RSS Higher Certificate in Statistics, Specimen

Module 8: Survey sampling and estimation

Solutions

Question 1

(a)

(i) X (ml) を缶に含まれている飲料水の容量とする．標本の大きさ (500) はとても大きいので，中心極限定理により標本平均の分布が近似的に正規分布となることが保証される．(問題文に X 自身が正規分布にしたがうことが仮定されていなくても，正規性を想定することは不合理ではないと思われる．)

$\bar{x} = 302$，$s = 0.5$ とおく．μ を（長期間見たときの）X の真の平均とする．μ の 95%信頼区間は $302 \pm (1.96 \times 0.5\sqrt{500}) = 302 \pm 0.044 = (301.96, 302.04)$ で与えられる．99%信頼区間は 1.96 を 2.576 に置き換えて 302 ± 0.058 = (301.94, 302.06) となる．

(ii) ばらつきはとても小さいので，機械は良好に作動している．信頼区間は指定された最小量 300 ml よりも十分に大きいところにあり，ばらつきに影響を与えることなく平均の設定を変更できるとすれば，缶の中身が規定の量を下回ることなく，平均 301（またはこれよりもいくらか小さい値）となるように設定を変えることができるであろう．

(iii) n を標本サイズとする．$1.96 \times 0.5/\sqrt{n} < 0.05$，すなわち

$$\sqrt{n} > \frac{1.96 \times 0.5}{0.05} = 19.6, \quad n > 384.16$$

が満たされなければならない．したがって $n = 385$ となる．（n は依然として大きいな値であるため，中心極限定理が十分適用できることに注意したい．n が小さな値になることが分かっていたならば，N (0, 1) の 1.96 を用いるよりも t 値を用いることを考えなければならなかったであろう．)

(b) 欠陥のあるリングプルをもつ缶の真の母割合を p とすると，p の推定値は $\hat{p}$ = 12/500 = 0.024 となる．$\hat{p}$ の分散は

$$\frac{\hat{p}(1-\hat{p})}{n} = \frac{0.024 \times 0.976}{500} = 0.00004685$$

と推定されるので，$\hat{p}$ の標準偏差の推定値は $\sqrt{0.00004685} = 0.00684$ となる．

したがって p の近似 95%信頼区間は (0.024 ± 1.96×0.00684) = (0.011, 0.037) で与えられる．1.1% から 3.7%の範囲が欠陥のあるリングプルをもつ缶の真の割合を 95%の確率でカバーしているといえる．

(c) 系統抽出は，例えば 10 個ごとなど，規則的な間隔で生産ラインから缶をとっていくというものである．これは簡単に行えそうである．(これに対して単純無作為抽出はまず缶に何らかの方法で番号付けをし，それらを「無作為」に選ぶ必要がある．これを生産ライン上で行うのならば，抽出を行うのに連続する缶の間で，ときにとても小さな間隔ができたり，大きな間隔ができたり

するであろう．これはとても不便であるかもしれないし，選ばれる缶はラインに沿ってやってくるのでとりそこないを引き起こしてしまうかもしれない．）しかしながら，缶に注入される量に何らかの傾向があるのならば，系統抽出によって同じような傾向をもつサンプルとなり，そのために推定に偏りを引き起こしてしまう可能性が出てくる．無作為抽出の際，求められた分散は非常に小さかったので，そのような傾向は無なさそうである．このような場合，系統抽出は受け入られるであろうし，サンプルは単純無作為抽出のようにふるまいそうである．単純無作為抽出の結果や公式はどの分析にも用いられるべきあろうが，これは理にかなっているようである．

Question 2

(i) 母集団とは調査される個体（人々など）の完全な集合のことである．標本とは母集団から選ばれたもののことである．フレームとは母集団の全ての個体のリストのことである．ここでの母集団は 1 ヶ月で発生した全請求であり，標本はいくつかの請求からなっており，フレームはその月のデータベースの全ての個体である．無作為抽出は各特定の個体の選ばれる確率が既知であるように母集団（フレームを用いており，個体は識別するために番号がつけられている）から選択することである．（単純無作為抽出では各個体が等しい実現確率を持っているが，他の無作為抽出方法も可能である．） 層化無作為抽出は，まず個体がグループや「層」に分けられ，標本（しばしば単純無作為抽出標本）が各グループから選ばれるようなときに行われる．層は機関の部門や個人の請求額の大きさ （おおまかにクラス分けされた）のようにグループによって異なるかもしれないような特徴を基準として形成される．

(ii) 各請求はデータベースに固有の照会番号を持っているため，優良な標本フレームが存在する．問題文の (1) (2) の 2 つの目的を達成するには複数の質問（情報の項目）を設定し，それらの回答がデータベース上のデータから得られるものとする必要がある．おそらく機関の部門によって層を形成するのが合理的であろう．また各部門において請求額の大きさによってさらに層を形成することも可能であろう．出張の必要性は部門間で異なるため，2 つ以上の請求のクラスを指定するのは有用ではあるが， おそらく請求額は単に「大きい」と「小さい」となりうる．もし大きな請求額に特定の関心があるのならば，この層はさらに集中的に抽出されるかもしれない．（すなわち抽出する上でその層にいるメンバーがさらに大きな割合で選ばれる．）集められた詳細情報は連続的な測定値（例えば，出張にかけた時間やポンド，ドル，ユーロ単位の請求額）であったり，いくつかの分類（例えば．出張するきっかけとなった仕事の種類）などである．明らかに移動方法は重要になるであろう．機関のどのような性質のものであるかが分からなければ，更なる詳細は与えられないのかもしれない．

提案された方法をテストするための事前調査は質問を洗練するのに非常に役に立つであろう．

分析はまず主な層の区分，すなわち部門によってなされるべきである．下層（請求額など）は各層内で調査することができる．全体の調査で平均化された結果を与えるのは無駄ではないが，層による違いや層の中での違いはさらに有用となるであろう．予期せぬ発見についても言及するかもしれないが，結論としては調査の目的に関係するものを述べるべきである．

(iii) 最初の 3 カ月後には請求の一般的な特徴について多くの情報が得られるであろうし，のちの研究で追加情報として用いられるかもしれない．回帰推定法は比推定法よりも前提条件が少なくて済む．もし有用な結果が与えられていなさそうならば，その情報は捨てればよく，そうすると予備分析にはその情報を含めるべきである．

Question 3

(i) p_S を公立の教員で「はい」と答えた割合とする． (同様に p_I を私立の教員のそれとする．) p_S の推定値は $\hat{p}_S = 120/300 = 0.4$ であり，その分散は $\hat{p}_S(1-\hat{p}_S)/300 = 0.4\times 0.6/300 = 0.0008$ と推定できる．したがって p_S の近似 95%信頼区間は $0.4 \pm 1.96\times\sqrt{0.0008}$ = (0.345, 0.455) となる．（この計算のさらなる詳細は Question 1 (b) の解答を見よ．）

(ii) 同様に $\hat{p}_I = 32/50 = 0.64$ であり，その分散は $0.64\times 0.36/50 = 0.004608$ と推定される．差 $p_I - p_S$ は 0.24 と推定され，この分散は $0.0008 + 0.004608 = 0.005408$ となる．したがって差の 95%信頼区間は $0.24 \pm 1.96\times\sqrt{0.005408}$ = (0.096, 0.384) となる．

区間内に値 0 は含まれないので p_I と p_S の差は有意である．つまり 9.6%から 38.4%の区間が真の割合の差を含んでいることを 95%の確率で信頼できるということである．

(iii) P をその地域の全ての教員の中で「はい」といった割合とする． P は重みづけ平均 $W_I P_I + W_S P_S$ であり， $\left(\frac{800}{5800}\times 0.64\right) + \left(\frac{5000}{5800}\times 0.4\right) = 0.433$ と推定される．重みはその地域の教員の「総数」に基づいて計算されることに注意したい．この推定値の分散は

$$W_I^2 \mathrm{Var}(\hat{p}_I) + W_S^2 \mathrm{Var}(\hat{p}_S) = \left(\frac{800}{5800}\right)^2 \times 0.004608 + \left(\frac{5000}{5800}\right)^2 \times 0.0008 = 0.0006822$$

で与えられる．したがって P の近似 95%信頼区間は $0.433 \pm 1.96\times\sqrt{0.0006822}$ = (0.382, 0.484) となる．

(iv) 層別化することで，調査されている特徴に関して，お互いが著しく違うと予想されるようなグループに母集団は細分される．それにより，すぐれた情報は各グループで別々に得られるかもしれないし，パラメータの母集団全体の推定値は単純無作為抽出からの標本よりもさらに正確になるであろう．この調査では，p_I と p_S にはっきりと有意差が見られるように，明らかにグループは著しく異なっている．すぐれた情報がそれぞれの層から得られたため， P は単純無作為抽出を行うよりもよい推定がなされている．

(v) 最適な配分は $n_h \propto N_h\sqrt{p_h(1-p_h)}$ となる．ここで h は層を表す． p_h の推定値を用いると，

$N_I\sqrt{p_I(1-p_I)} = 800\sqrt{0.64\times 0.36}$ となり， $N_S\sqrt{p_S(1-p_S)} = 5000\sqrt{0.4\times 0.6} = 2449.5$ となる．

総サンプルサイズは 250 であるから，それを $384.0 : 2449.5$ の割合で分けると，それぞれ 33.88 と 216.12 となる．したがって公立の区分から 216，私立の区分から 34 をとればよい．

Question 4

(i)

(a) N 個の個体からなり，分散が S^2 である有限母集団から抽出を行うとき， n を標本の大きさであるとする．このとき $f = n/N$ は「抽出比率」，$(1-f)$ は「有限修正」となる．その使用法を説明するために，抽出された n 個の個体で観測された連続変数を X とし，その平均を $\bar{x}$ とすると，標準的な結果として $\mathrm{Var}(\bar{x}) = (1-f)S^2/n$ が得られる．一方，無限母集団に対しては同じ記号を用いて $\mathrm{Var}(\bar{x}) = S^2/n$ となる．この $(1-f)$ を「有限修正」という． f がとても小さい（例えば，実用的には5% 未満と言われている）とき，有限修正を用いるのは計算上でほとんど数的に差異はなく，不都合な影響がほとんどないような状況では無視することができる．しかし f が大きいときには，常に有限修正を用いるべきである．

(b) S^2 が未知の場合， $\bar{x}$ の信頼区間を求めるとき（またはなんらかの形式的な仮説検定を行うとき）標本からの推定値 s^2 を用いなければならない．したがって N(0, 1) <u>ではなく</u>， t 統計量が必要となる．標準正規分布 N(0, 1) の限界点を用いている「無限母集団」における信頼区間の式は t_{n-1} 分布からの適切な値を用いたそれに対応する式で置き換えられる． t 分布を用いる基準の一つとして $n < 30$ があげられる．大きい n に対しては N(0, 1) を用いても十分であろう．

問題(a)の場合, 有限修正をしなければ分散の推定値は大きくなりすぎてしまう. 問題(b)の場合, t 値を用いなければ信頼区間は小さくなりすぎてしまう.

(ii) 経理部門: $f = 15/40 = 0.375$ となる． t_{14} の両側 5%点は 2.145 である．したがって 95%信頼区間は $30 \pm 2.145\sqrt{(1-0.375)36/15} = 30 \pm 2.73$，すなわち (27.27, 32.73) となる．

研究開発部門: $f = 10/20 = 0.5$ となる． t_9 の両側 5%点は 2.262 である．したがって 95%信頼区間は $15 \pm 2.262\sqrt{(1-0.5)16/10} = 15 \pm 2.02$，すなわち (12.98, 17.02) となる．

マーケティング部門: $f = 3/10 = 0.3$ となる． t_2 の両側 5%点は 4.303 である．したがって 95%信頼区間は $20 \pm 4.303\sqrt{(1-0.3)9/3} = 20 \pm 6.24$，すなわち (13.76, 26.24) となる．

(iii) 標本標準偏差は同じままであると仮定する． t の自由度を大きくするため，さらに大きな標本が必要となり，このとき信頼区間の幅を決めている 5%点が小さくなる． (ii)では $n = 3$ のときに ±6.24 という値が出された． $n = 4$ を試してみると，（t_3 から）±3.70 となるような 3.182 の t 値が与えられる． $n = 5$ を試してみると，（t_4 から）±2.63 となるような 2.776 の t 値が与えられる．したがって $n = 5$ を用いることになる．

(iv) 全体の平均は以下で与えられる通常の層化標本平均 $\bar{x}_{st}$ によって推定される:

$$\bar{x}_{st} = \sum_i \frac{N_i}{N}\bar{x}_i = \left(\frac{40}{70}\times 30\right) + \left(\frac{20}{70}\times 15\right) + \left(\frac{10}{70}\times 20\right) = 24.29$$

（重みは母集団の層の大きさによるということに注意すること.）この分散は

$$\begin{aligned}\mathrm{Var}(\bar{x}_{st}) &= \sum\left(\frac{N_i}{N}\right)(1-f_i)\frac{S_i^2}{n_i} \\ &= \left(\frac{40}{70}\right)^2(1-0.375)\frac{36}{15} + \left(\frac{20}{70}\right)^2(1-0.5)\frac{16}{10} + \left(\frac{10}{70}\right)^2(1-0.3)\frac{9}{3} = 0.59796\end{aligned}$$

ここで S_i^2 は標本から推定される．このプールされた分散の推定値の自由度は 25 である．

t_{25} の両側 5%点は 2.060 であるので，95%信頼区間は $24.29 \pm 2.060\sqrt{0.59796} = 24.29 \pm 1.59$，すなわち $(22.70, 25.88)$ となる．