RSS Higher Certificate in Statistics, 2010

Module 8 : Survey sampling and estimation

Solutions

Question 1

(i) £000 単位では，“5 未満”の企業に対して $\bar{y} = 3.5$, $\mathrm{SE}(\bar{y}) = s/\sqrt{n} = 3/\sqrt{100} = 0.3$ が与えられている．

従って，広告費の真の平均値に対する 95%信頼区間は $3.5 \pm (1.96 \times 0.3) = 3.5 \pm 0.588$ ，すなわち $(2.912,\ 4.088)$（単位は 1000£）である．

(ii) 従業員が 5 人以下のキングスタウンの中小企業の母集団全体に対して，広告費の真の平均値がこの区間に含まれていると 95%確信できる，という意味である．

(iii) 再び£000 単位を用いると，幅が 0.5 以下の信頼区間には，標本サイズ $n$ が $1.96 \times 3/\sqrt{n} \le 0.25$ を満たすことが必要である．

よって $\sqrt{n} \ge 1.96 \times 3/0.25 = 23.52$ ， $n \ge 553.19$ ．

この区間は幅が 0.5 以下なのを確かめるために切り上げると，$n$ =554 が得られる．

(iv) 再び£000 単位で考える．

全体の平均は

$$\bar{\bar{y}} = \frac{1}{18}(10 \times 3.5 + 6 \times 10.0 + 2 \times 35.0) = \frac{165}{18} = 9.167$$

$\bar{\bar{y}}$ の分散の推定量は

$$\sum_h \left(\frac{N_h}{N}\right)^2 \frac{{s_h}^2}{n_h} = \left(\frac{10}{18}\right)^2 \frac{9}{100} + \left(\frac{6}{18}\right)^2 \frac{25}{60} + \left(\frac{2}{18}\right)\frac{225}{15}$$
$$= 0.2777 + 0.04630 + 0.18519$$
$$= 0.25926$$

より $\mathrm{SE}(\bar{\bar{y}}) = \sqrt{0.25926} = 0.5092$ ．

従って，キングスタウンの全ての中小企業の広告費の真の平均値に対する 95%信頼区間は $9.167 \pm 1.96 \times 0.5092 = 9.167 \pm 0.998$ ，すなわち $(8.169,\ 10.165)$（単位は 1000£）である．

(v) 補助変数は，各企業を測定できなければならず，極めて密接に $y$ と関連していそうな（相関関係がありそうな）変数であるはずだ．$x$ としてありうる変数には，前年の会社の利益または売上高――これら両方が有用そうだ――が含まれる．（たとえば）従業員数の増加は活動の増加につながるためより多くの広告の必要性を示しそうなので，従業員数も有用な変数となりそうだ．仕事の全く新しい路線も広告の増加につながりそうでもあるから，おそらくダミー (0, 1) 変数を用いることによってモデル化することもできるだろう．どんな特殊なプロモーションでも同様の方法でモデル化することができる．

[試験では，可能な変数についての有効かつ関連する全てのコメント，提案に対して点数が与えられた．]

Question 2

(a)

(i) 調査では，測定値は標本の抽出単位ごとに求められる．この質問の例では，聞かれた質問に対し，はい/いいえ，と投票することである．標本の「はい」と「いいえ」の投票の総数が分かれば，「はい」と答えた投票の割合は分かる．この割合は調査されている母集団全体において「はい」と答える割合の推定に用いられる．優れた抽出方法によって (未知である) 真の母集団の値に (ある意味) 近い推定値を導出すべきである．もしある抽出方法を用いるときに，常に体系的に大きすぎたり小さすぎたりといった不十分な推定値が導出されるのならば，それを偏りという．

(ii) 多くの批判がされそうである．重要な点としては調査のもととなる母集団の定義の欠如，理にかなった抽出の枠の欠如，非常に低い回答率，誘導的で単純すぎる質問の性質，重要な追加情報を捉えることができない，といったものがある．概して，報告された「はい」と答えた割合の推定において非常に大きい方に偏りが起きそうである．

調査のもととなる母集団の定義の欠如は深刻な欠陥である．その母集団は地元のものだけであろうか? もしこれを意図するのならば，「地元」とは正確には何を意味するのか? 地元紙はどれくらいの範囲に流通しているのか? 母集団の定義が非常に限られたものであったとしても，回答率は非常に低い (町の母集団が 25,000 にもかかわらず，集められた標本サイズは 150 である)．結果として求められた推定値はほとんど信頼できない．

または乗り物に乗って町に入ってきたり，町を通過したりするドライバー (または乗客? ) を含んでいるつもりなのであろうか? もしそうならば，集められた標本サイズに関してさらに心配にさえなる．これはおそらく驚くべきことではないが，そのような人々，特に乗り物で通過する人々は地元紙を見ていなさそうである．彼らはその町の交通問題を減らしたり避けたりするために，おそらくバイパスを支持すると思われる．しかし，調査ではそのことを明らかに述べられていない．

したがって調査の抽出の枠はかなり不完全なものである．実際に存在すると言うこともほとんどできない．回答者は偶然新聞を読んだり質問を見かけたりした人に限られるし，さらにその新聞にテキストメッセージを送りたいと思っていて，かつ送ることのできる人に限られる．

他の重要な点として，聞かれている質問の性質が挙げられる．この質問は「誘導性」が非常に高い．「はい」と答えるのは当然である．危険な交通状態を町に残すべきと考えることのできる回答者がはたしているだろうか? さらに悪いことに道路の開通によってもたらされうる景観の美しい地域への損害について言及すらされていない．回答者の中にはこれに気付かず，議論の対象になっていないバイパスに対して単純に投票していると感じるものもいるかもしれない．また，全ての交通状態を「危険」とする暗黙の思い込みから生じた，非常に些細ではあるが重要な点が指摘される．危険な交通状態のみを町からなくすべきということを意図している本当の回答が全ての交通をなくすべきと支持しているものと誤解されるだろう．

一方でチップタウンから (危険な) 交通状態をなくすための道路を一般的には支持するが，提案されている特定の道路を支持しない回答者もいるかもしれない．この質問はそのような回答者からの適切な回答に対する全土地が全く与えられていない．

さらに中間の「分からない」という回答の設定がなされていないため，中心街をすっきりしたり景観の美しいエリアを通るといったりした問題に非常に強く意見を持っている人しか，投票しなさそうである．これは多数派の視点を代表しなさそうな極端な意見を捉えてしまうというありふれた問題を引き起こしそうだ．

年齢層の指示が設定されていない．異なった年齢層で明らかに異なった視点を持つかもしれない．または意見に影響を及ぼすような他のいかなる付随の情報も異なるかもしれない．例として回答者がどれくらい町に近いか，たいてい町を歩いて回るか，車を所有しているか，というのがあるし，他の可能性も多くありうる．はい/いいえ，という工夫の足りない回答はあまりにも単純すぎて価値のある指標にはなりえない．

考えるべきさらに重要な追加の質問は町の性質による．例えば，鉄道の駅，ショッピングセンター，ビジネスパークはあるだろうか? もしこれらの理由で人々が町にやってくるのならば，駐車場のスペースはバイパスに関係なく彼らすべての関心ごとになりそうだ．ラッシュアワーがあり，それは通過交通によってさらにひどいものになっていそうか? そこにはどれくらい通過交通があるか? 道路によって景観の美しいエリアにどれくらいの損害がもたらされるであろうか?

[上で簡単に挙げたものに加え，この質問への回答において指摘されそうな多くの関連する点がある．この調査は非常に欠陥が多い! 他の点に関する証明も明らかに可能である．試験では関連する点やちゃんとした証明には加点を行った．]

(b)

集落抽出では，チップタウン全体がクラスタと呼ばれるグループに分けられる．各クラスタは，調査された意見に関して，町全体にざっと似たものと思われている．特に，各グループは町全体を表すような全ての変数をできるだけとらえたものであるべきである．各エリアが町全体にざっと似ているように思われるならば，質問で述べられているエリアはこれらのグループとして用いるのに適切であろう．

適切なクラスタが定義されたとすると，集落抽出によってそれらのうち限られた数だけを調査する必要があるということが分かる．したがってクラスタの標本は調査によって抽出される．この標本は通常はクラスタすべてから単純無作為抽出によって抽出される．典型的にはほんの少数のクラスタから成り立つだろうが，たったひとつで成り立っているということも実際かなりよくあることである．そうすると，調査はこれら抽出されたクラスタから行われる．特に，直接訪問のインタビューによって調査が行われるのならば，これはたいてい管理を行うのに簡潔なものとなっている．それによって必要な時間が短縮されそうだ．さらに抽出の枠は抽出されたクラスタに限られるという点が挙げられる．そうすることによって他のクラスタの抽出単位の最新のリストが必要でなくなる．

したがって，クラスタ抽出は直接インタビューに役立つ．今回の例では，これにより適切な調査票を設計することができ，優れた結果を与えてくれそうだ．

集落抽出の基本的な概念をいろいろ拡張することも，この例では適切なものかもしれない．例えば，チップタウンの中心街に近いエリア (おそらく古いエリアだろうか? ) は市街地に離れた郊外のエリアではない市街地に非常に近い郊外のエリアとは違いを示すと思われるかもしれない．もしそうであるならば，中心街，市街地に近い郊外のエリア，市街地から離れた郊外のエリアの少数のクラスタを抽出し，最終的に分析を行う上でその結果を組み合わせるのが適切であろう．事実上，集落抽出と層化抽出の組み合わせである．

Question 3

(i) 層化によって母集団全体を調査できるだけでなく部分母集団をそれぞれ別々に調査することができる．もし層が適切に抽出されるのならば，母集団全体に対する分散の推定値は層化抽出のもとでは単純無作為抽出のもとで行ったものよりも小さくなるであろう．

現在の調査では，層を都会と地方のエリアにするのが理にかなっていると思われる．割合はこの 2 つでかなり異なる可能性があると考えてもよく，もしそうならば，分散の推定値は単純無作為抽出に比べて減少するはずである．いずれにせよ，その国全体についてだけでなく，都会と地方のエリアについて別々に情報を得ることは役立つように思える．

(ii) 問題の文字を用いると，各割合の推定値 $p_h$ の標準誤差は $\mathrm{SE}(p_h)=\sqrt{\dfrac{p_h(1-p_h)}{n_h}}$ と推定され，求める 95%信頼区間は $p_h \pm 1.96\mathrm{SE}(p_h)$ で与えられる．

したがって都会のエリアの区間は $0.3\pm\left(1.96\times\sqrt{0.3\times0.7/300}\right)=0.3\pm0.052$ で，すなわち 0.248 から 0.352 となる．

地方のエリアの区間は $0.6\pm\left(1.96\times\sqrt{0.6\times0.4/150}\right)=0.6\pm0.078$ で，すなわち 0.522 から 0.678 となる．

(iii) $p_r-p_u=0.3$ で，推定される分散は

$$\frac{p_r(1-p_r)}{n_r}+\frac{p_u(1-p_u)}{n_u}=\frac{0.6\times0.4}{150}+\frac{0.3\times0.7}{300}=0.0023$$

となる．したがって，割合の真の差が 0 かどうかを調べるための検定統計量の値は

$$\frac{0.3}{\sqrt{0.0023}}=6.255$$

これは $\mathrm{N}(0,1)$ から得られた観測値としては非常に有意であるため，等しい割合であるという帰無仮説に反する圧倒的な根拠となる．

(iv) 総数の推定値は $(8000\times 0.3)+(4000\times 0.6)=4800$ となる.

全体の割合に対する分散の推定値は

$$\sum_h \left(\frac{N_h}{N}\right)^2 \mathrm{Var}(p_h)=\left(\frac{8000}{12000}\right)^2 \frac{0.3\times 0.7}{300}+\left(\frac{4000}{12000}\right)^2 \frac{0.6\times 0.4}{150}=0.0004889$$

したがって総数の分散の推定値は

$$(12000)^2\times 0.0004889=70400$$

ゆえに求める全体の総数の 95%信頼区間は $4800\times\left(1.96\times\sqrt{70400}\right)=4800\pm 520.05$，すなわち 4280 から 5320 である.

(v) 比例配分は層の標本サイズ $n_h$ を層の母集団のサイズ $N_h$ と同じ割合で抽出するというものである. 母集団の層のサイズが 2:1 で標本の総数が 500 のとき，$n_h$ はそれぞれ $333.33[=(2/3)\times 500]$ と $166.67[=(1/3)\times 500]$ となるであろう. これより層の標本サイズはそれぞれ 333 と 167 をとる.

最適配分 (今回のようなコストの考えられていない簡単な状況では，しばしばネイマン配分と呼ばれる.) は与えられた総標本サイズ $n$ に対して，全母集団，この場合では総数の推定値の分散の最小化を目的としている. したがって各層の分散に注意する. ばらつきの大きな層ほど標本サイズは大きくなる. このとき, 2009 年の推定値を用いて標本サイズが大きいことに注意すると，$n_h$ は $N_h s_h$ に比例するようにとられる. ここで

$s_h=\sqrt{n_h p_h(1-p_h)}$ とする.

$N_h s_h$ の値は都会のエリアでは $8000\sqrt{300\times 0.3\times 0.7}=63498.03$ で，地方のエリアでは $4000\sqrt{150\times 0.6\times 0.4}=24000.00$ である.

したがって標本総数 500 に対して，$n_h$ はそれぞれ 362.85 と 137.15 となるであろう. このため層の標本サイズはそれぞれ 363 と 137 となる.

Question 4

(i) 見出しは誤っている.

見出しは南イングランド商工会議所がカバーする地域の企業の母集団全体についての意見であるという全く違う印象を与えるので, 正しいものでなく, 懸念されるものでさえある. 実際, 商工会議所に属している企業の標本調査を参照しているに過ぎない. それが商工会議所に属していない企業を含めた全体としての企業群を反映しているかどうかということに関しての情報がない. 返答は 100 社しかなく, 商工会議所に所属している企業が非常に多くあるような, おそらく大きな地域であるのに対し, 達成された標本サイズは非常に小さいように思える. したがってカバーし切れていない, 無返答といった多くの問題がある.

また, 実際は 40%が全従業員を増やすと予想しているという情報であるにもかかわらず, 「ほとんどの企業では雇用を減らすであろう」と報告したのも正しいものではなく, 懸念されるものである. 全従業員を増やしていないところは必ずしも減らしているというわけではない. 雇用に関しては安定させたままと予想しているのかもしれない.

さらに情報の詳細について全体的に欠けていることに関する問題がある.

40%という数字自体, 母集団全体の推定としてはかなり不正確なものとなりやすい. なぜならば, 無返答という問題だけでなく, 達成された標本サイズが小さいためである. 例えば, 信頼区間を与えることによって (おそらく 50%くらいの幅の区間となるだろう) その精度について何か指示があるべきである.

また企業の数に関して問題がある. 例えば, 製品を他の場所へたくさん運び出している大事業主はおそらくその地域全体に非常に大きな影響を持っているであろう. 一方で少数の小さな事業主の閉店は直接雇用している従業員以外にはほとんど影響がないと思ってよい. 関連するが, 別の点として, 企業はたとえ同じサイズであっても, 非常に異なった計画を持っているのかもしれない. 非常に小さな変化はほとんど影響を持たないのかもしれない一方で, 大規模な縮小 (または拡大) は地域の結果に対して影響を持ちうる.

さらに，非常勤労働の量は様々であるかもしれず，このことは次の 12 ヵ月後の間に変化すると予想されたり，されなかったりするかもしれない．調査がこのことを把握しているか，していないかに関して情報がない．

[試験では妥当で関連するコメントに関しては全て評価をさらに与えた．]

(ii) まず母集団がその地域の全ての企業から成るものなのか，または商工会議所に所属する企業のみから成るものなのかについて定義する必要がある．もし商工会議所に所属している企業にしか関係していないと結果を適切に報告すれば，後者の場合は批判はされないであろう．「企業」を構成しているものは正確には何かということや，企業はその地域にあるものとみなしていいのかどうか (例えば，国内の大企業や，さらには多国籍の企業のうちの非常に小さな部門にすぎないかもしれない．) ということに関して問題があるかもしれない．

もし調査が商工会議所に所属している企業に限られているのならば，それらのリストは存在するだろうから抽出の枠は直接利用可能である．もしさらに広い母集団を調査しているのならば，公的な (例えば政府の) ある種のリストを用いることができるかもしれない．

分けられた各層についての情報や，うまくいけば母集団全体のさらに正確な推定を得るために，層化を行うのはほとんど確実に賢明なことであろう．層化のいくつかの基準は，例えば，サイズや企業区分や地理的な位置のようにかなり自明なものである．

郵便による調査は可能であるが，ときどきそれを適切な回答者に送れたかどうか知るのが難しい．督促状もしばしば必要になってくるが，それでも返答率は低いものとなるかもしれない．調査票は前もって設計しておく必要があり，もし可能だとするならば，試験的に調査をすべきである (おそらく商工会議所に所属している企業のうち少数の企業を使うであろう．)．

または標本として抽出された各企業の適切な人に対し直接訪問を行う．これにはコストがかかるが，さらに優れた結果をもたらせてくれるかもしれない．さらに多くの質問をすることができるだろうし，より広い意見が引き出される．標本として抽出された企業と最初にコンタクトをとるのは難しいかもしれないので，おそらく電話をするのが最良な試みである．

母集団が何かという慎重な定義を含めて，母集団全体だけでなく各層についての結果を分析によってもたらすべきである．結果は点推定値に限定すべきではなく，信頼区間も含めるべきである．無返答の問題はいかなるものでも適切な議論を行うべきである．分析は詳細な情報や適切な要約を含めて行うべきであり，付随する報告では目立った結果を強調し，可能な解説を示すべきである．