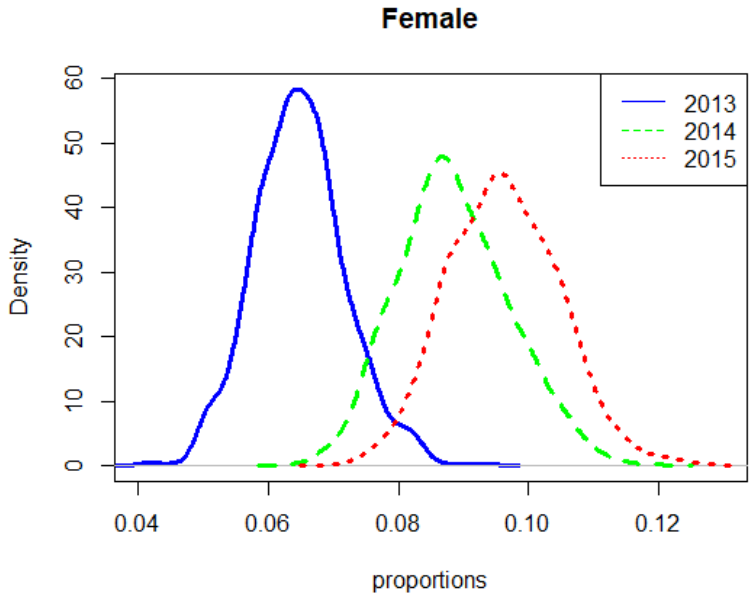
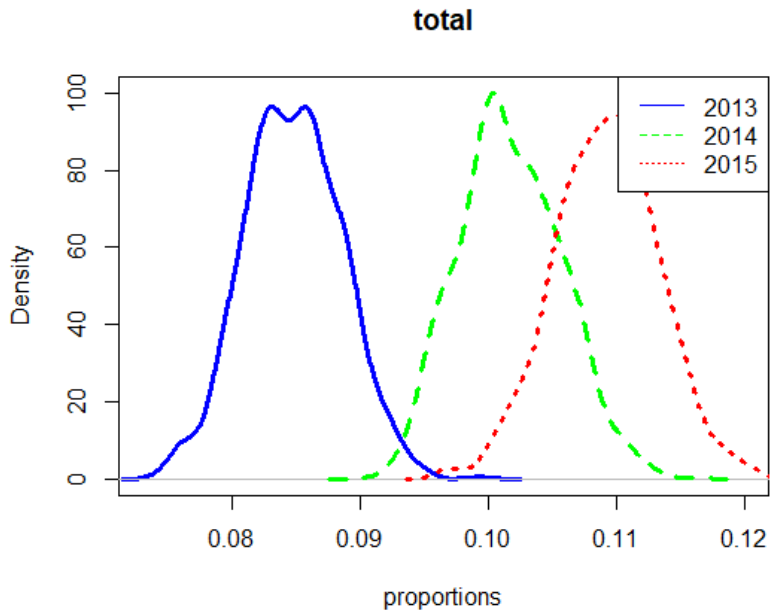


Figure 10 :Genderwise cancer Distribution of patients having Urinary bladder as the affected organ site







Results

- The interpretation of SDI is that values less than 1 denotes low diversity and thereby prevalence of cancer due to tobacco is uneven or not spreaded across both the genders and among sites of cancer. While SDI values more than 1 shows higher diversity and spread of cancer cases for the same.

- Table 1.1,1.2,1.3 depicts classical estimates of SDI for 2013,2014,2015 showing higher diversity among sites affected by cancer due to tobacco but low diversity between gender for each of the sites.

- The location wise diversity range from (1.86 to 1.97) and between gender for each of the sites ranges from(0.302 to 0.666) across 3 years

Table 1.1 has been shown that SDI estimates for 2013 across each sites ranging (0.302-0.642), where the lowest value observed(0.302) at Larynx while the highest value observed(0.642) at Oesophagus .

Table 1.2 has been shown that SDI estimates for 2014 ranging(0.343-0.660), where the lowest value observed(0.343) at Larynx while the highest value observed(0.660) at Oesophagus.

Table 1.3 has been shown that SDI estimates for 2015 ranging(0.344-0.666), where lowest value observed (0.344) at Oropharynx while the highest value observed(0.666) at Oesophagus

Thus, the classical estimates exhibit lowest diversity(closer to 0.3) in Larynx implying that mostly males are experiencing cancer in this specific site and Oesophagus has shown more diversity(closer to 0.6)

Under the Bayesian paradigm it is observed that SDI estimates show improved values compared to classical for various priors considered. Here the different prior combinations represent variations in the posterior distributions of the proportions.

- The table [2.1,2.2,2.3] depicts that Bayesian estimate seem to have a systematic pattern of derived SDI estimates.
- The diversity estimates for year 2013 for male, female and total ranged from (1.86-1.97)and the Bayesian estimates seem to range from(1.87-2.008) and almost similar for 2014 and 2015.
- So it is evident that Bayesian estimates have showed better results in terms of estimates and also with reduced variance in the estimates.

Density plots have been obtained in the fig 1 -10

- From fig 5, with passing consecutive years male , female and total proportions of cancer cases at Hypopharynx declined.
- From fig 7, with passing years male proportions declined but female proportions inclined cancer cases at Oesophagus. While overall cases declined with consecutive years.
- From fig 8, male cases in Larynx have declined but female cases increase from 2013 and remained same for both 2014 and 2015, while overall cases also declined.
- From fig 9 and 10, both male ,female and total lung cancer cases increase over the years.
- The following sites follows a common pattern but the rest of the sites show irregular pattern in the consecutive years.

References

[1] *Doll R.* The first reports on smoking and lung cancer. In: Lock SR, Tansey EM, editors. Ashes to ashes: the history of smoking and health, vol. 46. Rodopi: Amsterdam; 1998. p. 130–40.

[2] *Asthana S*, Patil RS, Labani S. Tobacco related cancers in India: A review of incidence reported from population-based cancer registries. Indian J Med Paediatr Oncol 2016;37:152-7.

[3] GlOBOCAN 2012: Word Cancer Facts – International Agency for Research on Cancer, WHO;2012.Available:http://www.globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

[4] Magurran AE, Ecoogical Diversity and its Measurement. Princeton:Princeton University Press; 1988.

[5] Verma, V., Mishra, A. K., Dhawan, A., & Nath, D. C. (2020). Diversity in substance use behaviour among street children of Delhi under Bayesian paradigm. *BMC Medical Research Methodology*, *20*, 1-9.

[6] Arnold, M., Razum, O., & Coebergh, J. W. (2010). Cancer risk diversity in non-western migrants to Europe: an overview of the literature. *European journal of cancer*, *46*(14), 2647-2659..

[7] Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American journal of political science*, 375-404

[8] Andradóttir, S.,Applying Bayesian ideas in simulation,Simulation Practice and Theory, Volume 8, Issues 3–4,2000,Pages 253-280

THANK YOU