Estimating household healthcare expenses for treatment of locally important diseases by sample size constrained network sampling

Arijit Chaudhuri and Shankar Dihidar[1]

Indian Statistical Institute

Kolkata, West Bengal, India

Abstract

We present a theory plus empirical report in estimating in a cost-effective way the household healthcare expenses for treatment of locally important diseases requiring institutional care within a compact area where getting access to remote households is difficult. Approaching traditionally we may encounter households containing no sufficient relevant data on healthcare expenses. So, we think it prudent to apply 'Network Sampling Technique' to get sampled households rich in relevant information. For this we start selecting various health centres. On gathering addresses of in-patients treated there within a specified time period we may locate households within the geographical area specified for the enquiry relatively easily. But an inherent hazard in this procedure is that total sample size may be extraordinarily large. So, to keep within our budget, with a self-inflicted bound on total sample size, we work out here necessary modifications on available literature on Network Sampling.

1. Introduction

It cannot be gainsaid that a motivating issue for us to ponder over undertaking this particular study was how 'Network Sampling as a statistical tool, may yield an appreciably substantial information content in 'Sample Survey- Based' data on healthcare possibly in excess over a possible traditionally alternative scheme of sampling. Thompson (1990, 1992) and Thompson and Seber (1996) have mentioned this survey situation which may be described as follows.

[1] Communicating author; email id: dshankar@isical.ac.in

We may have a list of an identifiable set of a known number of units, called selection units (SU). A listed frame for these SU's is available enabling one to make a selection out of them by any sophisticated method of sampling. But our interest is to estimate the total value of a variable defined on an unknown number of individuals called observational units. These observational units are not directly identifiable but may be contacted through the above selection units to one or more of which each observational unit is linked in a well defined manner. Each collection of observational units so linked is called a network. Each network formed this way is disjoint from every other and together they exhaust all the observational units of interest. Through a sample of selection units one may observe the variate values for the observational units in the networks linked to the sampled selection units. This approach of reaching samples of observational units through such networks is called network sampling. The above authors have given theories for estimation of total and variance estimation in case of simple random sampling without replacement and in stratified SRSWOR methods. Chaudhuri (2000) extends this problem to general sampling schemes with unequal probabilities. Chaudhuri and Stenger (2005, p-314) derive the gain in efficiency induced by network sampling in comparison to traditional survey.

The motivation of applying the above technique in our study project, is that, with a traditional approach the households might be selected in sample most of which do not contain any relevant data on expenses for healthcare. In order to remove this difficulty and to get the sampled household units rich in relevant information content we thought it appropriate to apply 'Network Sampling Technique' with the following description of selection and observational units. For this we choose a compact district, namely West Tripura District of the Hill state Tripura of India to collect data from both urban and rural areas where getting access to remote households is not easy. In our study the selection units are the hospitals, nursing Homes, health Centres etc located in that geographical region. We were aware that we could get hold of the residential household addresses of the in-patients treated in them for one or more of the three diseases of our interest, namely Heart disease, Gall Bladder and Cancer, during the period of our study period. So, all the unidentified and unknown number of households in the above mentioned geographical location, the inmates of which were in-patients in these healthcare centres during this period constitute our second category of units called Observational Units (OU). Our interest is to estimate certain parameters relating to these OU's.

In a traditional sample survey one may take an appropriately designed sample of households and ascertain the healthcare expenses incurred by them for treatment of their members for these diseases if there have been any. In the Network Sampling adopted by us we take a sample out of the SU's which are the healthcare centres using a frame for them. Then we establish a link between them and the OU's namely the households with their members having been in-patients in them and are thereby linked to the former. This scheme ensures our observational units to be rich in content because every one of them when visited must have incurred healthcare expenses over one or more of these three diseases. But sometimes it may be difficult to survey all the observational units linked to the sampled selection units because of having limited financial budget as well as limited time to complete the work. In that situation, it requires to take a suitable sub-sample of the observational units and hence needs the modification of available literature on network sampling to serve our purpose. Our experience of above situation is unfolded in details in subsequent sections.

In Section 2 we describe the technique of Network sampling and its uses in estimation. In Section 3 we discuss in detais the sampling scheme of the survey design employed by us. In Section 4 we give the explanatory notes on estimation procedure of the average expenses and also the unbiased estimator of the variance of the required estimator. In Section 5 we present our findings with some concluding remarks.

2. An outline of the technique of Network Sampling and its use in Estimation.

Let M ≡ the known total number of "Selection Units (SU)" which are identified as possibly linked to the entire set of an unknown and unidentified second category of units called " Observational Units (OU)". Let N be the unknown but true number of OU's.

Let Aj be the set of OU's 'linked to the jth SU', j=1,……,M.

m_i ≡ the number of SU's linked to the ith OU; of course i =1,…., N.

Let s be a sample of m SU's drawn from the population U of M SU's labelled j = 1,…., M. Let y be a real- valued variable of interest with values y_i for i=1,…, N. Our immediate object is to suitably estimate the population total $Y = \sum_{i=1}^{N} y_i$. Let $w_j = \sum_{i \in A_j} \frac{y_i}{m_i}, j \in s$.

Then, $W = \sum_{j=1}^{M} w_j = \sum_{j=1}^{M} \sum_{i \in A_j} \frac{y_i}{m_i} = \sum_{i=1}^{N} (\frac{y_i}{m_i}) \sum_{(j|A_j \ni i)} 1$ = Y.

So, it is enough to devise a suitable estimator for W using (s, w_j| j∈s) in our effort to estimate Y. This approach of estimating the total of the values defined on the set of OU's on establishing a 'link' with a well-defined set of SU's and selecting a sample of SU's is called 'Network sampling' used to estimate a parameter defined on the population of SU's , shown to equal a parameter of interest related to the OU's .

By way of illustration suppose, $p_j (0 < p_j < 1, \sum_{j=1}^{M} p_j = 1)$ be available as M normed size measures for the SU's. Then, a sample of m SU's from U may be selected on applying the celebrating scheme given by Rao, Hartley and Cochran (RHC, 1962). For this m mutually exclusive random groups are formed out of all the M SU's taking , say Mj SU's in the jth group (j=1,…,m), such that Σ_mMj = M, writing Σ_m to denote sum over the m groups. Optimal group sizes Mj's are suggested by Rao. Hartley and Cochran (1962) as follows:

$$M_j = \left[\frac{M}{m}\right] \quad , j=1,\ldots..,k$$

$$= \left[\frac{M}{m}\right] +1, j= k+1,\ldots\ldots.m$$

k is to be determined satisfying Σ_mMj = M. Writing Qj ≡ the sum of the values of pj's for the SU's falling in the jth group, j=1,…,m it follows that

$$t = \sum_m (\frac{Q_j}{p_j}) w_j$$

is an unbiased estimator for W . Writing Σ_M Σ_M as the sum over the pairs of distinct units with no repetition, RHC (1962) have shown that the variance of that estimator is $V(t)=A\sum_{M}\sum_{M}p_j p_{j'}\left(\frac{w_j}{p_j}-\frac{w_{j'}}{p_{j'}}\right)^2$ where $A=\frac{\sum_{m}M_j^2-M}{M(M-1)}$. Writing Σm Σm as the sum over the pairs of distinct groups with no repetition they have given a uniformly non-negative unbiased estimator of that variance V(t) as $v(t)=B\sum_{m}\sum_{m}Q_iQ_j(\frac{w_i}{p_i}-\frac{w_j}{p_j})^2$ writing $B=\frac{\sum_{m}M_j^2-M}{M^2-\sum_{m}M_j^2}$.

If s be a simple random sample taken without replacement (SRSWOR) in m draws from U, then, an unbiased estimator for W may be taken as $t^{'}=\frac{M}{m}\sum_{j\in s}w_j$

Then, an unbiased estimator for its variance $V(t')=M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{M-1}\sum_{j=1}^{M}\left(w_j-\frac{\sum_{j=1}^{M}w_j}{M}\right)^2$ is

$$v(t^{'})=M^2(\frac{1}{m}-\frac{1}{M})\frac{1}{m-1}\sum_{j\in s}\left(w_j-\frac{\sum_{j\in s}w_j}{m}\right)^2$$

In the respective cases we shall take estimated coefficients of

Variations, **cv** = $100\times\frac{\sqrt{v(e)}}{e}$, as measures of accuracy in estimation with e=t, t', as the case may be.

Whatever may be a sample s of SU's the corresponding 'Network Sample' namely the sample of OU's linked to s, may be quite large. As these OU's are also to be actually surveyed, remembering that s is also to be surveyed to establish the link with the corresponding OU's for the SU's in s, sometimes with the limited budget it may be difficult to implement the actual survey of the OU's contained in the 'Network Sample'. Then, we modify network sampling in the following way.

Let Cj ≡ the number of OU's in Aj. Then $C=\sum_{j\in s}C_j$ is the total number of OU's to be surveyed. If C is prohibitively large, let from the respective Aj's independently across j in s, SRSWOR’S of sizes dj (2 ≤ dj ≤ Cj) be taken as Bj’s such that $D=\sum_{j\in s} d_j$ is kept within a manageable number. This set namely $S_B=\bigcup_{j\in s} B_j$ is our

‘Modified Network Sample’ with a ‘constrained sample size’. Then the estimator t is to be modified as $u=\sum_{j\in s}\frac{Q_j}{p_j}\left(\frac{C_j}{d_j}\sum_{i\in B_j}\frac{y_i}{m_i}\right)=\sum_{j\in s}\frac{Q_j}{p_j}u_j$, say, writing $u_j=\left(\frac{C_j}{d_j}\sum_{i\in B_j}\frac{y_i}{m_i}\right)$. Then an unbiased estimator of the variance of u is

$$v(u)=(1+B)\sum_{m} v_L(u_j)\left(\frac{Q_j}{p_j}\right)^2+B\left[\sum_{j\in s}(u_j)^2\frac{Q_j}{p_j}-u^2\right]$$

writing $v_L(u_j)=C_j^2\left(\frac{1}{d_j}-\frac{1}{C_j}\right)\left(\frac{1}{d_j-1}\right)\sum_{i\in B_j}\left(\frac{y_i}{m_i}-\frac{1}{d_j}\sum_{i\in B_j}\frac{y_i}{m_i}\right)^2$.

Similarly, t' is to be replaced by $\hat{t}=\left(\frac{M}{m}\right)\sum_{j\in s}u_j$. Then an unbiased estimator of t' is

$$v(\hat{t})=\left(\frac{M}{m}\right)^2\left[\sum_{j\in s}v_L(u_j)+\left(\frac{M}{M-1}\right)\left(\frac{1}{m}-\frac{1}{M}\right)\sum_{j<j'}\sum_{\in s}(u_j-u_{j'})^2\right].$$

3. Sampling Scheme and Survey Design Employed

This study was undertaken during November, 2005 – March, 2006 by us of the Population Studies Unit of the Indian Statistical Institute , Kolkata. The time period covering the data specification was as 1.11.2004 – 31.10.2005.

First we divide the entire geographical area of the West Tripura district of the state of Tripura into 3 strata.

The stratum 1 consists of the Agartala City which is the capital territory of the state. This stratum is supposed to consist of 2 sub-strata which are labeled as follows:

1.1 Cancer Hospital is the only SU in the sub-stratum 1.1.

1.2 The Government Hospitals and the Nursing Homes in Agartala city are the SU’sin the sub-stratum 1.2.

The stratum 2 consists of a single sub-stratum 2.1 of Government Hospitals in other cities in the district.

The stratum 3 consists of all the Primary Heath Centres (PHC) and the General PHC's called CHC's in the district within urban and rural areas and denoted together as the sub-stratum 3.1

Since only one SU is selected from the single SU in sub-stratum 1.1 , the multiplier used in estimation is 1. Since from the 17 SU's in sub stratum 3.1 only 4 are selected by SRSWOR method the common multiplier is 17/4 for each of the 4 PHC/CHC's selected. From sub-stratum 2.1 only 2 SU's are selected employing RHC scheme using the number of beds as the size – measure. From sub-stratum 1.2 no SU could be selected as none was found. So, the multiplier symbolically is Q/p for the SU's selected.

From each selected SU whatever stratum / substratum it may belong to, the residential addresses, if they are within West Tripura district, are gathered of all the in-patients treated therein, during 1.11.04 – 31.10.05, of Heart trouble, Gall bladder irregularity and Cancer separately or in combination. The Households thus identified are the sampled OU's with their union as the 'Network Sample'. Deciding that more than 700 households may not be feasibly covered to ascertain the household expenditure during the specified period for treatment of these diseases in details and are mentioned in the attached Questionnaire we chose SRSWOR'S sub-stratum-wise from the sampled OU's which are the households as identified in the above manner. Thus we arrive at our modified Network Sample with constrained sizes. Then we proceed to estimate parameters related to expenses on treating the patients in respect of these 3 diseases through Health Care establishments.

4. Explanatory Notes for Estimation

(i) Besides estimating $Y=\sum y_i$, the total of a real-valued variable of interest over all the OU's in the entire population, we also do it strata and sub-strata-wise. Adding the estimates across the strata and sub-strata we arrive at the estimate for the population as a whole. Estimated variances of the estimates of the totals sub-strata and strata-wise are also added together to produce variance estimates for the population.

More importantly, introducing indicator functions

I_A (i) = 1 if ith unit belongs to a section A of a population (or stratum or sub-stratum)

= 0, otherwise

we define parameters

$Y_A = \sum y_i I_A$ (i)

and derive estimates and variance estimates quite easily.

Also we considered estimating ratio parameters

R =Y/X, where X is the total of a related variable x vis-à-vis y.

Then, if $\hat{Y}$ is an estimator for Y, applying the same formula we estimate X by $\hat{X}$ and hence R by $\hat{R} = \frac{\hat{Y}}{\hat{X}}$, then a variance –estimator, rather Mean Square Error (MSE) – estimator for R is taken as $mse = \frac{1}{\hat{X}^2} \hat{V}(\hat{Y})\Big|_{y=y-\hat{R}x}$.

Here $\hat{V}(\hat{Y})$ means a variance – estimator for $\hat{Y}$ and $\hat{V}(\hat{Y})\Big|_{y=y-\hat{R}x}$ is the value of $\hat{V}(\hat{Y})$ derived on replacing y in $\hat{V}(\hat{Y})$ throughout by $y - \hat{R}x$ values.

Keeping all these in mind we implemented our estimation plan as per the parameters defined consistently with our questionnaire attached. The results are tabulated as given below.

5. Tabulation of our findings

Table 1

Age, Sex and Disease-wise observed distribution of the households sampled

Age of In-patient treated	Heart Trouble (H)		
	Male	Female	Total
≤ 40	15	15	30
41 – 50	25	12	37
51-60	35	17	52
61-	31	34	65
Total	106	78	184
Cancer (C)			
	Male	Female	Total
≤ 40	14	2	16
41 – 50	28	18	46
51-60	55	19	74
61-	98	15	113
Total	195	54	249
Gall Bladder (G)			
	Male	Female	Total
≤ 40	7	63	70
41 – 50	16	66	82
51-60	35	53	88
61-	8	23	31
Total	66	205	271

Table 2

Estimated age-sex wise percentage distribution of the in-patients treated in Healthcare Centres in West Tripura District.

Heart Trouble (H)									
	Male			Female			Total		
Age of in-patient treated	Estimate (%)	cv (%)	Sample of HH's visited	Estimate (%)	cv (%)	Sample of HH's visited	Estimate (%)	cv (%)	Sample of HH's visited
≤ 40	23.1	34.4	15	20.8	14.9	15	22.7	21.0	30
41-50	22.0	28.0	25	20.0	22.9	12	21.0	23.6	37
51-60	32.2	18.1	35	15.7	30.1	17	26.7	15.6	52
61	22.7	22.4	31	43.5	9.1	34	29.6	13.5	65
Total	100.0		106	100		78	100.0		184
Cancer (C)									
	Male			Female			Total		
≤ 40	6.4	22.6	14	3.7	54.4	2	5.6	26.9	16
41-50	14.8	14.4	28	35.2	15.1	18	19.3	10.8	46
51-60	29.5	9.1	55	37.0	14.5	19	31.1	7.9	74
61	49.3	5.8	98	24.1	19.2	15	44.0	5.7	113
Total	100.0		195	100.0		54	100.0		249
Gall Bladder (G)									
	Male			Female			Total		
≤ 40	107	35.7	7	33.4	13.0	63	30.2	15.2	70
41-50	14.8	44.6	16	27.8	19.5	66	25.1	14.2	82
51-60	60.8	11.7	35	19.5	20.2	53	26.2	15.6	88
61	13.7	28.6	8	19.3	28.3	23	18.5	25.3	31
Total	100.0		66	100.0		205	100.0		271

Table 3

Estimated per household expense during 1.11.2004- 31.10.2005 for treatment in healthcare centers (place of residence wise, religion wise and social class-wise)

Heart Trouble (H)				
	Rural	Urban	Total	Total sample size
Estimated expenses (Rs.)	33668	57084	44267	
cv (%)	23.7	22.1	18.5	
	Hindu	Muslim	Others	
Estimated Expenses (Rs)	44652	12496	46320	
	Scheduled Caste	Scheduled Tribe	Others	
Estimated Expenses (Rs)	34867	25463	45875	184
Cancer (C)				
	Rural	Urban	Total	Total sample size
Estimated expenses (Rs.)	75440	89214	79912	
cv (%)	7.9	12.1	5.8	
	Hindu	Muslim	Others	
Estimated Expenses (Rs)	84032	37704	48915	
	Scheduled Caste	Scheduled Tribe	Others	
Estimated Expenses (Rs)	64602	30697	84595	244

Table 3 continued.

Gall Bladder (G)

	Rural	Urban	Total	Total sample size
Estimated expenses (Rs.)	12215	13193	12585	
cv (%)	21.5	19.9	19.9	
	Hindu	Muslim	Others	
Estimated Expenses (Rs)	12627	3935		
	Scheduled Caste	Scheduled Tribe	Others	
Estimated Expenses (Rs)	14051	5074	12348	270

Table 4

Estimated medical and non-medical expenses per household

(inside and outside district separately)

Cancer					
	Medical		Non-medical		n
	Est. (Rs)	CV	Est. (Rs)	CV	
Inside	40112	3.5	3932	6.7	244
Outside	26651	11.3	9217	16.3	51
Total	66763	5.2	13149	11.7	244
Heart					
	Medical		Non-medical		n
	Est. (Rs)	CV	Est. (Rs)	CV	
Inside	34187	18.3	10080	19.5	184
Outside	15635	18.3	2235	19.8	184
Total	18552	20.1	7845	20.1	44
Gall Bladder					
	Medical		Non-medical		n
	Est. (Rs)	CV	Est. (Rs)	CV	
Inside	10373	26.6	2214	12.5	270
Outside	-	-	-	-	-
Total	10373	26.6	2214	12.5	270

Concluding Remarks

We set a target of covering a total of 700 households in the West Tripura District. But we could actually get responses to our queries from 698 households in all. This is clear from our Table 2. But the Table 1 accounts for a total of 704 households claimed to be covered. This is not an enigma. We would gather data on 704 household members who received treatments as in-patients for Heart trouble, Cancer and/ or Gall Bladder irregularities altogether out of a total of 698 households. No contradiction is apparent.

Among those treated for Cancer only about 22% are females and only 41% are female among those who are treated for Heart related diseases. But this picture is reversed in case of Gall Bladder- here 76% are females. In the younger ages females treated for Gall Bladder far outnumber the male households members. These relate to the facts experienced for the sampled households only.

When we turn to estimation for the entire district we find that among those aged 50-60 years half the people are women while among those who are elderly still, half of the people are men- this is among those who suffer from heart trouble. Among the cancer patients treated in healthcare institutions the female / male ratios are about ½ among those aged 50-60 years and 2/5 among those aged higher still respectively. Among the gall bladder patient these ratios are 1/3 and 1:1 roughly in these 2 age-groups.

Per household expenses for treatment of Cancer is quite huge both in the urban and rural areas, not quite worthy of attention the same for heart trouble and for Gall bladder treatment. The expenses are not great enough to cause any serious alarm. The expenses however vary significantly across villages versus cities, Hindus Vs Muslims and caste-wise. About the accuracies in estimation we may only mention that except when the sample –size is too small the coefficient of variation is quite low in magnitude indicating desirable accuracy level.

Possibly contrary to popular anticipation even in case of treatment of cancer the costs incurred within the particular district exceed those outside. But incidental expenses incurred outside far exceed those within, possibly for travel costs of the patients and their attendants for heart-treatment also the expenses incurred outside the district are quite less compared with those inside. But the same is the case for those on incidental items. This probably implies that for treatment of heart trouble people do not greatly move out.

For treatment of gall bladder irritation probably the people do not seek attention from outside the district. Also incidental expenses vis-a-vis those for actual treatment are rather insignificant.

References

Chaudhuri, Arijit (2000): *Network and Adaptive Sampling with unequal probabilities*. Cal. Stat. Assoc. Bull-50,237-253.

Chaudhuri, Arijit and Stenger, Horst (2005): *Survey Sampling: Theory and Methods (2nd edition).* Chapman and Hallm New York.

Rao, JNK, Harteley, H.O. and Cochram, W.G. (1962): *On a simple procedure of unequal probability sampling without replacement*. Jour. Roy. Stat. Soc. 24, 482-491.

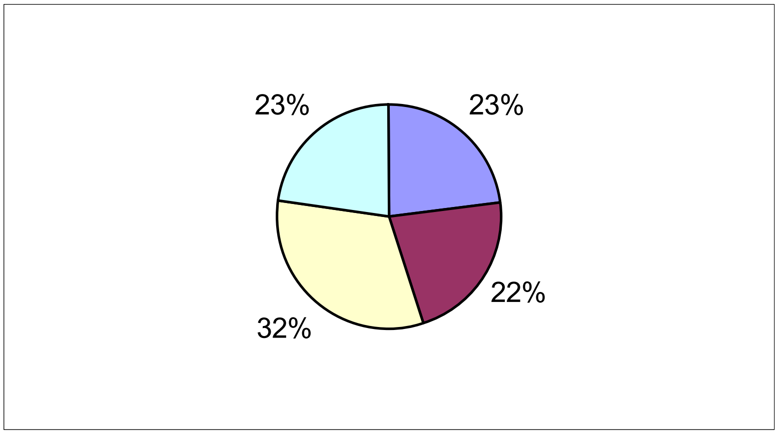
Thompson, S.K. (1990): *Adaptive Cluster Sampling.* Jour. Amer. Assoc.85, 1050-1059.

Thompson, S.K. (1992): *Sampling*- John Wiley & Sons. N.Y.

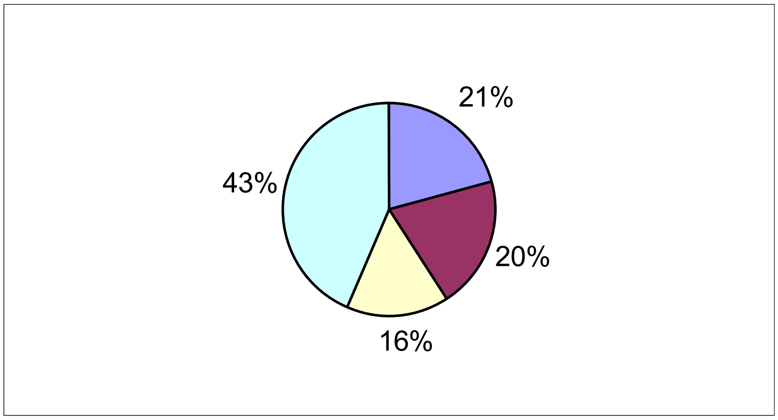
Thompson, S.K. and Serber, G.A.F. (1996): *Adaptive Sampling*. John Wiley & Sons. New York.

Graphical representations of our results

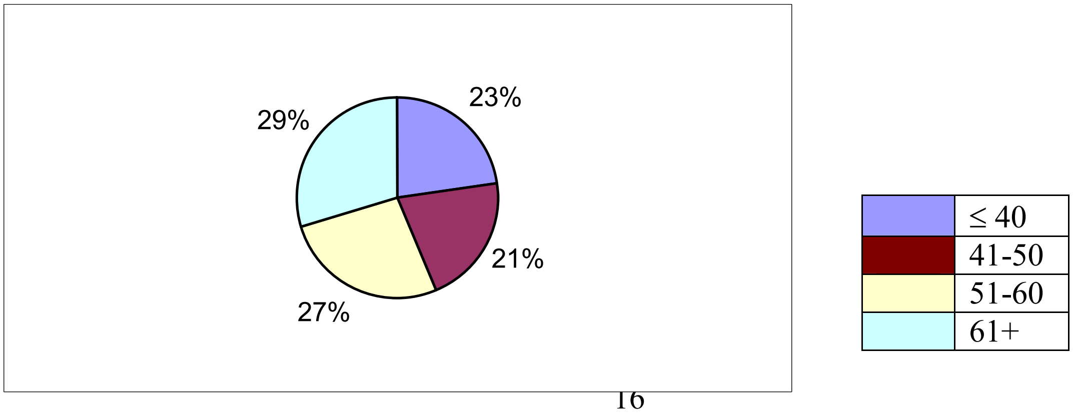
Age-wise distribution of the male in-patients treated in Healthcare Centres in West Tripura District for Heart Trouble



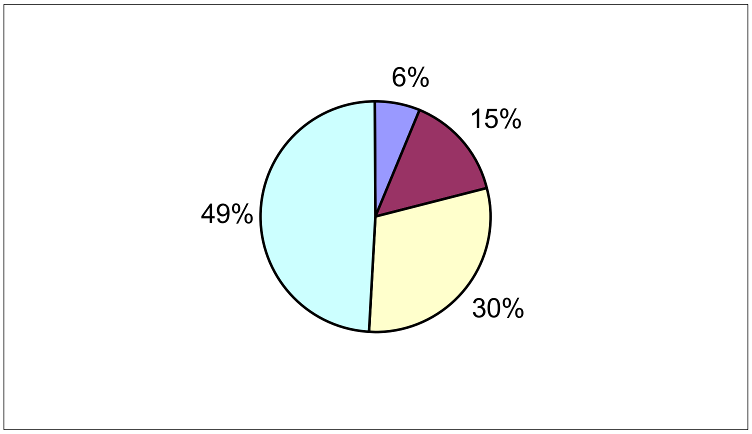
Age-wise distribution of the female in-patients treated in Healthcare Centres in West Tripura District for Heart Trouble



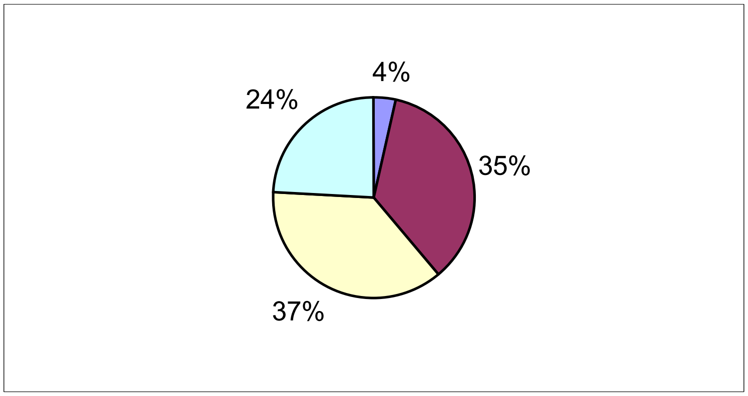
Age-wise distribution of all the in-patients treated in Healthcare Centres in West Tripura District for Heart Trouble



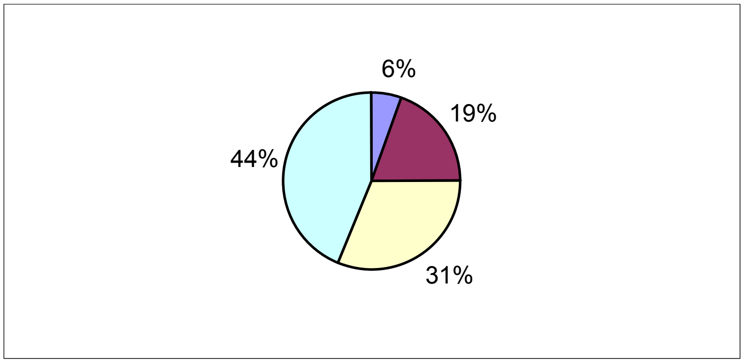
Age –wise distribution of the male in-patients treated in Healthcare Centres in West Tripura District for Cancer



Age-wise distribution of the female in-patients treated in Healthcare Centres in West Tripura District for Cancer

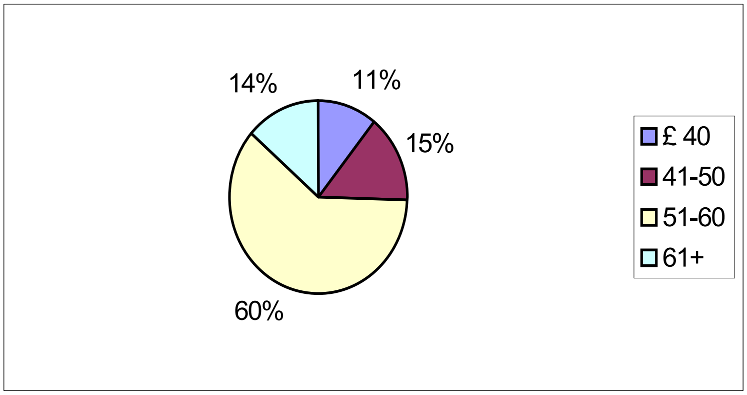


Age-wise distribution of all the in-patients treated in Healthcare Centres in West Tripura District for Cancer

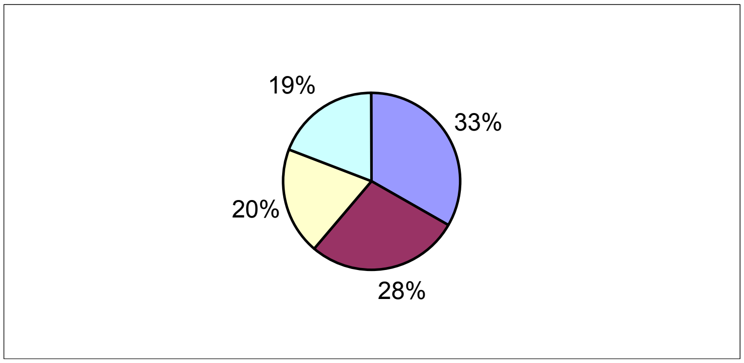


🟪	≤ 40
🟥	41-50
🟨	51-60
🟦	61+

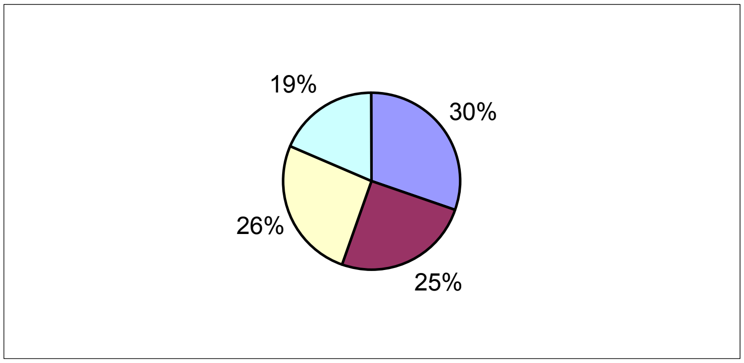
Age-wise distribution of the male in-patients treated in Healthcare Centres in West Tripura District for Gall Bladder



Age-wise distribution of the female in-patients treated in Healthcare Centres in West Tripura District for Gall Bladder

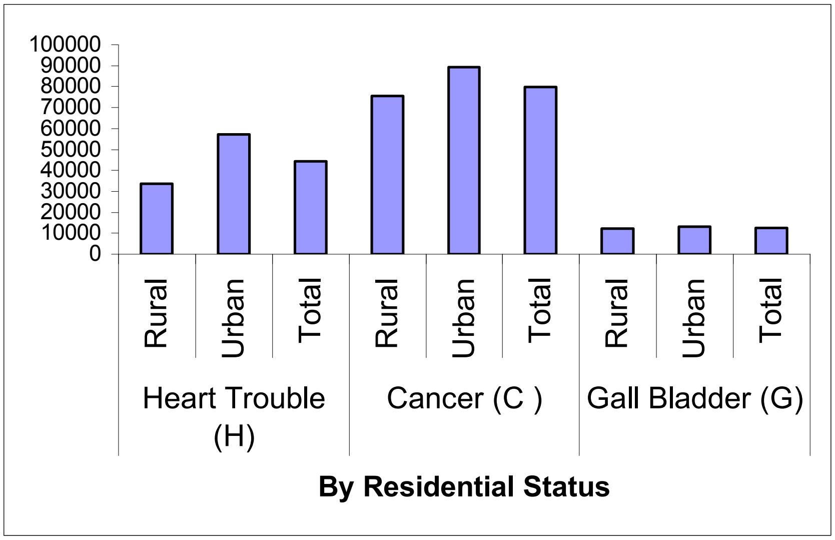


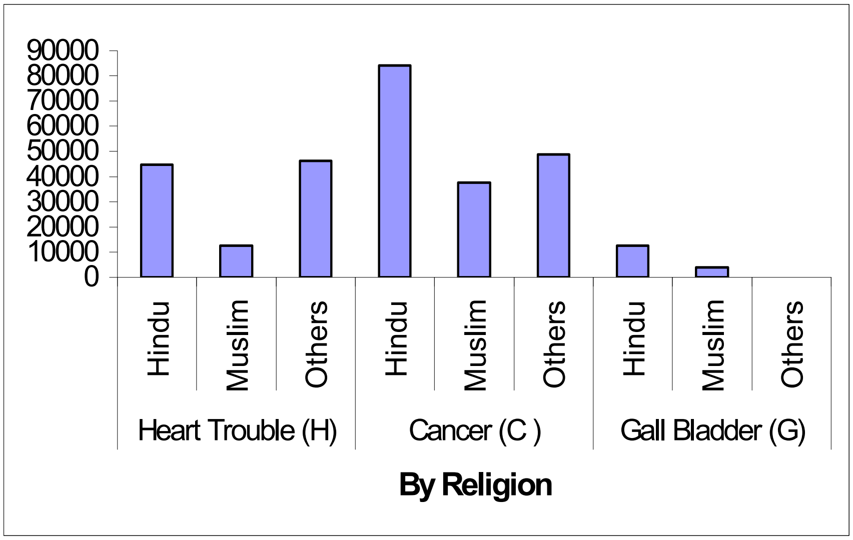
Age-wise distribution of all the in-patients treated in Healthcare Centres in West Tripura District for Gall Bladder

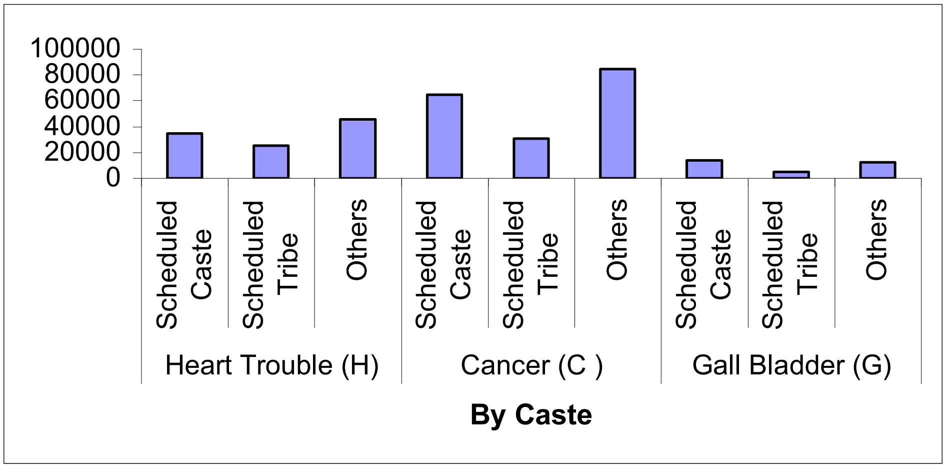


🟪	≤ 40
🟥	41-50
🟨	51-60
🟦	61+

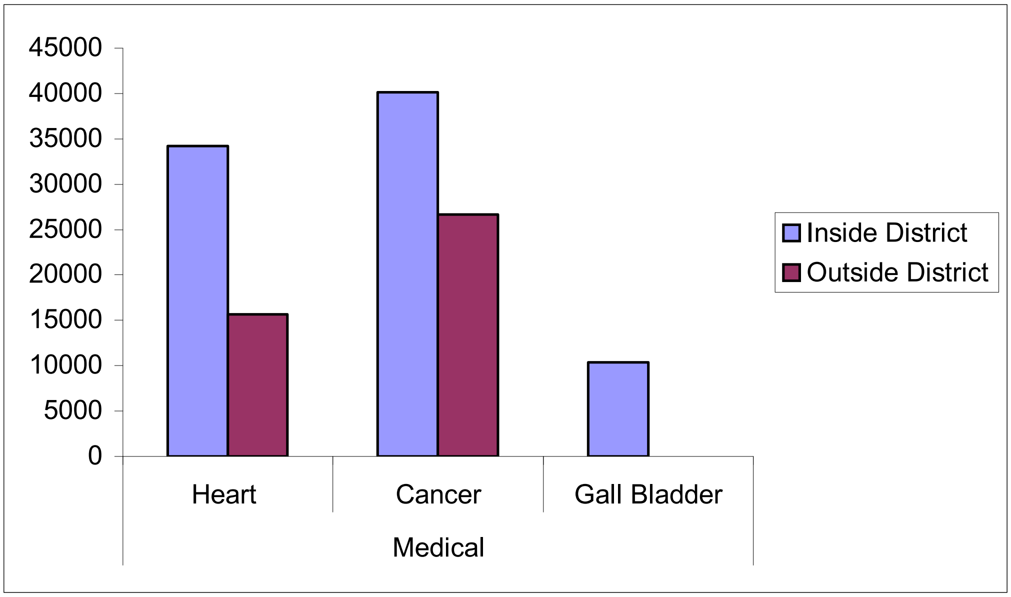
Distribution of Household expense (Rs.) for treatment in Healthcare centers disease and class-wise during 1.11.2004- 31.10.2005



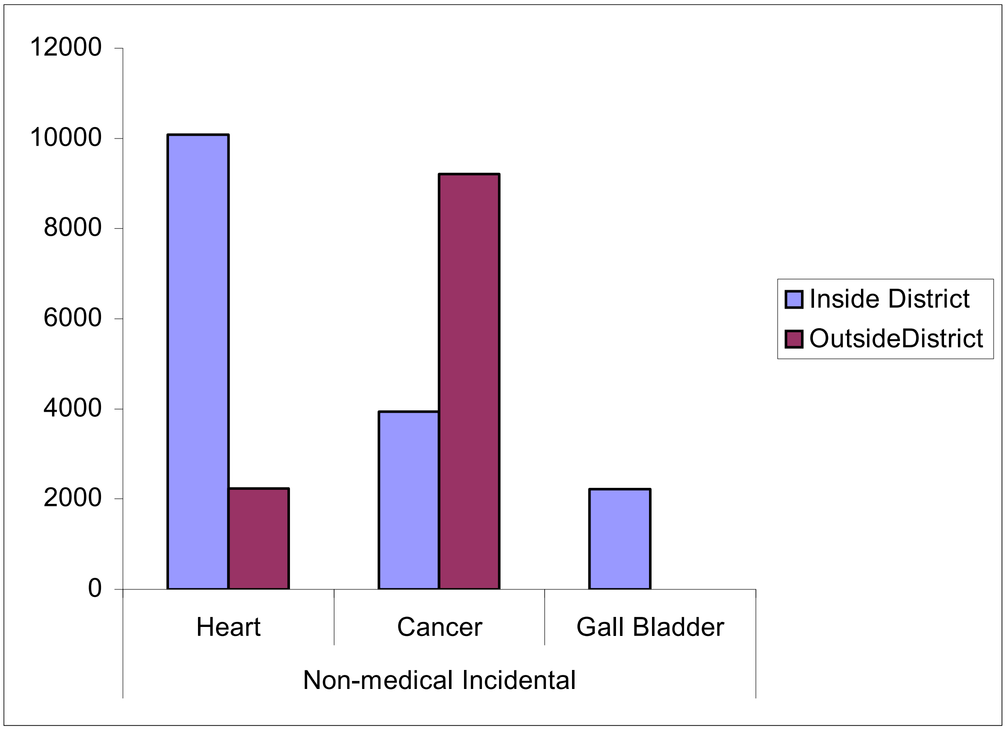




Medical expenses per household by disease-type



Non-medical expenses per household by disease-type


Heart
Cancer
Gall Bladder
Non-medical Incidental
Inside District
OutsideDistrict