RSS Higher Certificate in Statistics, Specimen B

Module 1: Data Collection and Interpretation

Solutions

Question 1

(i) 並び替えられた図は次のようになる（幹は万単位）.

| STEM | |
|---|---|
| 0 | 1 1 2 3 3 4 4 6 7 8 8 9 9 |
| 1 | 4 4 7 |
| 2 | 1 4 4 7 8 |
| 3 | 3 6 6 7 8 9 |
| 4 | 2 6 9 |
| 5 | 0 0 2 7 7 |
| 6 | 1 |
| 7 | 0 1 3 8 9 |
| 8 | 2 7 8 |
| 9 | 3 6 9 |
| 10 | 9 |
| 11 | 5 |
| … | |
| 16 | 0 |

これはかなりゆがみがあり，幹が 0 のものが多く，分布が右に裾をひいている．また途切れた部分もある.

(ii) 中央値は 25 番目と 26 番目の間であり，$M=\frac{37+38}{2}=37.5$ となる．四分位数は 13 番目と 38 番目であり，第 1 四分位数は $q=9$ で，第 3 四分位数は $Q=71$ である．（四分位数を求める際に他の定義を用いても良い.）これは千単位なので，$q=9000$，$M=37500$，$Q=71000$ となる．四分位範囲はしたがって 62000 となる.

平均は $\bar{x}=\frac{2217}{50}=44.34$ （千），標本分散は $s^2=\frac{1}{49}\left(16441-\frac{2217^2}{50}\right)=\frac{66139.22}{49}=1349.78$ であるから，標準偏差は $s=36.74$ （千）となる.

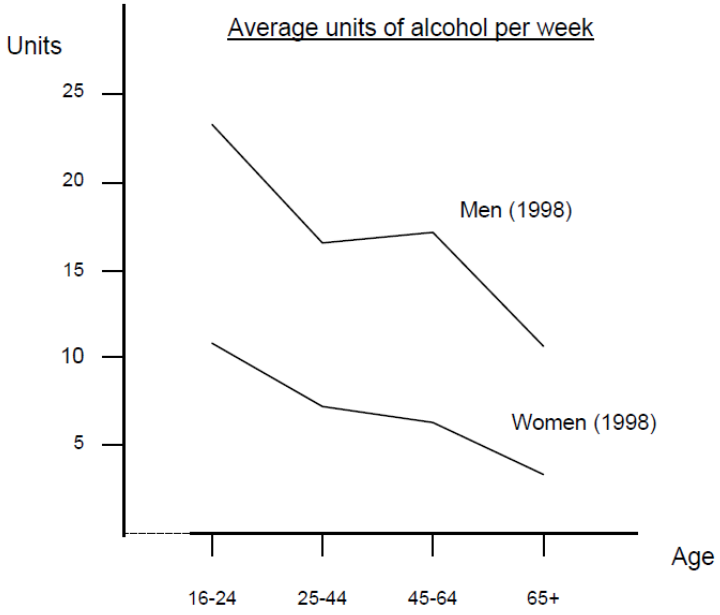
(iii) 上で述べた理由から平均と標準偏差は位置とばらつきを表す測定値としてはあまりよくないようである．中央値 37500, 四分位範囲 62000（または半分の 31000）が好まれるであろう．観測値の真ん中 50%の幅は 62000 である．小さい方 50%の幅は 37500 かそれより小さい値である.

Question2

(i) (a)


（電子的に複写した影響により，多少ギザギザしているかもしれない.）

(b)



(ii) 全体的にアルコールの消費量は期間を経るごとに少しずつ増加しているが，あまり規則的なパターンはない．

男性と女性の両方において 16-24 歳のグループは最初の 2 つと後半 2 つのデータ間(1992/94 と 1996/98)ではっきりとした増加を示している．

男性と女性の両方において 65 歳以上のグループは後半 2 つの期間（すなわち 1996 年から 1998 年）で減少している．これは 25-44 歳のグループにも言えることであるが，他のグループではかなり大きな増加を示している．

全体的には，図から 16-24 歳のグループでアルコールの消費が極めて多く，65 歳以上で低いということから大いに説明されることであるが，概して，年齢が上がるにつれアルコールの消費は減少している．概して，男性は女性の 2 倍から 3 倍飲酒している．

Question3

(i) (a) 無回答者が母集団全体に占める部分が無作為ではない傾向があり，特定のタイプの人はより返答に失敗したり，拒絶したりしやすい．彼らの反応は，聞くことができれば，母集団の他の部分とはかなり異なりそうである．もし彼らについて説明されなければ，調査の結果を母集団全体に適用するのは妥当ではないであろう．この偏りを生じることに加え，意図している標本サイズは無回答者の分だけ減り，それによって正確さに欠けてしまう．

(b) 考えられる手順は次のようである．

・1 度か 2 度以上再びアンケートを送る．
・督促の手紙を送る．（アンケート抜きで）
・まだ回答していない人たちに電話をする．
・まだ回答していない人々，あるいは彼らのうちの抽出された人々のもとへ訪れる．

これらは全て身元確認が必要であり，匿名性を保つために回答内容とは切り離されて考えられているアンケート番号によってたいてい確認される．

(ii) (a) 戦略 A では捜し出すのがとても簡単な人々の標本になってしまうと思われるので，たとえリストが適切に無作為な方法で成り立っていたとしても，実際に用いられる人々はリストから無作為に選ばれたとは言えないであろう．最初の依頼で拒絶する人に対しては，回答を説得しようとするよりもむしろ無視するであろう．インタビュアーのチームで行うため，これらの効果は 600 のうちのより高い割合で（より早く）起きそうである．あるいは逆に，早く終わったインタビューはそれほど徹底的に行われていなかったのかもしれない．できるだけ多くのインタビューをしようとせず，600 で止める理由などあるだろうか？

(b) 戦略 B もまた乗り気だったり，簡単に都合がついたりする世帯主が選ばれることになるであろう．すると，最初の $(600-m)$ 個の標本にはかなりの偏りが生じてしまう．次の $m$ 個の標本は，母集団をより表すものでなければならないが，$m$ の大きさによって 2 回目に用いられるデータの質は影響されるであろう（$m$ が大きいほど偏りが避けられる）．事務作業もこの方法により増える．

Question4

約 5600 人の会員が Grade I に属し，1400 人が Grade II に属し，約 5250 人が ABC の地域で，1750 人が世界のその他の地域にいることに注意しよう．

会員のアルファベット順のリストからの単純無作為抽出は計画しやすいであろう．アンケートは雑誌とは別に配られることができるだろうし，あるいはもしかしたら，次に発行される雑誌（あるいはニュースレターのような他の定期的な発行物）のうちの適当な冊数にアンケートを入れることもできるであろう．それはあまりいい方法ではないかもしれない．なぜなら Grade II に属し，「世界のその他の地域」に属している会員の割合はとても小さいからである．これらのグループはあまり抽出されないという危険にさらされるかもしれない．

層化無作為抽出は，簡便さに欠けてしまうが，よりはるかに満足のいく方法であろう．なぜならば，小さなグループだけでなく A/B/C の各グループから適正な抽出数を得ることができるとともに，適当なばらつき，コスト，そして割合が満たされているので，満足のいく調査が可能となるからである．さらに再分されたリストは有用であり，最新のデータ保存法によって下位グループに簡単に識別名を与えることができる．

割当標本抽出は，完全に実現不可能である．上で示されたようにいくつかのグループに分けるのが望ましい．もしそれが可能であるならば，割当標本抽出によって，必要となるサイズの標本を得ることができたであろう．しかし，それは不可能であるので，下位グループで理想的なサイズの標本を得るためには，無回答者に催促をするくらいしか方法はないであろう．

集落抽出は実現可能ではない．なぜならばクラスタに分ける明確な方法もなければ，クラスタからの抽出が妥当な方法となるように十分なクラスタを生成することもできないからである．クラスタから抽出を行う必要性についての理論的な根拠もまた全くなさそうである．

元のアルファベット順のリストからの系統抽出はとても簡単であり，おそらく単純無作為抽出と同じくらい満足のいくものとなるであろう．しかしながら，いくつかの名字が，ある地域ととりわけ関連しているというまた別のリスクが存在するので，層化がより望ましい．もしその抽出方法によって，抽出を行う異なったグループのリストを作成することになりそうならば，系統抽出の方が早いので，無作為抽出にかわって系統抽出が用いられうる．

可能なグループは Grade I と II で，それぞれ A，B，C，「その他」に分けられるであろう．「優れた」方法によってクラスを十分比較しなければならない．これがもっとも重要なことであろう．