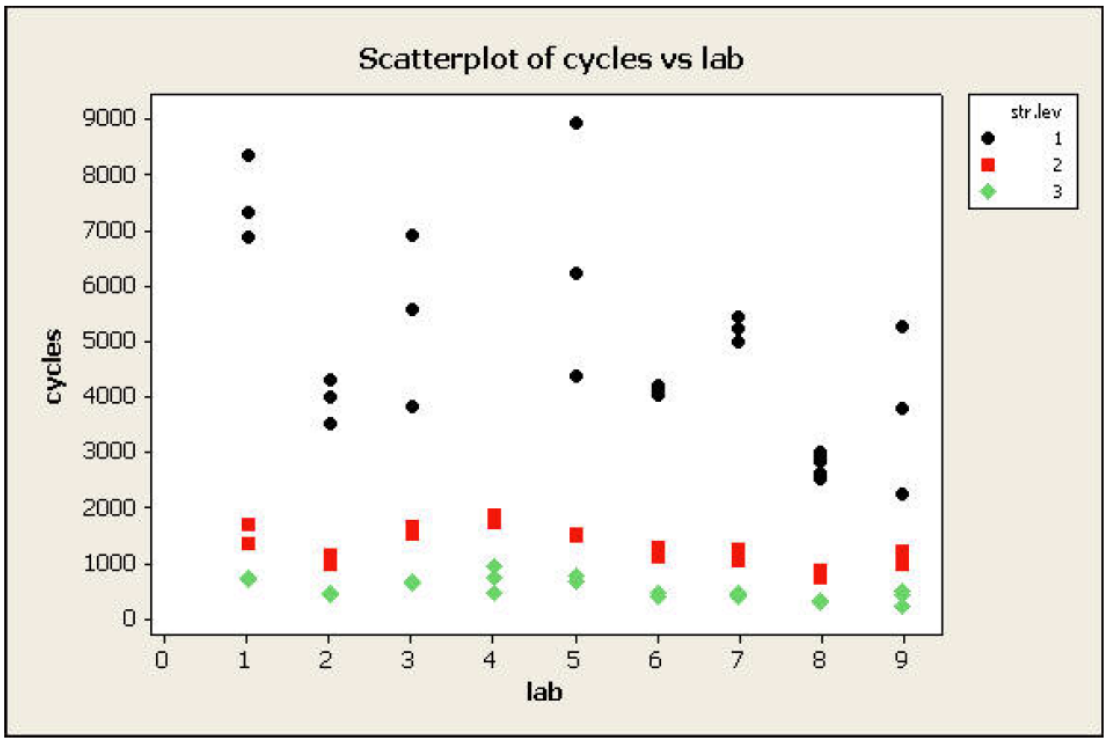
RSS Higher Certificate in Statistics, Specimen A

Module 1: Data Collection and Interpretation

Solutions

Question 1

各張力レベルについてそれぞれの研究室で測定されたものをグラフで示すと良い(ここでは研究室名 *A*-*J* を 1-9 に変えている).



3 つの張力レベル間の平均レベルに本質的な違いがあるため, 適当な尺度上で結果を示すのは難しい. しかし, いくつかの点が明らかになっている.

(1) いくつかの研究室では常に他の研究室より低い記録を出している. 例えば, *H* (8) はいたるところで最も低い平均値を出している. *A* (1), *C* (3), *E* (5) は高く, *B* (2), *J* (9) は低い傾向にある. 研究室内でのばらつきも大きく異なる. これは特に張力レベル 1 で見られ, *C* (3), *E* (5), *J* (9) のレンジはとても広いが, *F* (6), *G* (7), *H* (8) はそうでもない. 用いられた基本物質にはそれほどばらつきがあったようには見えない. なぜなら, 研究室内での全てのばらつきが大きいわけではないからである. 設備の資源や人材の点での技術的な理由がありえそうである.

(2) レベル 1 が最も低い張力レベルであり，レベル 2 や 3 よりもはるかに大きな平均値とばらつきを示している．明らかに疲労のサイクルと張力レベルの間には逆の関係がある．しかし，等分散を想定したモデルは不適切であろう．したがって変数 y（サイクル）を変換したものが調べられ，もしかしたらそれは $\log y$ かもしれない．予測は高い張力レベルで行った方が正確になり，すでにテストを行った範囲内でのみ行われるべきである．したがってレベル 1 以下やレベル 3 より高いレベルに外挿するのは賢明ではない．

全体的に，研究室は一貫性が欠如している．もし目的が研究室とは独立な予測をするということならば，研究室でのやり方をもっと一貫する必要がある．もし，それができないのならば，用いたモデルに研究室の影響を組み入れる必要がある．

Question 2

(i) 単純無作為抽出はサイズ N の母集団からサイズ n の標本を抽出するとして，そこから全ての標本が同じ確率で選ばれるというものである．（この確率はもちろん $1/\binom{N}{n}$ である．）この結果，目標母集団から選ばれたそれぞれの個体は標本として選ばれるのに同じ確率をもつ．

もし，母集団が全体としては均一でないが，それぞれで均一なグループに分けられるならば，それぞれのグループで無作為に抽出するのがよさそうである．（すなわち層化無作為抽出である．）これによって全体としての評価の精度が増すだけでなく，異なったグループを調べることができる．また，例えば名簿などの非常に大きな母集団を抽出することになったとき，系統抽出によってはるかに整理しやすくなったり，もしリストの傾向や周期が回避できるのならば，無作為に扱えたりするであろう．

(ii) (a)集計者や調査員の訓練不足，不注意や回答者の回答の誤解による解答の記録の誤り．郵送によるアンケートでは，質問の言い回しが不十分なために回答者が意図した質問に回答しないかもしれない．

(b)データがフォームから選ばれたり，入力システムに記録されたりする際の送信エラー．郵送による調査アンケートでは解読しにくい回答でも起こりうる．

(c)郵送による調査での無回答やインタビューへの拒否．これは調査されているトピックに興味がなかったり，質問の言い回しやインタビュアーの話の持ちかけ方にさしさわりがあったり，回答に時間を費やすのに気が進まなかったり，単純に質問が多すぎて調査できないことによって起きる．

(d)調査に選ばれた個人や集団をつきとめるのに失敗した．これは例えばリストの欠陥，インタビュアーが電話をした時間に都合がつかない，すでに引っ越しをしていたため空家であった，違う仕事や休暇のために調査を予定していないようなまれな時間に会う必要があるといった理由で起こる．

(iii) 電話調査は都合がつく人々や，電話をした時間帯に回答をしてくれる人々，調査のトピックにいくらか興味がある人々，電話番号が電話帳に記載されている（もし，電話帳がサンプルフレームとして使えるのならば）人々にしかできない．そのような調査に頻繁に電話がかかってきた人からは高い確率で回答を拒否されそうである．さらにいくつかの国では全員が電話を持っているというわけではない．

Question 3

Section1: Question 2 は未亡人，離婚や別居，同棲のボックスを増やす必要がある．

Question 3 はもっと明確にし，大人が何人，子供が何人家にいるかを聞くことができそうである．

Question 4 は家族全体の年収と書く必要がある．また，例えば£10000 がどこに入るか「£5000-£9999」，「£10000-£19999」などとして明らかにすべきである．

Section2: Question 2 は関係する全てのボックスにマークすることを説明すべきである．

Question 3 はかなり前の期間に対して記憶に頼ったり当てずっぽうになったりするのを避けるために 1 週間あたり と書くべきである．また今の question 3 は「主な買い物」とできる限り呼ぶことにして，他の質問で「それ以外の買い物」としてカテゴリーを例えば「£10 未満」，「£10-£19.99」，「£20 以上」とするような質問をしたほうがよい．

Question4 は過去 1 ヶ月のみにしたほうがよく（同様に記憶に頼ったり当てずっぽうになったりするのを避けるためである），「5 回より多く」，「3〜5 回」，「1 回か 2 回」，「全くない（またはこれまでほとんどない）」のようにいくつかの数を与えるべきである．

Question5 は 1 週間あたり と書き，おそらく明白に先週に限るべきである．「0」と，あとできる限り「4 より多く」というボックスを含めるべきである．

Question4

(i) 調査を計画する上で重要なことは，結果が適用されるうる母集団（これを目標母集団という）が何か，これを得るためにはどのような資源が必要かを正確に決めることである．もし目標母集団の一部を扱うことが難しかったり，費用がかかったりするならば，時間やお金や人材などの資源が満足なサイズの調査を行うのに十分でないかもしれない．目標母集団の一部を実用的な理由で省略し，省略された残りで調査が行われたとき，この残りを調査母集団という．調査母集団と目標母集団とは異なる．

もし，目標母集団の省略された部分が調査母集団とかなり異なっているようならば，そのことを報告書に明示しなければならない．

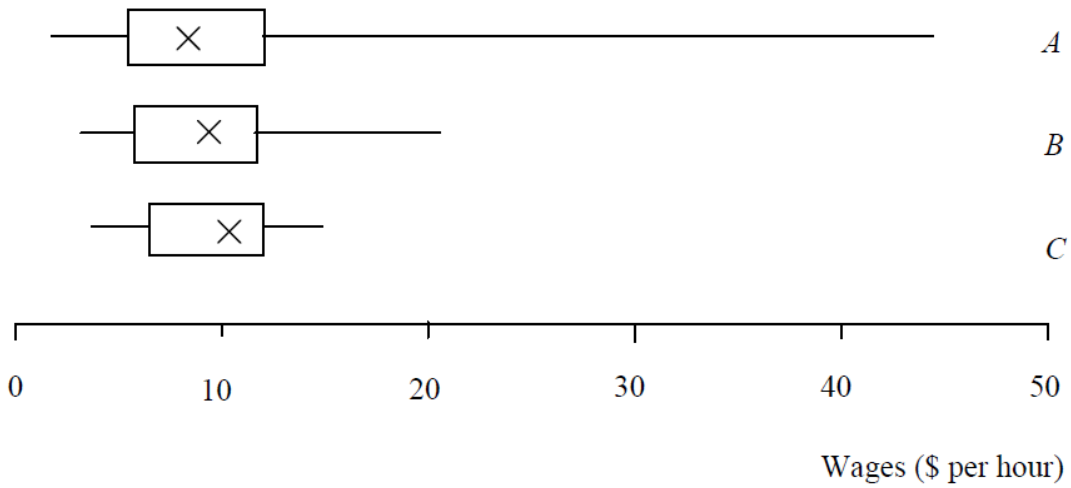
大規模な農家と地域の小自作農で育てられている作物の農業調査はそれぞれの選ばれた小自作農のもとへ行く必要があるため，非常に費用がかかる．しかし，2 つのタイプの農家が異なった結果を与えそうであると，たとえ小自作農がその作物の産出の割合が比較的小さくても 2 つのグループを調査しなければならない．

一方で，もし作物が程良く気候が一様である広大な地域で育てられているとするならば，その地域のいくつかの中心部でサンプリングするような方法を行い，それらの地域の近くで調査を集中できそうである．

抽出枠はできる限り最新のもので，完全なものである必要がある．重複や欠落は避けるべきである．上の最初の状況では，大規模な農家と同様に全体の地域での小自作農の良いリストが必要である．これは基本コストがかかりそうである．2 つ目の状況では，リストは全体の地域の限られた領域にのみ必要である．

(ii) 記述統計量の概略のみ与えられていただけでは，全ての有用な図が描けるというわけではない．（しかしアウトプットの一部として，散布図や幹葉図を求めることができる．）

統計的検定を何も行わないとするならば，最も有用な図は箱ひげ図である．
（ヒストグラムはアウトプットの一部として求めることができるが，*B* と *C* はヒストグラムを有用となるために十分な観測値が得られていない．）



C は左に歪んでおり，*A* は多少右に歪んでいるが，大きな外れ値を少なくとも 1 つは持っているように見える．全てのグループの中心値はほぼ同じで，*A* は *B* や *C* よりもばらついている．右の長いひげ部分を除いて，*B* はかなり対称的である．*C* は小さなグループであるが，中央値上にデータがいくらか集中しているように見える．