Tunisia comments on the UNCTAD manual for the production of Statistics on the Information Economy

Mr. Mohamed Amine JALLOULI
Senior Engineer / Head of Unit of ICT statistics
jallouli.med.amine@gmail.com

Republic of Tunisia
Ministry of Development and International Cooperation
National Institute of Statistics

International Seminar on Information and Telecommunication technologies
Statistics, Seoul/Republic of South Korea, 19-21 July 2010

Outlier

1. Introduction

2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation

3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response

4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Outlier

1. Introduction
2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation
3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response
4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Outlier

1. Introduction

2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation

3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response

4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Outlier

1. Introduction
2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation
3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response
4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Introduction

- Tunisian business ICT access and use survey has been realized with the UNCTAD technical assistance and an expert from INSEE-France.
- This survey is the fruit of cooperation between the National Institute of Statistics (technical aspect) and the Ministry of Telecommunication Technologies (financial aspect).
- Data collection process has been realized in 3 months.

Outlier

1. Introduction

2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation

3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response

4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Population frame

Frame. National Business Register (2008)

Sources. It is the result of matching and activity codification operations on two administrative files: Tax Office and National Fund of Social Security.

Variables. identification variables (ID, enterprise name, address) + stratification variables (economic activity, class of salaries, class of turnover) + number of salaries

Size. About 550,000 enterprises

Outlier

1 Introduction

2 Sampling
- Population frame
- Target population & sample design
- Sample allocation

3 Data editing
- Treatment of internal inconsistencies & errors
- Treatment of missing data
- Treatment of non-response

4 Measurement of ICT impact indicators
- Existing works
- Recommendations

Target population & sample design

Population frame. Tunisian Business Register (2008)

Target population. Enterprises not operating in

- agricultural activities
- activities of membership organisations

and having more than 6 salaries.

Stratification variables. Economic activity and class of salaries.

- **Employment breakdown:** [6,9], [10,19], [20,49], [50,99], [100,199], 200+
- **Activities breakdown:** 11-37, 40-41, 45, 50-52, 55, 60-63, 65-67, 70-75, 80, 85, 90, 92, 93

Target population size. 17,418

Sample design. Stratified Simple Random Sampling without replacement.

Outlier

1. Introduction
2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation
3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response
4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Sample Allocation Algorithm

Neymann allocation algorithm

prerequisites: availability of good quantitative auxiliary variable for all individuals in the population frame ($y = employment$).

- Choose a sampling error rate[a] c^* and determine its correspondent global sample size n_{opt}

$$\begin{array}{rrl} & n_{opt} = & \frac{N^2 S_y^2}{(t_y c^*)^2 + N S_y^2} \\ \text{where} & c^* : & \text{sampling error} \\ & t_y : & \text{total of y} \\ & N : & \text{population frame size} \end{array} \qquad (1)$$

[a] The sampling error rate is chosen to be the CV.

Sample Allocation Algorithm (Cnt'd)

Algorithm (Cnt'd)

- For each stratum h, $n_h = n_{opt}\frac{N_h S_{yh}}{\sum_{i=1}^{H} N_i S_{yi}}$
- **If** $n_h > N_h$ **then** $n_h = N_h$
- **If** $n_h < n_h^*$ **then** $n_h = n_h^*$ where $n_h^* = \frac{N_h^2 S_{yh}^2}{(t_{yh} c_h^*)^2 + N_h S_{yh}^2}$
- **If** $n_h < 3$ **then** $n_h = min\{3, n_h^*\}$

Advantages

- This allocation algorithm is easily implemented.
- All strata are representative.
- This allocation algorithm minimizes of sampling variance.

Sample Allocation Algorithm (Cnt'd)

Disadvanges

- It requires the availability good quantitative auxiliary variable in the population frame.

Recommendations

- This sample allocation algorithm is described in the UNCTAD manual for the production of Statistics on Information Economy (P. 79, para. 208). In order to be make this paragrah easier for understanding, I recommend a deeper emphasis on the mathematical aspect.
- For being more pratical, I recommend an R package for sample allocation: **bethel**. Bethel alg. (1989) allows to determine total sample size and allocation of units in strata, so to minimize costs under the constraints defined in terms of estimates precision levels in the multivariate case (more than one auxiliary variable, for example: number of employees and turnover). `http://cran.r-project.org/web/packages/bethel/`

Tunisia sample allocation

- Eurostat recommendations for members states specify a maximum coefficient of variation for overall proportions of $c^* = 2\%$, $c_g^* = 5\%$ at the level of sub-groups (one dimensional-breakdown) and $c_h^* = 10\%$ at the level of strata (two-dimensional breakdown). (see P. 78 para. 208 and Box 13).
- Due to budget constraints, $c^* = 6.02\%$ and $c_h^* = 10\%$. No constraints are taken at the level of sub-groups.

	error	optimal size
1	0.01	14949.15
2	0.02	10506.16
3	0.03	7065.64
4	0.04	4883.07
5	0.05	3527.93
6	0.06	2661.44
7	0.07	2084.71

Population frame size: 17,418

Sample size: 2,618

Sampling rate: 15%

Sampling error: 6.2%

Outlier

1. Introduction
2. Sampling
   - Population frame
   - Target population & sample design
   - Sample allocation
3. Data editing
   - Treatment of internal inconsistencies & errors
   - Treatment of missing data
   - Treatment of non-response
4. Measurement of ICT impact indicators
   - Existing works
   - Recommendations

Treatment of internal inconsistencies & errors

- It consists in establishing a set of consistency rules for being checked at the level of individuals. Examples:

  Rule 1 If a firm does not use LAN, then it does not use Intranet.

  Rule 2 If a firm does not use Internet, then it does not place order over the Internet.

- This set of internal consistency rules depends on the country questionnaire.
- This set of internal consistency rules should be integrated in data collection instrument in order to reduce the data processing.
- When an individual does not verify a rule, one of these treatment should occur:
  1. correcting the erroneous value by a plausible value.
  2. putting the erroneous value to missing data.

Outlier

1 Introduction

2 Sampling
- Population frame
- Target population & sample design
- Sample allocation

3 Data editing
- Treatment of internal inconsistencies & errors
- Treatment of missing data
- Treatment of non-response

4 Measurement of ICT impact indicators
- Existing works
- Recommendations

Imputation

Existing methods

Deterministic methods. If the imputation method reproduces always the same results for a multiple executions.

1. mean (at level of strata or groups)
2. ratio (at level of strata or groups)
3. regression (at level of groups)
4. cold-deck (it is the use of other sources (other surveys, previous realizations of the survey, business register) for imputation)

Probabilistic methods. If the imputation method could reproduce different results for each execution.

1. hot-deck (at level of strata or groups)
2. k-nearest neighbor

Imputation methods used by the Tunisian ICT survey

Imputation methods used by the Tunisian ICT survey

Cold-deck imputation. Number of employees and the turnover based on the National Business Register and the Annual Survey on Economic Activities.

Hot-deck imputation. All qualitative variables and all quantitative variables (exception for the Number of employees and the turnover).

Imputation under constraints

- When imputing constrained quantitative variables, it is recommended to convert them to ratios and then impute them.
  Example of constraint: #internet usr $\leq$ #computer usr$\leq$ #employees
  1st imputation: #computer usr becomes $\frac{\text{\#computer usr}}{\text{\#employees}}$
  2nd imputation: #internet usr becomes $\frac{\text{\#internet usr}}{\text{\#computer usr}}$

Recommendations relative to imputation

Recommendations

1. Avoid to impute big enterprises (200+) and specially public enterprises.
2. These enterprises should be interviewed at the beginning of the data collection process in order to recontact them during this process if some internal inconsistencies or missing data occur.
3. **R packages for imputations:** rrp and yaImpute.
   `http://cran.r-project.org/web/packages/rrp/`
   `http://cran.r-project.org/web/packages/yaImpute/`

Outlier

1 Introduction

2 Sampling
- Population frame
- Target population & sample design
- Sample allocation

3 Data editing
- Treatment of internal inconsistencies & errors
- Treatment of missing data
- Treatment of non-response

4 Measurement of ICT impact indicators
- Existing works
- Recommendations

Re-weighting

Re-weighting methods

- Correction based on Response Homogeneity Group (RHG)
  $w_{RHG_i} = \frac{w_i}{r_{RHG_g}}$ where enterprise $i \in$ group g.
  Groups can be the class of employees.
- Correction based on individual probabilities of responding
  $w_{logit_i} = \frac{w_i}{p_{logit_i}}$ for each enterprise i

$$\begin{aligned} P(response_i = 1|Employment = x_i) &= logit(x_i) \\ reponse &= \begin{cases} 1 & \text{if} \quad \text{the enterprise answered} \\ 0 & \text{if} \quad \text{the entreprise did not answer} \end{cases} \end{aligned} \tag{2}$$

- Calibration

$$\begin{array}{lll} (P) : min_{w_{calibration\,i}} \sum_{i=1}^{r} d(w_i, w_{calibration\,i}) & & \\ \text{u.c.} \quad \sum_{i=1}^{r} w_{calibration\,i} = N & & \\ \qquad \sum_{i=1}^{r} X_i w_{calibration\,i} = X_{frame} & & \\ \text{where} \quad r & : & \text{Number of responses.} \\ \qquad d & : & \text{a distance} \\ \qquad N & : & \text{sampling frame size.} \end{array} \tag{3}$$

Recommendations

- For countries having a good auxiliary variable in their sampling frame, it is recommended to use the calibration method for correcting non-response errors.
- The re-weighting method based on RHG has been described by UNCTAD manual for the Production of Statistics on Information Society (P. 85, Box 16). This method works well if and only if these conditions are checked [Lynn 2005]:

  cond 1. Response rates vary over the groups.

  cond 2. The values of survey estimates (e.g. means, proportions, regression coefficients, etc) vary over the groups.

  cond 3. The values of survey estimates must be similar for both respondents and non-respondents within each group.

  cond 4. classes should not be too small (min of 30 or 50 sample units).

  It is recommended to mention these conditions in the UNCTAD manual.

Lynn, P. (2005), *"Weighting"*, in (ed. K. Kempf-Leonard) Encyclopedia of Social Measurement, Academic Press, pp. 967-973.

Outlier

1 Introduction

2 Sampling
- Population frame
- Target population & sample design
- Sample allocation

3 Data editing
- Treatment of internal inconsistencies & errors
- Treatment of missing data
- Treatment of non-response

4 Measurement of ICT impact indicators
- Existing works
- Recommendations

Thailand experience

Cobb-Douglass production function

$$\begin{array}{lrcl} \multicolumn{4}{c}{ln(P_i) = \lambda + \alpha ln(L_i) + \beta ln(K_i) + \gamma ICT_i + \theta ln(M_i)} \\ \text{where} & P & : & \text{Production (approximated by turnover)} \\ & L & : & \text{employment} \\ & K & : & \text{capital (approximated by immobilizations)} \\ & ICT & : & \text{composite variable indicating ICT use} \\ & M & : & \text{Spending on materials} \end{array} \qquad (4)$$

"Measuring the impact of ICT use in business: the case of manufacturing in Thailand", Prepared jointly by the UNCTAD secretariat and the Thailand National Statistical Office, pp. 28, 2008.

Outlier

1 Introduction

2 Sampling
- Population frame
- Target population & sample design
- Sample allocation

3 Data editing
- Treatment of internal inconsistencies & errors
- Treatment of missing data
- Treatment of non-response

4 Measurement of ICT impact indicators
- Existing works
- Recommendations

Recommendations

- The UNCTAD model questionnaire (P. 127, Module C: Other information about your business) does not allow the measurement of ICT impact indicators, due to the absence of two variables: spending on materials (M) and capital (K).
  $\Rightarrow$ It is recommended to add two questions in the UNCTAD model questionnaire in order to allow ICT impact indicators measurement.
- Shall the spending concern all materials or ICT materials only?
- Is it better to explain the production by the spending on material and intangible relative to ICTs?

Thank you for your attention.

jallouli.med.amine@gmail.com