A New Version of the Item Count Technique

Eleni Manoli and Tasos C. Christofides

University of Cyprus

2016 20 – 21 October Budapest

1. Introduction

1. Introduction

- Classical survey procedures are not appropriate for the investigation of sensitive items

1. Introduction

- Classical survey procedures are not appropriate for the investigation of sensitive items
- People are reluctant to participate in surveys dealing with matters of privacy

1. Introduction

- Classical survey procedures are not appropriate for the investigation of sensitive items
- People are reluctant to participate in surveys dealing with matters of privacy
- People participating often provide untruthful responses

Stigmatizing Issues

Stigmatizing Issues

- Sexual Behavior

Stigmatizing Issues

- Sexual Behavior
- Drug Abuse

Stigmatizing Issues

- Sexual Behavior
- Drug Abuse
- Political affiliation (in certain societies)

Stigmatizing Issues

- Sexual Behavior
- Drug Abuse
- Political affiliation (in certain societies)
- Engagement in criminal activities

Indirect Questioning Techniques

Indirect Questioning Techniques

- Indirect questioning has been used by psychologists and merchants since the **1950s**

Indirect Questioning Techniques

- Indirect questioning has been used by psychologists and merchants since the **1950s**
- Warner (1965) was the first researcher who published a research paper concerning Indirect Questioning Techniques

Indirect Questioning Techniques

- Indirect questioning has been used by psychologists and merchants since the **1950s**
- Warner (1965) was the first researcher who published a research paper concerning Indirect Questioning Techniques
- Techniques which allow the respondent to provide an answer from which it is not possible for the interviewer to infer whether the specific respondent belongs to the stigmatizing category

Indirect Questioning Techniques

- Indirect questioning has been used by psychologists and merchants since the **1950s**
- Warner (1965) was the first researcher who published a research paper concerning Indirect Questioning Techniques
- Techniques which allow the respondent to provide an answer from which it is not possible for the interviewer to infer whether the specific respondent belongs to the stigmatizing category
- **Protection of Privacy**

Item Count Technique (ICT)

- Raghavarao and Federer (1979), Miller (1984) and Miller et al. (1986)

Item Count Technique (ICT)

- Raghavarao and Federer (1979), Miller (1984) and Miller et al. (1986)
- Estimate the proportion θ of people belonging to the stigmatizing category

Item Count Technique (ICT)

- Raghavarao and Federer (1979), Miller (1984) and Miller et al. (1986)
- Estimate the proportion θ of people belonging to the stigmatizing category
- Two lists are used:
  a) the long item list which includes **G + 1 items**
  (G are non sensitive and one is sensitive)
  and
  b) the short item list which includes the **G non sensitive items**

The Item Count Technique

- In both samples, the respondents should report the total number of items that apply to them without disclosing which ones

The Item Count Technique

- In both samples, the respondents should report the total number of items that apply to them without disclosing which ones
- Let X_i be the number reported by person i from the first sample $(i = 1, \ldots, n_1)$ and let Y_j be the number reported by person j from the second sample $(j = 1, \ldots, n_2)$. Then

$$\hat{\theta} = \bar{X} - \bar{Y}$$

is an estimator of θ

Sample Questionnaires

What follows are sample lists which can be used to estimate the prevalence of illegal doping among professional athletes.

Questionnaire 1		
a/a	*Statement*	*Score*
1	I am on a high protein diet.	
2	My mother was/is allergic to fish.	
3	I make use of illegal doping.	
4	I have taken antibiotics during last year.	
5	I have never been hospitalized.	
6	Before I became a professional athlete I used to take vitamins on a daily basis.	
7	After retirement, I will become a trainer for professional athletes.	
	Total Score:	

Questionnaire 2		
a/a	*Statement*	*Score*
1	I am on a high protein diet.	
2	My mother was/is allergic to fish.	
3	I have taken antibiotics during last year.	
4	I have never been hospitalized.	
5	Before I became a professional athlete I used to take vitamins on a daily basis.	
6	After retirement, I will become a trainer for professional athletes.	
	Total Score:	

- **Disadvantages:**
  a) Only the treatment group gives information on the sensitive characteristic we are investigating

  b) In cases where the interviewee reports the number G + 1 as his/her item score, then the interviewer can be sure that the respondent has the sensitive item

The Item Count Technique

- **Disadvantages:**
  a) Only the treatment group gives information on the sensitive characteristic we are investigating

  b) In cases where the interviewee reports the number G + 1 as his/her item score, then the interviewer can be sure that the respondent has the sensitive item

- **Advantages:**
  a) It can be incorporated into regular questionnaires

  b) It does not require respondents to conduct a randomization experiment

An improved version of the Item Count Technique

- Purpose of the new version: To **protect the privacy** of the participants and to estimate the population proportion θ of people who belong to the sensitive category

An improved version of the Item Count Technique

- Purpose of the new version: To **protect the privacy** of the participants and to estimate the population proportion θ of people who belong to the sensitive category
- How is this achieved?

An improved version of the Item Count Technique

- Purpose of the new version: To **protect the privacy** of the participants and to estimate the population proportion θ of people who belong to the sensitive category

- How is this achieved? The interviewer **cannot infer** from any response that the participant does or does not belong to the stigmatizing category

2. The method

2. The method

- Two *independent* random samples of sizes n_1 and n_2 are drawn with replacement from the population

2. The method

- Two *independent* random samples of sizes n_1 and n_2 are drawn with replacement from the population

- **First Sample:**

2. The method

- Two *independent* random samples of sizes n_1 and n_2 are drawn with replacement from the population

- **First Sample:**

  ✓Every participant is presented with a list of $G + 1$ items of which G are non sensitive and one is sensitive

2. The method

- Two *independent* random samples of sizes n_1 and n_2 are drawn with replacement from the population

- **First Sample:**

  ✓Every participant is presented with a list of $G + 1$ items of which G are non sensitive and one is sensitive

  ✓He/she has to study the items and to count how many of them are applicable to him or her without disclosing which ones

2. The method

- Two *independent* random samples of sizes n_1 and n_2 are drawn with replacement from the population

- **First Sample:**

  ✓Every participant is presented with a list of $G+1$ items of which G are non sensitive and one is sensitive

  ✓He/she has to study the items and to count how many of them are applicable to him or her without disclosing which ones

  ✓If all $G+1$ items are applicable then he or she must report the number 1 and if zero items are applicable then the participant must report the number G. In any other case he/she must report the exact number of the applicable items

- **Second Sample:**

- **Second Sample:**

  ✓Every participant is presented with a list of the G non sensitive items that are included in the first list

- **Second Sample:**

  ✓Every participant is presented with a list of the G non sensitive items that are included in the first list

  ✓He/she has to study the items and to count how many of them are applicable to him or her without disclosing which ones

- **Second Sample:**

  ✓Every participant is presented with a list of the G non sensitive items that are included in the first list

  ✓He/she has to study the items and to count how many of them are applicable to him or her without disclosing which ones

  ✓He/she must report the exact number of applicable items

2. The method

2. The method

- Let q_k be the probability that an individual from the first sample reports the number k, $k = 1, \ldots, G$

2. The method

- Let q_k be the probability that an individual from the first sample reports the number k, $k = 1, \ldots, G$
- Let p_k be the probability that exactly k of the non stigmatizing items are applicable to an individual, $k = 0, 1, \ldots, G$

2. The method

- Let q_k be the probability that an individual from the first sample reports the number k, $k = 1, \ldots, G$
- Let p_k be the probability that exactly k of the non stigmatizing items are applicable to an individual, $k = 0, 1, \ldots, G$
- Then

$$
\begin{aligned}
q_1 &= p_0\theta + p_1(1-\theta) + p_G\theta, && (1)\\
q_k &= p_{k-1}\theta + p_k(1-\theta), \textit{for} \quad k = 2, \ldots, G-1, && (2)\\
q_G &= p_{G-1}\theta + p_G(1-\theta) + p_0(1-\theta) && (3)
\end{aligned}
$$

- Let $n_k^{(1)}$ be the number of individuals from the first sample reporting the number k, $k = 1, \ldots, G$

- Let $n_k^{(1)}$ be the number of individuals from the first sample reporting the number k, $k = 1, \ldots, G$
- Then the vector $\left(n_1^{(1)}, \ldots, n_G^{(1)}\right)$ follows the multinomial distribution with parameters $n_1, q_1, \ldots, q_G$

- Let $n_k^{(1)}$ be the number of individuals from the first sample reporting the number k, $k = 1, \ldots, G$
- Then the vector $\left(n_1^{(1)}, \ldots, n_G^{(1)}\right)$ follows the multinomial distribution with parameters $n_1, q_1, \ldots, q_G$
- Let $n_k^{(2)}$ be the number of individuals from the second sample reporting the number k, $k = 0, \ldots, G$

- Let $n_k^{(1)}$ be the number of individuals from the first sample reporting the number k, $k = 1, \ldots, G$
- Then the vector $\left(n_1^{(1)}, \ldots, n_G^{(1)}\right)$ follows the multinomial distribution with parameters $n_1, q_1, \ldots, q_G$
- Let $n_k^{(2)}$ be the number of individuals from the second sample reporting the number k, $k = 0, \ldots, G$
- Then the vector $\left(n_0^{(2)}, \ldots, n_G^{(2)}\right)$ follows the multinomial distribution with parameters $n_2, p_0, \ldots, p_G$

Consider the following quantities:

$$C_1 = \frac{1}{n_1} \sum_{k=1}^{G} k n_k^{(1)}$$

$$C_2 = \frac{1}{n_2} \sum_{k=1}^{G} k n_k^{(2)}$$

Consider the following quantities:

$$C_1 = \frac{1}{n_1}\sum_{k=1}^{G} k n_k^{(1)}$$

$$C_2 = \frac{1}{n_2}\sum_{k=1}^{G} k n_k^{(2)}$$

Using properties of the multinomial distribution it can be shown that

$$\theta = \frac{E(C_1 - C_2) - G p_0}{1 - G(p_G + p_0)}, \tag{4}$$

assuming that $G(p_G + p_0) \neq 1$

- Equation (4) suggests as an estimator for the parameter θ the quantity

$$\hat{\theta} = \frac{C_1 - C_2 - G\hat{p}_0}{1 - G(\hat{p}_0 + \hat{p}_G)}, \tag{5}$$

where $\hat{p}_G = \frac{n_G^{(2)}}{n_2}$, $\hat{p}_0 = \frac{n_0^{(2)}}{n_2}$ and assuming that $(\hat{p}_0 + \hat{p}_G) \neq \frac{1}{G}$

- Equation (4) suggests as an estimator for the parameter θ the quantity

$$\hat{\theta} = \frac{C_1 - C_2 - G\hat{p}_0}{1 - G(\hat{p}_0 + \hat{p}_G)}, \tag{5}$$

where $\hat{p}_G = \frac{n_G^{(2)}}{n_2}$, $\hat{p}_0 = \frac{n_0^{(2)}}{n_2}$ and assuming that $(\hat{p}_0 + \hat{p}_G) \neq \frac{1}{G}$

- Ratio estimator

- Equation (4) suggests as an estimator for the parameter θ the quantity

$$\hat{\theta} = \frac{C_1 - C_2 - G\hat{p}_0}{1 - G(\hat{p}_0 + \hat{p}_G)}, \tag{5}$$

where $\hat{p}_G = \frac{n_G^{(2)}}{n_2}$, $\hat{p}_0 = \frac{n_0^{(2)}}{n_2}$ and assuming that $(\hat{p}_0 + \hat{p}_G) \neq \frac{1}{G}$

- Ratio estimator

- Biased estimator BUT asymptotically unbiased

The dominating term of its bias is given by the following result:

Theorem

For the estimator $\hat{\theta}$ given by (5), we have that

$$E(\hat{\theta} - \theta) \approx \frac{G(p_0 + p_G)[\theta G(1 - p_0 - p_G) + A - G(1 - p_0)]}{n_2[1 - G(p_0 + p_G)]^2},$$

where $A = \sum_{k=1}^{G} k p_k$

Theorem

$$Var\left(\hat{\theta}\right) \approx \frac{1}{[1-G(p_0+p_G)]^2} \; x \left(\frac{Var(W_1)}{n_1} + \frac{Var(W_2)}{n_2} + \frac{G^2 p_0(1-p_0) - 2Gp_0A}{n_2}\right),$$

where W_1 and W_2 are random variables with probability mass function

$$P(W_1 = k) = q_k, \quad k = 1, \ldots, G$$

and

$$P(W_2 = k) = p_k, \quad k = 0, 1, \ldots, G$$

respectively

- Obviously the approximate variance is **unknown**

- Obviously the approximate variance is **unknown**
- Thus, an estimator of the variance can be given as:

- Obviously the approximate variance is **unknown**
- Thus, an estimator of the variance can be given as:

$$\begin{aligned} \hat{Var}(\hat{\theta}) &= \frac{1}{[1 - G(\hat{p}_0 + \hat{p}_G)]^2} \\ x &\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + G^2 \frac{\hat{p}_0(1 - \hat{p}_0)}{n_2} - \frac{2GC_2\hat{p}_0}{n_2} \right), \end{aligned}$$

where S_1^2 and S_2^2 are the sample variances of the numbers reported by the first and second sample respectively

3. Simulation Results

3. Simulation Results

- **Assumptions:** $n_1 = 500,\ n_2 = 800,\ G = 5,\ \theta = 0.25$
  Population prevalences for the five non sensitive items: $0.10, 0.15, 0.20, 0.25$ and 0.30 respectively

3. Simulation Results

- **Assumptions:** $n_1 = 500,\ n_2 = 800,\ G = 5,\ \theta = 0.25$
  Population prevalences for the five non sensitive items:
  $0.10, 0.15, 0.20, 0.25$ and 0.30 respectively

- **Table 1** provides some revealing results for the revised version and **Table 2** provides some revealing results for the original ICT

Table 1			
Point Estimate	Left Limit	Right Limit	Length of Conf.Interval
0.2232500	0.11923745	0.3272625	0.2080251
0.2945000	0.18995337	0.3990466	0.2090933
0.2945000	0.18856967	0.4004303	0.2118607
0.2015000	0.09792109	0.3050789	0.2071578
0.3090000	0.20162572	0.4163743	0.2147486
0.2166038	0.10961250	0.3235951	0.2139826

Table 2			
Point Estimate	Left Limit	Right Limit	Length of Conf.Interval
0.22325	0.11923745	0.3272625	0.2080251
0.29450	0.18995337	0.3990466	0.2090933
0.29450	0.18856967	0.4004303	0.2118607
0.20150	0.09792109	0.3050789	0.2071578
0.30900	0.20162572	0.4163743	0.2147486
0.21525	0.10892742	0.3215726	0.2126452

- According to Table 1 and Table 2 the mean value of the point estimates for θ is: 0.2549394 in the `improved version` and 0.2548125 in the `classical ICT`

- According to Table 1 and Table 2 the mean value of the point estimates for θ is: 0.2549394 in the `improved version` and 0.2548125 in the `classical ICT`
- Also, in the `improved version`, for the generated confidence intervals, the minimum is 0.1989, the maximum length is 0.2210 and the mean value of the length is 0.2111

- According to Table 1 and Table 2 the mean value of the point estimates for θ is: 0.2549394 in the `improved version` and 0.2548125 in the `classical ICT`
- Also, in the `improved version`, for the generated confidence intervals, the minimum is 0.1989, the maximum length is 0.2210 and the mean value of the length is 0.2111
- The corresponding values for the `classical ICT` are: 0.1989, 0.2210 and 0.2109

- For the revised method **Table 3** gives actual coverage proportions in the case of the repetition of the procedure one hundred times

- For the revised method **Table 3** gives actual coverage proportions in the case of the repetition of the procedure one hundred times

- **Table 4** provides summary statistics of the coverage proportions of Table 3

- For the revised method **Table 3** gives actual coverage proportions in the case of the repetition of the procedure one hundred times

- **Table 4** provides summary statistics of the coverage proportions of Table 3

- The corresponding results for the classical ICT are given in **Table 5** and **Table 6**

Table 3

Actual Coverage Proportions

0.94	0.97	0.91	0.96	0.94	0.94	0.95	0.93	0.96	0.99	0.96	0.94
0.93	0.91	0.90	0.96	0.88	0.90	0.98	0.96	0.94	0.98	0.97	0.97
0.94	0.94	0.92	0.95	0.98	0.89	0.97	0.93	0.99	0.94	0.97	0.95
0.95	0.94	0.93	0.97	0.93	0.98	0.97	0.95	0.94	0.94	0.96	0.95
0.94	0.94	0.93	0.95	0.98	0.95	0.96	0.98	0.97	0.94	0.97	0.92
0.92	0.91	0.95	0.92	0.90	0.93	0.89	0.94	0.95	0.94	0.97	0.94
0.99	0.97	0.96	0.94	0.95	0.97	0.96	0.96	0.94	0.93	0.90	0.95
0.97	0.94	0.94	0.97	0.97	0.94	0.94	0.92	0.92	0.98	0.97	0.94
0.96	0.97	0.98	0.91								

Table 4

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8800	0.9375	0.9500	0.9468	0.9700	0.9900

Table 5
Actual Coverage Proportions
0.94 0.97 0.91 0.96 0.93 0.94 0.95 0.93 0.96 0.99 0.96 0.94
0.93 0.91 0.89 0.96 0.88 0.90 0.98 0.96 0.94 0.98 0.97 0.97
0.93 0.94 0.92 0.95 0.98 0.89 0.97 0.93 0.99 0.94 0.98 0.95
0.95 0.94 0.93 0.97 0.94 0.98 0.97 0.95 0.94 0.94 0.96 0.95
0.94 0.94 0.93 0.95 0.98 0.94 0.96 0.98 0.97 0.94 0.97 0.92
0.93 0.90 0.95 0.92 0.90 0.93 0.89 0.94 0.95 0.94 0.97 0.94
0.99 0.97 0.96 0.94 0.95 0.97 0.95 0.96 0.94 0.93 0.90 0.95
0.97 0.94 0.94 0.97 0.97 0.94 0.94 0.92 0.92 0.98 0.97 0.94
0.96 0.97 0.98 0.91

Table 6					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8800	0.9300	0.9450	0.9465	0.9700	0.9900

- Finally, we generated the appropriate code in R in order to find the Mean Square Error (MSE) of the estimator $\widehat{\theta}$ for the `improved version`

- Finally, we generated the appropriate code in R in order to find the Mean Square Error (MSE) of the estimator $\widehat{\theta}$ for the `improved version`
- The results are presented in **Table 7**. The mean value of the MSE is 0.002954452

- Finally, we generated the appropriate code in R in order to find the Mean Square Error (MSE) of the estimator $\widehat{\theta}$ for the `improved version`
- The results are presented in **Table 7**. The mean value of the MSE is 0.002954452

Table 7			
MSE			
0.002994515	0.002913339	0.002922922	0.003003098
0.002944866	0.002891367	0.002930708	0.003148336
0.002784625	0.002916452	0.002874766	0.003285987
0.002843892	0.002825660	0.002980677	0.002992889
0.003039482	0.003101447	0.002886134	0.002807886

- For the `classical ICT` the MSE is obviously equal with the variance of the estimator θ. The results are shown in **Table 8**

- For the `classical ICT` the MSE is obviously equal with the variance of the estimator θ. The results are shown in **Table 8**
- Taking the mean value of the generated results we find that the MSE is 0.002941434

- For the `classical ICT` the MSE is obviously equal with the variance of the estimator θ. The results are shown in **Table 8**
- Taking the mean value of the generated results we find that the MSE is 0.002941434

Table 8
MSE
0.002994515 0.002877035 0.002922922 0.003003098
0.002944866 0.002855338 0.002930708 0.003070118
0.002784625 0.002916452 0.002874766 0.003285987
0.002843892 0.002790449 0.002980677 0.002992889
0.003039482 0.003062799 0.002850170 0.002807886

4. Discussion

- Equation (4), namely equation

$$\theta = \frac{E\,(C_1 - C_2) - Gp_0}{1 - G(p_0 + p_G)}$$

  is well defined if we assume that $1 - G(p_0 + p_G) \neq 0$

- The assumption will always be true if the list of non sensitive items contains two items for which the summation of their population prevalences is less than $\frac{1}{G}$

- It is achievable to have two non sensitive items for which the summation of their population prevalence is less than $\frac{1}{G}$
- In this case the probability that all G non sensitive items are applicable to an individual will also be less than $\frac{1}{G}$
- If the list of the non sensitive items includes two items with the summation of their population prevalences less than $\frac{1}{G}$, then it is expected that $(\hat{p}_0 + \hat{p}_G)$ will be less than $\frac{1}{G}$, so that equation (5) is well defined
- In case however that $1 - G(\hat{p}_0 + \hat{p}_G) = 0$, it is clear that equation (5) is not applicable

- Instead, equation (1) suggests that θ can be estimated from its sample equivalent,

$$\begin{aligned} \hat{q}_1 &= \frac{n_1^{(1)}}{n_1} \\ &= \left(\frac{1}{G} - \hat{p}_1\right)\hat{\theta} + \hat{p}_1, \end{aligned}$$

where $\hat{p}_1$ is the estimator of p_1 obtained from the second sample. Then

$$\hat{\theta} = \left(\frac{n_1^{(1)}}{n_1} - \hat{p}_1\right)\left(\frac{1}{G} - \hat{p}_1\right)^{-1},$$

which is well defined if $\hat{p}_1 \neq \frac{1}{G}$

- However if $(\hat{p}_0 + \hat{p}_G) = \hat{p}_1 = \frac{1}{G}$, then equation (2) for $k = 2$ suggests that θ can be estimated from:

$$\begin{aligned} \hat{q}_2 &= \frac{n_2^{(1)}}{n_1} \\ &= \hat{p}_2(1 - \hat{\theta}) + \frac{1}{G}\hat{\theta}, \end{aligned}$$

where $\hat{p}_2$ is the estimator of p_2 obtained again from the second sample. As previously, from the above calculations we get that:

$$\hat{\theta} = \left(\frac{n_2^{(1)}}{n_1} - \hat{p}_2 \right) \left(\frac{1}{G} - \hat{p}_2 \right)^{-1},$$

which is well defined if $\hat{p}_2 \neq \frac{1}{G}$

- If $(\hat{p}_0 + \hat{p}_G) = \hat{p}_1 = \hat{p}_2 = \frac{1}{G}$, then if we continue in the same manner we will get an estimator by using the sample equivalent of equation (2) for some k
- The entire process will only fail if

$$(\hat{p}_0 + \hat{p}_G) = \hat{p}_1 = \hat{p}_2 = \ldots = \hat{p}_{G-1} = \frac{1}{G}$$

- **But this scenario is unreasonable and it is not to be expected!**

Concluding Remarks

Concluding Remarks

- Practically, there is **not** any potential difference between the MSE of the 2 methods

Concluding Remarks

- Practically, there is **not** any potential difference between the MSE of the 2 methods
- The new version presented offers **better protection** than the original ICT

Concluding Remarks

- Practically, there is **not** any potential difference between the MSE of the 2 methods
- The new version presented offers **better protection** than the original ICT
- The new version does not give unbiased estimators BUT the main point is to get accurate results

Thank you very much for your attention!