International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 3, Issue 1, pp: (105-122), Month: April 2015 - September 2015, Available at: www.researchpublish.com

Parametric and Non-Parametric Design Based Tests Analysis of the Level and Differentials of Household Consumption Expenditure in Rwanda (2010-2011)

Roger KAMANA [1], Dr. Joseph K. Mung'atu [2], Dr. Marcel Ndengo [3]

[1] Students at Jomo Kenyata University of Agriculture and Technology (JKUAT/KIGALI CAMPUS), Master of Science in applied statistics
[2] Lecturer at Jomo Kenyata University of Agriculture and Technology/Nairobi, Kenya
[3] Lecturer at Jomo Kenyata University of Agriculture and Technology/ Kigali, Rwanda

***Abstract*: The purpose of this project was to analyze the household consumption expenditure in Rwanda. The main objective of this study was to measure the level of household consumption expenditure in Rwanda and to show the contribution of each household consumption expenditure component in the total household consumption expenditure.**

The government wants to know where people spend much money for some items for them to be able to budget for food security, non food and housing consumption expenditure. This information is not always available and people keep changing their spending.

The project studied the distribution of the household consumption expenditure data and the mean differences between household consumption expenditure by component, by urban and rural areas, by sex of head of household and by household composition size and household adult equivalent size.

The research found out that the distribution of household consumption expenditure data was not approximately normally distributed, and Kigali city was the province with higher household consumption expenditure mean and the last was southern province. The research also revealed that, in Rwanda, households spent much money in food consumption, followed by non-food and the last was alcohol. The statistical hypothesis tests found out that the household consumption expenditure mean differences between components, urban and rural areas, household headed by male and female, and between household adult equivalent and household composition size were statistically significant at 95% confidence level.

***Keywords:* Parametric and non-parametric Test, Test of Normality, Household Consumption expenditure, Household size, Adult equivalent household size.**

I. INTRODUCTION

1.1 Statement of the problem:

Households have personal needs and wants that are directly satisfied through consumption of goods and services resulting from activities that are reproductive in an economic (SNA, 1993). These goods and services are referred to as consumer goods and services and their individual value is defined as the consumption expenditure on these goods or services. Household consumption expenditure (HCE) is the value of consumer goods and services that were acquired (used or paid for) by a household for the direct satisfaction of the needs and wants of its members through direct monetary purchases in the market; through the market-place but without using any money as means of payment (barter, income in kind) and from production within the household (own-account production). Households also acquire (or use) consumer goods and services that satisfy the needs and wants of its members through social transfers in kind from government and non-profit institutions or through transfers from other households. The sum of HCE and the value of these transfers are referred to as actual final consumption (AFC) of the household. The non-consumption expenditures of households include current transfers of cash, goods and services to other households such as gifts donated, remittances, alimony, child supported. Other items included are irregular contributions to non-profit institutions; compulsory transfers to governments such as income and other direct taxes (e.g. wealth taxes), compulsory fees and fines; and pension and social security contributions. Expenditures on goods and services for use in the operation of unincorporated enterprise as well as the occupational expenses of employees are excluded from the measurement of household expenditure. In addition, capital expenditures such as savings, reduction of liabilities, amounts loaned, purchases of financial assets, life insurance premiums are excluded. Expenditure on valuables (Works of art, jewellery, gemstones, etc.) is also excluded from household expenditure (ICLS-2003-06-0049-1-EN.Doc/v1).

Therefore household consumption expenditure is a big challenge in Rwanda. The government need to know where people spend much money for some items for them to be able to budget for food security, non food, education, health and etc. This information is not always available and people keep changing their spending.

This study want to analyses the distribution of household consumption expenditure, the level of household consumption expenditure by province, household consumption components, sex of head of household, urban and rural status and to analyze the house size in terms of consumption and composition.

1.2 Objectives:

1.2.1General Objective:

The main objective of this study was to measure the level of household consumption in Rwanda and to show the contribution of each household consumption expenditure component in the total household consumption expenditure.

1.2.2Specific Objectives:

This study had the following Specific Objectives:

1. To analyze the distribution of the household consumption expenditure data
2. To determine the level of households consumption expenditure in Rwanda
3. To analyze the mean households consumption expenditure by component in Rwanda
4. To determine the mean households consumption expenditure by Urban and Rural area
5. To analyze the households consumption between households headed by Males and household headed by females.
6. To determine the mean of household size with the mean of adult equivalent household size

1.3 Research Questions:

1. Is household consumption expenditure data normally distributed?
2. What is the level of household consumption expenditure in Rwanda?
3. Is the mean of household consumption expenditure component equal in Rwanda?
4. Does household consumption expenditure in urban area equal to mean household consumption expenditure in rural area?
5. Do households headed by Male spend much more than those household headed by female?
6. Does the mean of household composition size is different from the mean of household adult equivalent size?

1.4 Justification:

This study first of all was design in the purpose of obtaining a Degree in Master Science in Applied Statistics and it will be used by different people such are:

It can help the government to know, where to allocate resources in mostly used consumption expenditure component.

It can help many investors to know where they can invest according to the classification of the most household consumption expenditure component in Rwanda.

It can help other researchers to understand the methodology used in Rwanda to compute the household consumption expenditure.

1.5 Scope:

This study used household consumption expenditure variables from secondary data of integrated household living condition survey dataset conducted in 2010/2011 by NISR with a sample of 14310 households which was selected country wide.

II. METHODOLOGY

2.1 Sampling technique:

2.1.1. Sampling frame:

The sampling frame for the Integrated household living condition survey (EICV3) was based on a database of villages (umudugudu) that cover all of the households in Rwanda. This database includes information on all the geographic codes and the approximate number of households in each village. The geographic hierarchy of the villages in the sampling frame was based on the new administrative divisions of Rwanda: 5 provinces, 30 districts, 416 sectors, 2148 cellules and 14837 villages. The average number of households per village was 132 (168 for urban villages, 129 for rural villages). The urban and rural classification was based on the 2002 Rwanda Census of Population.

In each sample village all the households were listed. This provided an updated sampling frame for the second stage of selection.

2.1.2. Stratification:

A stratified random sample is one obtained by separating the population elements into no overlapping groups, called strata, and then selecting a simple random sample from each stratum. See Richard L. Scheaffer, William Mendenhall III, R. Lyman Ott and Kenneth Gerow (1996).

Regarding to the design of the study we found that the stratification sampling technique was the best sampling technique compare to other methods because the estimates for the survey were estimated up to district level considered as strata. “According to Richard L. Scheaffer, William Mendenhall III, R. Lyman Ott and Kenneth Gerow (1996)” they mentioned three reason for using stratification sampling technique and the nature of this study satisfy those conditions.

The sampling frame of villages was stratified by district. Within each district the villages were ordered by urban and rural classification, then by geographic codes to provide an implicit stratification by urban and rural classification, and geographic location. This resulted in a proportional distribution of the sample villages by urban and rural classification.

2.1.3. Sample Size and Allocation:

According to De Landsheere .G. (1982,p. 382), sampling is the fact of choosing a limited number of individuals, objects, events which the observation allows to draw conclusions applicable to the whole population from which the choice has been made. The sample size was designed for measuring the poverty reduction in Rwanda. To determine the sample of our study, we have used the following formula:

Let define:

p : Proportion of population below poverty line in 2005 (was the key indicator, 56%)

q : Proportion of population above poverty line (46%)

n : Sample size

RR : Response Rate (96%)

h : Expected household size (5 members)

E : Margin error (2%)

Z : Confidence level at 95% (1.96)

Deff: Design Effect (1.5)

The sample size was computed by using the formula below:

$$n = \frac{Z^2 * p * (1-p) * Deff}{E^2 * RR * h}$$

This gives a minimum sample size of 442 households of which was increased up to 480 households in rural stratum and 450 households in urban stratum and the total strata was 30 (3 urban and 27 rural).

In the entire country, 1230 PSUs (Villages) were selected in the 30 strata with probability proportionally to size of PSUs in each of the 30 strata (District) and for each selected PSU, one second Stage sampling (SSU) unit or in this case households were randomly selected using a simple systematic sampling.

This was result in a sample of size 40 villages per district for the full year (10 cycles). In the case of the three districts in Kigali Province there has been 50 sample villages each; since 9 households have been selected in each sample village, the sample size for these districts were 450 households each. Although the sample size per district was slightly smaller for Kigali Province than for the other provinces, the 50 sample villages in each of these urban districts will ensure a more disperse sample, and the smaller number of sample households per village were reduced the design effects for the urban estimates.

2.1.4. Probability of selecting a primary sampling unit (P_α) and household (p_β):

Let define:

PPS: Probability proportional to size

MOS_α : Measure of Size (Number of households in a specific village)

P_α :Probability of selecting a specific village

n : Number of village to be selected in a specific stratum

The first stage sample of villages in each district was selected systematically with PPS, where the measure of size was based on the number of households in the frame for each village. According to PPS, the probability of selecting a certain village (p_α) is:

$$p_\alpha = \frac{n\ MOS_\alpha}{\sum MOS_\alpha}$$

At the second sampling stage the households within each sample village were selected with equal probability using systematic random sampling.

According to the simple random sampling (SRS), the probability of selecting household in primary sampling unit (p_α) is:

$p_\beta = \frac{\gamma}{MOS_\alpha}$ Where ‘ γ ’ is the sample size (households) to be selected from the district, and MOS_α is the census number of households in the α^{th} primary sampling unity (PSU).

2.1.5 Design Weight:

The Design weight of the sample denoted by W is merely the inverse of the selection probability. It was computed as follows:

$$W = \frac{1}{P_\alpha * P_\beta}$$

This weight was based on census household in PSU (Village) and for updating the number of households in PSU a listing operation was done and the number for households in PSU after listing was used to adjust the weight. The adjusted weight denoted by W_{adj} was computed as follows:

International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 3, Issue 1, pp: (105-122), Month: April 2015 - September 2015, Available at: www.researchpublish.com

Let define:

MOS_1 : Number of households listened in the PSU

MOS_2 : Number of household in PSU from census

$$W_{adj} = W * \frac{MOS_1}{MOS_2}$$

2.2 Estimation Methodology:

As the stratified random sampling techniques have been adopted for the survey, below are the formula used to estimate various statistical parameters. Some notation is required for stratified random sampling. The suffix h denotes the stratum and i the unit within the stratum. The following symbols all refer to stratum h. Let

N_h : Total number of units, n_h : Number of units in sample, y_{hi} : Value obtained for the i^{th} unit

$W_h = \frac{N_h}{N}$: Stratum weight, $f_h = \frac{n_h}{N_h}$: Sampling fraction in the stratum

$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$: True mean, $\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$: Sample mean, $S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2}{N_h - 1}$: True variance

2.2.1 Estimation of a Population Mean and Total:

Let $\bar{y}_{st}$ denote the sample mean for the simple random sample selected from

Stratum h, n_h the sample size for stratum h, μ_h the population mean for stratum h, and τ_h the population total for stratum i. Then the population total τ is equal to $\tau_1 + \tau_2 + ... + \tau_L$ We have a simple random sample within each stratum. Therefore, we know that $\bar{y}_{st}$ is an unbiased estimator of μ_h and $N_h \bar{y}_{st}$ is an unbiased estimator of the stratum total $\tau_h = N_h \mu_h$.

It seems reasonable to form an estimator of τ, which is the sum of the τ_h values, by summing the estimators of the τ_h. Similarly, because the population mean μ equals the population total τ divided by N, an unbiased estimator of μ is obtained by summing the estimators of the τ_h overall strata and then dividing by N. Cochran, Willam Gemmell (1909).

We denote this estimator by $\bar{y}_{st}$, where the subscript st indicates that stratified random sampling is used.

a. Estimator of the population mean:

$$\bar{y}_{st} = \frac{1}{N}[N_1 \bar{y}_1 + N_2 \bar{y}_2 + ... + N_L \bar{y}_L] = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$$

The estimate $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean, $\bar{y}$, can be written as $\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n}$.

The difference is that in $\bar{y}_{st}$ the estimates from the individual strata receive their correct weights $\frac{N_h}{N}$. It is evident that $\bar{y}$ coincides with $\bar{y}_{st}$ provided that in every stratum. $\frac{n_h}{n} = \frac{N_h}{N}$ or $\frac{n_h}{N_h} = \frac{n}{N}$ or $f_n = f$. This means that the

International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 3, Issue 1, pp: (105-122), Month: April 2015 - September 2015, Available at: www.researchpublish.com

sampling fraction is the same in all strata. If in every stratum the sample estimate $\bar{y}_h$ is unbiased, then $\bar{y}_{st}$ is unbiased estimate of the population mean $\bar{y}$. $E(\bar{y}_{st}) = E \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{Y}_h$, Cochran, Willam Gemmell (1909).

b. Estimator of the population total:

Procedures for the estimation of a population total τ follow directly from the procedures presented for estimating μ. Because τ is equal to $N\mu$, an unbiased estimator of τ is given by $N\bar{y}_{st}$.

$$N\bar{y}_{st} = N_1\bar{y}_1 + N_2\bar{y}_2 + \ldots + N_L\bar{y}_L = \sum_{i=1}^{L} N_h \bar{y}_h$$

c. Estimated variance of total:

If a simple random sample is taken within each stratum, an unbiased estimate of S_h^2 from theorem 2.4, Cochran, Willam Gemmell (1909) is

$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ With stratified random sampling, an unbiased estimate of the variance of $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = S^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h},$$

An alternative form for computing purposes is $S^2(\bar{y}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n} - \sum_{h=1}^{L} \frac{W_h s_h^2}{N}$

The second term on the right represents the reduction due to the fpc. In order to compute this estimate, there must be at least two units drawn from every stratum. Willam Gemmell, Cochran (1909)

d. Estimate confidence limits.

The formulas for confidence limits are as follows

Population mean: $\bar{y}_{st} \pm ts(\bar{y}_{st})$, Population total: $N\bar{y}_{st} \pm tNs(\bar{y}_{st})$

These formulas assume that $\bar{y}_{st}$ is normally distributed and that $s(\bar{y}_{st})$ is well determined, so that the multiplier t can be read from tables of the normal distribution. If only a few degrees of freedom are provided by each stratum, the usually procedure for taking account of the sampling error attached to a quantity like $s(\bar{y}_{st})$ is to read the t-value from the tables of student's t instead of from the normal table. Cochran ,Willam Gemmell(1909)

2.3 Statistical test:

A statistical test provides a mechanism for making qualitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. Not rejecting may be a good result if we want to continue to act as if we "believe" the null hypothesis is true. Or it may be a disappointing result, possibly indicating we may not yet enough data to "prove" something by rejecting the null hypothesis.

2.3.1 Shapiro-wilk test:

The Shapiro-will test utilizes the null hypothesis principle to check whether a sample $x_1 \ldots x_n$ came from a normally distributed population. The test statistic is:

$$w = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Where: x_i is the i^{th} order statistic; $\bar{x}$ is the sample mean; The constants a_i are given by: $(a_1,\ldots,a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$

Where:

$m = (m_1,\ldots,m_n)^T$ and $m_1,\ldots,m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics. We reject the null hypothesis if W is below a predetermined threshold.

The null hypothesis is this test is that the population is normally distributed. Thus if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not from a normal distributed population. In other words, the data are not normal. On the contrary, if the p-value is greater than the chosen alpha level, the null hypothesis that data came from a normal distributed population cannot be rejected, since the test is biased by sample size the test may be statistically significant from a normal distribution in any large samples. Thus a Q-Q plot is required for verification in addition to the test. Pearson, A. V., and Hartley, H. O. (1972).

2.3.2 Student's t-test:

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value.

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\text{Estimated standard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is zero. The hypothesized value is the null hypothesis that the difference between population means is 0. For the calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$m_1 - m_2$ Where m_2 is the mean for group 1 and m_2 is the mean for group 2

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means, so the estimated standard error of the statistic is

$$(s_{m_1 - m_2})$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2. Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances.

Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{S_1^2 + S_2^2}{2}$$

Where:

MSE is our estimate of σ^2; $S_{m_1-m_2} = \sqrt{\frac{2MSE}{n}}$ for equal sample size and $S_{m_1-m_2} = \sqrt{\frac{2SSE/df}{n_h}}$ for unequal sample size

Where: $n_h = \frac{2}{\frac{1}{n_1}+\frac{1}{n_2}}$; $SSE = \sum(X-m_1)^2 + \sum(x-m_2)^2$

The next step is to compute t

$t = \frac{m_1 - m_2}{S_{m_1-m_2}}$ (Gopal K. Kanji 2006).

2.3.3 One way ANOVA Test:

A one-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

Assumptions

i. The populations from which the samples were obtained assumed to be normally or approximately normally distributed.

ii. The samples must be independent.

iii. The variances of the populations are assumed to be equal.

Hypotheses:

The null hypothesis will be that all population means are equal; the alternative hypothesis is that at least one means is different. In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size. (Anderson, R.L. (1942)

Let

$\overline{X}GM = \frac{\sum x}{N} = \frac{\sum n\overline{x}}{\sum n}$: Grand mean, $SS(T) = \sum\left(x - \overline{X}GM\right)^2$: Total variation

$SS(B) = \sum n\left(\overline{X} - \overline{X}GM\right)^2$: Interaction between the samples (There are k-1 degrees of freedom)

$MS(B) = S_b^2 = \frac{SS(B)}{K-1}$: Mean square between groups

$SS(W) = \sum df * s^2 = \sum(N-K) * s^2$: Sum squares within groups

F-test statistic:

Recall that the F statistic is the ratio of two independent chi-square variables divided by their respective degrees of freedom. Also recall that the F-test statistic is the ratio of two sample variances, well, it turns out that's exactly what we have here. The F-test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group $(k-1)$ and the degrees of freedom for the denominator are the degrees of freedom for the within group $(N-K)$.

$$F = \frac{S_b^2}{S_w^2}$$

Decision Rule:

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with $K-1$ numerator and $N-K$ denominator degrees of freedom. see Gopal K. Kanji (2006)

2.3.4 Kruskal–Wallis test:

The Kruskal-Wallis H test is a non-parametric test which is used in place of a one-way ANOVA. Essentially it is an extension of the wilcoxon rank-sum test to more than two independent samples.

This test is appropriate for use under the following circumstances:

1. You have three or more conditions that you want to compare;

2. Each condition is performed by a different group of participants; i.e. you have an independent-measures design with three or more conditions

3. The data do not meet the requirements for a parametric test. (i.e. use it if the data are not normally distributed; if the variances for the different conditions are markedly different; or if the data are measurements on an ordinal scale).

The K samples are combined and arranged in order of increasing size and given a rank number. Where ties occur the mean of the available rank numbers is used. The rank sum for each of the K samples is calculated. Let Rj be the rank sum of the j^{th} sample, n_j be the size of the j^{th} sample, and N be the size of the combined sample. The test statistic is

$$H = \left\{ \frac{12}{N(N+1} \sum_{j=1}^{K} \frac{R_i^2}{n_j} \right\} - 3(N+1) \text{ Gopal K. Kanji (2006)}$$

This follows a $\chi^2 -$distribution with K-1 degrees of freedom. The null hypothesis of equal means is rejected when H exceeds the critical value. Critical values of H H for small sample sizes and K = 3, 4, 5 are given in Table 22(Gopal K. Kanji 2006). Each sample size should be at least 5 in order for χ^2 to be used, though sample sizes need not be equal. The K frequency distributions should be continuous.

Decision rule:

If $\chi^2_{\text{calculated}} > \chi^2_{\text{table}}$ reject the null hypothesis

2.3.5 Mann-Whitney Test:

The Mann-Whitney U test is essentially an alternative form of the Wilcoxon Rank-Sum test for independent samples and is completely equivalent. Define the following test statistics for samples 1 and 2 where n_1 is the size of sample 1 and n_2 is the size of sample 2, and R_1 is the adjusted rank sum for sample 1 and R_2 is the adjusted rank sum of sample 2. It doesn't matter which sample is bigger.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1;\ U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2;\ U = \min(U_1, U_2)$$

Hypothesis test:

H_0 : The samples are taken from identical populations

Decision rule:

If the observed value of U is < U_{Critical} then the test is significant (at the α level), i.e. we reject the null hypothesis. Neave, H.R. (1976b)

2.3.6 The wilcoxon Signed ranks test:

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed.

Test procedure:

Let N be the sample size, the number of pairs. Thus, there are a lot a total of $2N$ points.

For $i = 1, \ldots, N$, let $x_{1,i}$ and $x_{2,i}$ denote the measurements.

H_0 : Median difference between the pairs is zero

H_1 : Median difference is not zero.

For $i = 1, \ldots, N$, , calculate $|x_{2,i} - x_{1,i}|$ and $\mathrm{Sgn}(x_{2,i} - x_{1,i})$ where Sgn is the sign function.

Exclude pairs with $|x_{2,i} - x_{1,i}| = 0$. Let N_T be the reduced sample size. Order the remaining N_T pairs from smallest absolute difference to largest absolute difference, $|x_{2,i} - x_{1,i}|$.

Rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let R_i denote the rank.

Calculate the test statistic W : $W = \left| \sum_{i=1}^{N_r} \left[\mathrm{Sgn}(x_{2,i} - X_{1,i}) * R_i \right] \right|$, the absolute value of the sum of the signed ranks

as N_T increases, the sampling distribution of W converges to a normal distribution. Thus, for $N_r \geq 10$, a z-score can be calculated as:

$$Z = \frac{W - 0.5}{\partial_w}, \partial_w = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$ Neave, H.R. (1976b)

Decision rule: If $Z > Z_{critical}$, then reject H_0. For $N_r < 10$, W is compared to a critical value from table

III. RESEARCH FINDINGS AND DISCUSSION

3.1. Distribution of household consumption expenditure data:

The Q-Q plot, normally when the data are normally distributed the dots are along the line but with our data the dots are not following the line means that the household consumption expenditure data are not approximately distributed.

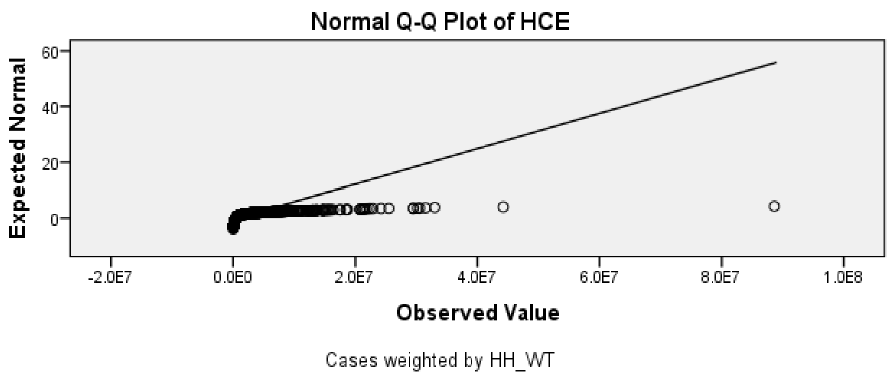

Figure 1 Q-Q plot of HCE

3.2. Normalization of HCE data:

The HCE data is not normally distributed as shown by the result of the test of normality. The parametric test is not allowed to use in testing the difference mean of HCE by domain or components. To use the parametric test we need first of all to normalize the HCE data by using the transformation function recommended in statistics. The log transformation was used to transform the HCE data and histogram was plotted to verify if the log household consumption expenditure (HCE) data are approximately normally distributed. After log transformation we found that the histogram has the approximate shape as a normal curve. This allows us to say that the log HCE data are approximately normally distributed.

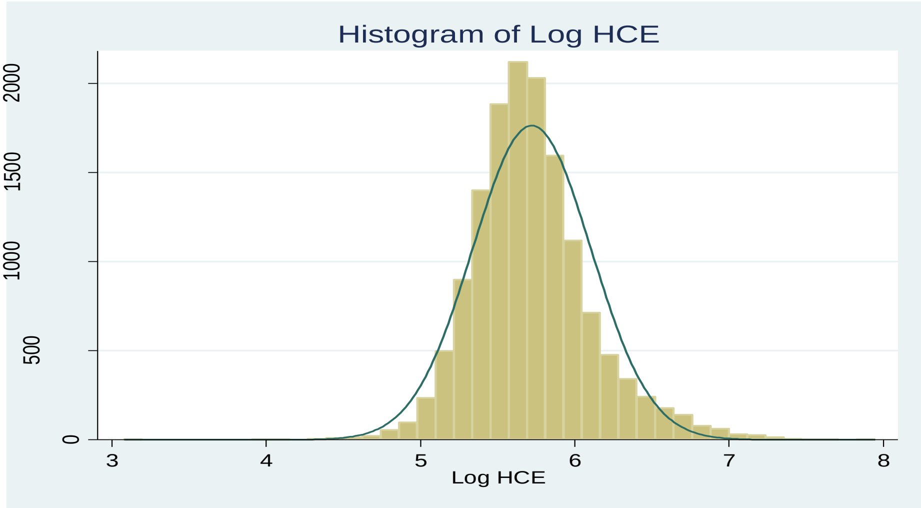

Figure1. Histogram of log HCE

3.3. Estimated HCE total, Mean and Median by Province:

Table 1 Level of household consumption expenditure by province in Rwandan

PROVINCE	Total	Mean	SE	CV	Median
Kigali City	553,845,419,933	2,487,489	99,662	1.47	1,238,955
Southern	308,498,106,448	563,144	12,617	1.39	396,293
Western	358,473,906,108	681,032	18,088	1.54	438,672
Northern	303,041,479,962	738,559	23,638	1.57	444,457
Eastern	428,225,052,754	791,820	16,684	1.22	554,900
Rwanda	**1,952,083,965,204**	**868,629**	**13,214**	**1.82**	**485,603**

Table 1 Shows the estimated total consumption expenditure, mean and median by province and at national level. The survey shows that the HCE was estimated up to 1,952,083,965,204 FRW, with a mean of 868,498 FRW and a median of 485,603FRW. Kigali city shows the highest estimated total HCE, mean, standard error and median.

Table 2 Mean and percentage of HCE by component

Component	Mean	SE	cv	Percentage	N
Food	375,892	4,115	1.3	43	14,308
Non food	256,052	5,210	2.4	30	14,308
Housing	120,321	3,611	3.6	14	14,308
Education	61,245	595	4.6	7	14,308
Health	33,697	2,344	10.0	4	14,308
Alcohol	19,288	2,826	3.7	2	14,308
Rwanda	**144,416**	**1,472**	**3.0**	**100**	**85,848**

Table2 shows the mean of the HCE by component. Food is the component with the highest mean, the second is Non Food, third is the Housing, fourth is the Education, fifth is Health and the last is the alcohol component.

International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 3, Issue 1, pp: (105-122), Month: April 2015 - September 2015, Available at: www.researchpublish.com

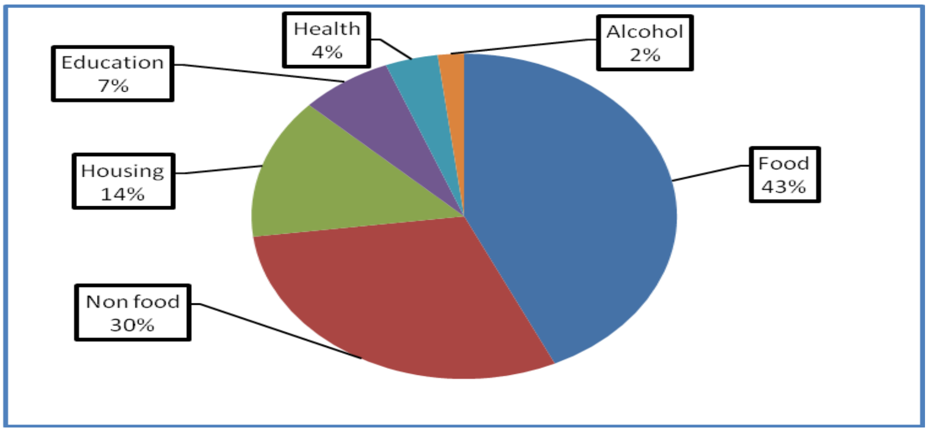

Figure2 Percentage of HCE by component

3.4. Test of the mean differences between HCE components:

3.4.1. Kruskal -wallis test:

As the household consumption expenditure data is not normally distributed, a non- parametric test called kruskal-wallis test was used to test the mean differences at confidence level of 95%. Running the test using STATA software,

Kruskal-Wallis equality-of-populations rank test		
component	Obs	Rank Sum
Housing	14308	6.98e+08
Education	14308	3.85e+08
Health	14308	4.11e+08
Food	14308	1.01e+09
Alcohol	14308	2.78e+08
Non Food	14308	9.01e+08
chi-squared = 51616.174 with 5 d.f.		
probability = 0.0001		
chi-squared with ties = 51942.015 with 5 d.f.		
probability = 0.0001		

We found that the p value for this test is 0.0001. Since the p value is less than 5%, we reject the null hypothesis and we accept the alternative hypothesis. There is a significance difference between means of HCE components. The chi square calculated is also greater than chi square table with n-1 d.f $\chi^2_{\text{Calculated}}(51616.2) > \chi^2_{\text{table}}(2.01)$ this also allow us to reject the null hypothesis "The difference mean of HCE is not statistically significant" and to accept the alternative hypothesis "The difference mean of HCE is statistically significant at significance level of 0.05"

3.4.2. Analysis of variance test (ANOVA):

The ANOVA is a parametric test used to test if the mean differences between more than two samples are statistically significant. As the researcher clearly showed earlier the household consumption expenditure are not normally distributed and the ANOVA is used only for approximately normally distributed data, that is the reason why the HCE was adjusted using the log transformation to normalize the HCE data.

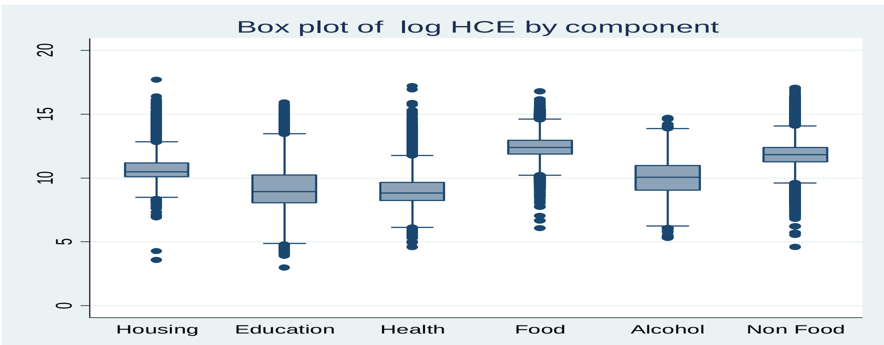

Figure3 Box plot of log HCE by Component

The box plot shows that, the log transformations for household consumption expenditure components are approximately normally distributed. Running the ANOVA using STATA software we found the output in the table below:

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	117446.44	5	23489.288	13861.18	0.0000
Within groups	118688.797	70039	1.69461011		
Total	236135.237	70044	3.37124147		
Bartlett's test for equal variances: chi2(5) = 8.6e+03 Prob>chi2 = 0.000					

The analysis of ANOVA shows that there is a statistical difference between mean of log HCE components since the p-value (prob>F) is less than 0.05 and $F_{Calculated}(13861.18) > F_{table\ \alpha=0.05}(2.21)$ this means that the mean differences between components are statistical significant at α level of 5% and degree of freedom of 5.

3.5. Test of differences in HCE mean between U/R households:

3.5.1. Mann-whitney test:

Running the mann-whitney test in STATA Software its give the output bellow:

Two-sample Wilcoxon rank-sum (Mann-Whitney) test			
Status	obs	rank sum	expected
Urban	2149	21847475	15375021
Rural	12159	80519112	86991566
combined	14308	1.024e+08	1.024e+08
unadjusted variance	3.116e+10		
adjustment for ties	-5.9993591		
adjusted variance	3.116e+10		
Ho: HCE(Status==Urban) = HCE(Status==Rural)			
z = 36.668			
Prob > z = 0.0000			

The mann-whitney test is the test used to test the differences mean between two independent samples. This test was used to test whether the mean differences household consumption expenditure between U/R households are statistically significant. The output of the mann-whitney test shows that the two means of household consumption expenditure are statistical different as well as their p-value is less than 5%. The test shows that the $Z_{Calculated}(36.668) > Z_{table\ \alpha=0.05}(1.96)$ this value also confirm the mean differences HCE between U/R households are statistically significant.

3.5.2. Student t-test:

The student t test is the test used to test the differences mean between two independent samples that are approximately normally distributed. To use this test the HCE was first of all adjusted using log transformation to normalize the data and histogram shows that the log HCE by urban and rural status are approximately normal distributed and this allows us to use the student t test for testing the differences mean HCE between Urban and rural households.

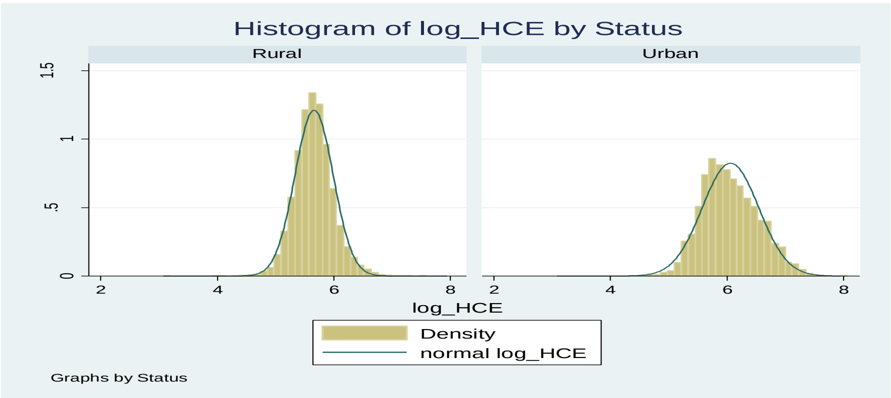

Figure4 Histogram of log HCE by urban and rural

Running the student t test in STATA Software its give the output bellow:

Two-sample t test with equal variances						
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Rural	12159	5.661434	.0029872	.329388	5.655578	5.667289
Urban	2149	6.059769	.0104529	.4845677	6.039271	6.080268
combined	14308	5.721262	.003213	.3843257	5.714964	5.72756
diff		-.3983359	.0083543			-0.7966718
diff = mean(Rural) - mean(Urban)					t = -47.6802	
Ho: diff = 0			degrees of freedom		= 14306	
Ha: diff < 0		Ha: diff != 0			Ha: diff > 0	
Pr(T < t) = 0.0000		Pr(T > t) = 0.0000			Pr(T > t) = 1.0000	

As t calculated is greater than t table, the difference mean between U/R area is statistically significant as well as the absolute $\left|t_{calculated}(-47.6802)\right| > t_{\alpha/2,ddf=14306}(1.645)$. The p-value also shows that the two mean differences are statistically significant at confidence level of 95% as well since it is less than 5%.

3.6. Estimated total, mean and percentage of HCE by sex of head of household:

The tables 4.6.1 indicate that the mean CHE for household headed by male are 967,496 FRW and for those household headed by female is 647,761 FRW. The table7 shows that the household headed by male consume 79.5% of the total household consumption expenditure while the household headed by female consume 20.5% of the total household consumption expenditure.

Table3. Estimated total, mean and percentage of HCE by sex of head of household

Sex of HH	Mean HCE	SE of HCE	Percentage of HCE
Male	967,494	16,740	79.5
Female	647,761	18,742	20.5

3.7. Test of the mean differences between household headed by male and female

3.7.1. The mann-whitney test:

Running the mann-whitney test in STATA Software, its gives the output below:

Two-sample Wilcoxon rank-sum (Mann-Whitney) test			
Gender	obs	rank sum	expected
Male	10330	79052073	73905985
Female	3978	23314513	28460601
combined	14308	1.024e+08	1.024e+08
unadjusted variance		4.900e+10	
adjustment for ties		-9.4348602	
adjusted variance		4.900e+10	
Ho: HCE(Gender==Male) = HCE(Gender==Female)			
z = 23.248			
Prob > z = 0.0000			

The mann-whitney test is the test used to test the difference mean between two independent samples. This test was used to test whether the difference means household consumption expenditure between households headed by male and households headed by female are statistically different. The output of the mann-whitney test shows that the difference in mean household consumption expenditure between household headed by male and household headed by female are statistically significant as well as their p-value is less than 5%. This means that the households headed by male spend more than households headed by female. The test show that the $Z_{Calculated}(23.248) > Z_{table}\ \alpha = 0.05(1.96)$ this value also confirm the differences in mean HCE between household headed by male and female are statistically significant.

3.7.2. Student t test:

The student t test is the test used to test the differences mean between two independent samples approximate normal distributed. To use this test, the HCE was first of all transformed using log transformation to normalize the data and histogram shows that the log HCE are approximately normally distributed and this allows us to use the student t test for testing the differences mean HCE between household headed by male and household headed by female.

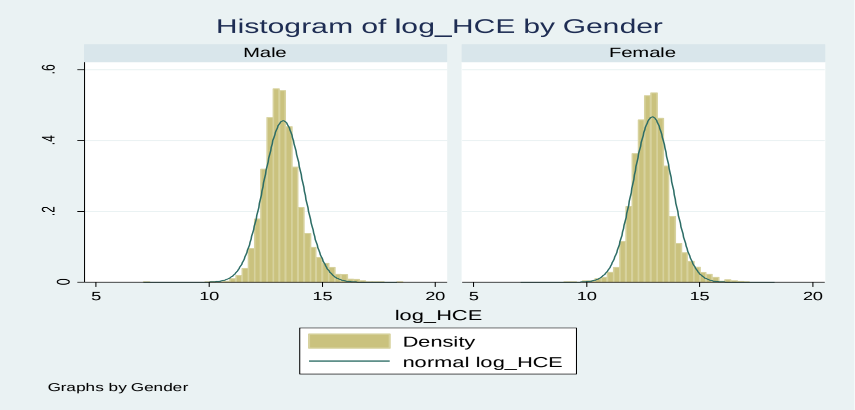

Figure 5 Histogram of log HCE by sex of HH

The shape of histogram for log HCE by sex of HH has approximately a shape of normal curve. This gives us the basis to confirm that the log HCE is approximately normally distributed therefore we are allowed to use the parametric test for testing the equality of two mean between household headed by male and female. Running the student t test gives the output below:

Two-sample t test with equal variances					
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
Male	10330	13.27511	.0086122	.8753138	13.25822 13.29199
Female	3978	12.91034	.0135582	.855135	12.88376 12.93692
combined	14308	13.17369	.0073982	.8849426	13.15919 13.18819
diff		.3647643	.0162294		.3329526 .396576
diff = mean(Male) - mean(Female)					t = 22.4756
Ho: diff = 0		degrees of freedom			= 14306
Ha: diff < 0		Ha: diff != 0			Ha: diff > 0
Pr(T < t) = 1.0000		Pr(T > t) = 0.0000			Pr(T > t) = 0.0000

The table above shows that the mean differences is statistically significant as well as the p value is less than 5% and the $t_{\text{calculated}}(22.46) > t_{\alpha = 0.05 \ \ ddf = 14306}(1.96)$.

3.8. Mean adult equivalent size and mean household composition size:

Table 4 Mean adult equivalent size and mean household composition size

Variable	Obs	Mean	SE
Adult equivalent HH size	14308	4.309	0.017
HH composition size	14308	4.780	0.018

Table4 shows that the mean household adult equivalent size is 4.3 and the mean household composition size is 4.8.

3.9. Test of mean differences between adult equivalent and household composition size:

3.9.1. Wilcoxon signed rank test:

This test is used for testing the mean differences for paired samples for non parametric test. Running the test of wilcoxon signed rank test, it gives the output where the wilcoxon signed rank test show that the differences mean between adult equivalent household size and household composition size is statistically significant at significance level of 0.05 as well as Z calculated is greater than z table.

Wilcoxon signed-rank test			
sign	Obs	Sum ranks	expected
positive	12786	1.01E+08	51043978
negative	776	1439958.5	51043978
zero	746	278631	278631
all	14308	1.02E+08	1.02E+08
unadjusted variance 2.441e+11			
adjustment for ties -10966730			
adjustment for zeros -34666340			
adjusted variance 2.441e+11			
Ho: HC = AE			
z = 100.405			
Prob > z = 0.0000			

3.9.2. Student t test for paired sample:

Running the test of t test paired sample test, it gives the output where the t test paired sample test shows that the mean difference between AE and HC size is statistically significant as well as the zero difference does not lie in the interval of 95%, $t_{calculated}(153.45) > t_{table\ \alpha=0.05}(1.96)$ and their p value also is less than 5%.

Paired t test		
Variable Obs Mean	Std. Err. Std. Dev.	[95% Conf. Interval]
HC 14308 4.780403	.0182171 2.179061	4.744695 4.816111
AE 14308 4.309808	.0171541 2.051908	4.276184 4.343433
diff 14308 .4705941	.0030667 .3668324	.4645828 .4766053
mean(diff) = mean(HC - AE)		t = 153.4505
Ho: mean(diff) = 0	degrees of freedom	14307
Ha: mean(diff) < 0 Ha:	mean(diff) != 0	Ha: mean(diff) > 0
Pr(T < t) = 1.0000 Pr(T > t) = 0.0000	Pr(T > t) = 0.0000	

IV. CONCLUSION AND RECOMMENDATIONS

4.1 Conclusion:

The research has shown that, the household consumption expenditure data is not normally distributed and the non-parametric analysis is better statistical methods to use for deep analysis of HCE which doesn't take any consideration of the distribution of the data during the analysis.

The research shows that, the annual estimated level of household consumption expenditure in Rwanda for six components (Food, Non Food, Education, Health, Housing and Alcohol) is 1,952,083,965,204FRW and the annual mean household consumption expenditure is 866,498 FRW.

The differences in mean Household annual consumption expenditure between household consumption expenditure components are statistically significant and households spent their money mostly on food with 43%, followed by non-food with 30%, Housing with 14%, Education with 7%, Health with 4% and the last is alcohol with 2%.

The research shows that, 65% of the Annual estimated total household consumption expenditure is consumed by rural households and 35% by urban households. Regarding the annual estimated mean household consumption expenditure, it was found that the urban households consume more than rural households with respectively estimated annual mean of 2,080,416 FRW and 657,261 FRW and it was proven that the two mean difference is statistical significant at confidence level of 95%.

The research found out that, the household headed by males spend much more (79.5%) than household headed by females (20.5%) with respectively annual estimated mean of 967,494 FRW and 647,761 FRW and the two mean difference was found statistically significant at significance level of 0.05. The research found that the mean of Adult

Equivalent household size is 4.3 and mean household composition size is 4.8. The test of hypothesis confirmed that the two mean difference is statistically significant at significance level of 95%.

4.2 Recommendations and Suggestions:

As it was statistically proven that, the data for household consumption expenditure are not approximately normally distributed any statistical analysis should take into consideration the following points:

1. To test the normality of the data before any statistical analysis
2. To use the non-parametric test for any statistical analysis
3. To transform the HCE data to normalize the HCE data before using the parametric test
4. The mean household consumption expenditure must be used carefully since the data are not normally distributed and therefore instead of using the mean it is imperative to replace it by the median.
5. The mean adult equivalent household size must be used when dealing with the household consumption expenditure and mean household composition size must be used when dealing with counting people in the house.
6. The government of Rwanda must invest in production of food as well as most of household consumption expenditure is food component.
7. As there is no methodology related to Rwanda for household consumption expenditure, we encourage people to continue working on it and to improve the methodology of collecting data for household consumption expenditure.
8. As the dataset is too large, the variable specific to household consumption expenditure must be separate with other variables to facilitate the users to find easily the data related to household consumption expenditure in the datasets
9. The household consumption expenditure for durable goods was excluded because the depreciation rate for durable goods was not available in the household living condition survey (EICV3) dataset. Therefore, we recommend the inclusion of the depreciation rate in the datasets.

REFERENCES

[1] Cochran, W. G. 1950. "Estimation of Bacterial Densities by Means of the 'Most Probable Number.' " Biometrics, 6: 105.

[2] Cochran, W. G. 1977. Sampling Techniques, 3rd ed. New York: Wiley.

[3] Bailey, A. D. 1981. Statistical Auditing. New York: Harcourt Brace Jovanovich.

[4] Richard L. Scheaffer, William Mendenhall III, R. Lyman Ott and Kenneth Gerow .1996. Elementary survey sampling, Seven Edition

[5] De Landsheere, G. (1982). Introduction to research in education, Paris.

[6] Bailey, N. T. J. 1951. "On Estimating the Size of Mobile Populations from Recaptive Data."Biometrika, 38: 292–306.

[7] Pearson, A. V., and Hartley, H. O. (1972). Biometrica Tables for Statisticians, Vol 2, Cambridge, England, Cambridge University Press

[8] Using Random Digit Dialing." American Statistical Association Proceedings of the Section on Survey Research Methods, pp. 239–243. Alexandria, VA: American Statistical Association.

[9] Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions", The American Statistician, 52(2), 119-126.

[10] Berenson M.L. and Levine D.M. (1996) Basic Business Statistics,Prentice-Hall, Englewood Cliffs,New Jersey.

[11] Bhattacharyya, G.K., and R.A. Johnson, (1997). Statistical Concepts and Methods, John wiley and Sons, New York

[12] Booth, J. G., Butler, R. W., and Hall, P. 1994. "Bootstrap Methods for Finite Populations."Journal of the American Statistical Association 89(428): 1282–1289.

[13] Bradburn, N. N., Rips, L. J., and Shevell, S. K. 1987. "Answering Autobiographical Questions:The Impact of Memory and Inference on Surveys." Science, 8 (April 10): 157–161.

[14] Bryson, M. C. 1976. "The Literary Digest Poll: Making of a Statistical Myth." American Statistician,30(4): 184–185.

[15] Campbell, C., and Joiner, B. 1973. "How to Get the Answer without Being Sure You Asked theQuestion." American Statistician, 27: 229–231.

[16] Chapman, D. G. 1952. "Inverse, Multiple and Sequential Sample Censuses." Biometrics, 8:286–306.

[17] Deming, W. E. 1960. Sample Design in Business Research. New York: Wiley.

[18] ABS. 2000. Household expenditure survey: User guide 1998-99. Canberra, Australian Bureau of Statistics.

[19] 1995. A provisional framework for household income, consumption, saving and wealth.Canberra, Australian Bureau of Statistics.

[20] Astin, J. 1999. "The European Union Harmonized Indices of Consumer Prices (HICP)",Proceedings of the Ottawa Group Fifth Meeting, Gudnason, R., and Gylfadottir, T.(eds.), held at Statistics Iceland, Reykjavik, Iceland, 25-27 Aug. Text available athttp://www.statcan.ca/secure/english/ottawa_group/, also published in Statistical Journal of the United Nations ECE 16 (1999), pp. 123-135.

[21] BLS. 2001. Consumer expenditure survey home page (http://stats.bls.gov/csxhome.htm).Washington, DC, Bureau of Labor Statistics.

[22] 1995. Household income and expenditure statistics, No. 4 (Geneva).

[23] Seventh ICLS (1949), in The Seventh International Conference of Labour Statisticians, Geneva, pp. 54-58.

[24] Johnson, M.; McKay, A.D.; Round, J.I. 1990. Income and expenditure in a system of household accounts: Concepts and estimation, Social Dimensions of Adjustment in Sub-Saharan Africa (Washington, DC, World Bank, Working Paper No. 10).

[25] STATCAN. 2000. 1999 Survey of household spending: User's guide (Ottawa, Statistics Canada).

[26] STATIN. 1999. Jamaica survey of living conditions 1998 (Kingston, Planning Institute of Jamaica and Statistical Institute of Jamaica).

[27] Tanner, S. 1998. "How much do consumers spend? Comparing the FES and national accounts," in How reliable is the family expenditure survey? Trends in incomes and expenditures over time, Banks, J. and Johnson, P. (eds.) (The Institute for Fiscal Studies, London).

[28] UN. 1989. National Household Survey Capability Programme, Household income and expenditure surveys: A technical study, United Nations, New York.

[29] 1977. Provisional guidelines on statistics of the distribution of income, consumption and accumulation of households. Studies in Methods, Series M, No. 61 (New York. United Nations ST/ESA/STAT/SER.M/72).

[30] 1964. Handbook of household surveys: A practical guide for inquiries on levels of living, Studies in Methods, Series F, No. 10, Chapters 7 and 8. (New York).

[31] ICLS-R-2003-06-0049-1-EN.Doc/v1 79

[32] UNDTCD. 1989. National household survey capability programme, Household income and expenditure surveys: A technical study (New York, United Nations Department of Technical Cooperation for Development, and Statistical Office).