RSS Higher Certificate in Statistics, Specimen A

Module 6 : Further Applications of Statistics

Solutions

Question 1

(i) 主要な仮定は，実験誤差が独立した正規分布 N(0, σ^2)に従う変数であることである（σ^2 は定数）.

各処理法（Method）と各対象（subject）における和は

Method 1	Method 2	Method 3
78.6	79.3	86.9

Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7	Subj. 8
23.2	30.8	26.8	26.5	32.9	39.5	27.2	37.9

であり，総計は 244.8，$\sum\sum y_{ij}{}^2 = 2585.22$ である．また「修正項」は $244.8^2/24 = 249696$ である．よって，全平方和は $2585.22 - 2496.96 = 88.26$ ，

処理法の平方和は $\dfrac{78.6^2}{8} + \dfrac{79.3^2}{8} + \dfrac{86.9^2}{8} - 2496.96 = 5.30$,

対象の平方和は $\dfrac{23.2^2}{3} + \dfrac{30.8^2}{3} + ... + \dfrac{37.9^2}{3} - 2496.96 = 78.47$

となる．また，残差平方和はこれらの引き算で得られる．したがって分散分析表は以下のようになる．

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	8.26 Compare $F_{2,14}$
Subjects	7	78.47	11.21	34.95 Compare $F_{7,14}$
Residual	14	4.49	0.3207	$= \hat{\sigma}^2$
TOTAL	23	88.26		

$F_{2,14}$ および $F_{7,14}$ の上側棄却点は以下のようである．

	5%	1%	0.1%
$F_{2,14}$	3.74	6.51	11.78
$F_{7,14}$	2.76	4.28	7.08

処理法に関する F 値は高有意である．したがって，全ての処理法で凝固時間の平均が同じではないという強い根拠がある．また，対象に関する F 値は非常に高有意である．したがって，全ての対象で凝固時間の平均が同じではないという非常に強い根拠がある．この分析によって，体系的な変動の要因を検出したり削除したりした．

処理法による違いを調べる．各処理法の平均は

Method 1: 9.825　　Method 2: 9.9125　　Method 3: 10.8625

であり，これらの平均の任意の組に関する最小有意差は

$$t_{14}\sqrt{\frac{2\times 0.3207}{8}} = 0.283t_{14} \quad \text{ただし}\ t_{14} = \begin{cases} 2.145 & at\ 5\% \\ 2.977 & at\ 1\% \\ 4.140 & at\ 0.1\% \end{cases}$$

なので，有意水準 5% における最小有意差は 0.607，1%では 0.842，0.1%では 1.172 となる．明らかに，method 1 と method 2 は平均凝固時間内に差が見られないが，method 3 の平均凝固時間は他の平均凝固時間よりも長いという強い根拠が認められる．

(ii)　ここでの分散分析では「処理法」と「残差」だけで，「対象」という要因が出てこない．以下の残差における平方和は，(i) で残差とされたものと対象とされたもの「両方」を含む．したがって，新しい分散分析表は以下のようになる．

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	0.67 Compare $F_{2,21}$
Residual	21	82.96	3.95	$= \hat{\sigma}^2$
TOTAL	23	88.26		

ここでは，処理法の F 値は有意でない．したがって，各処理法の凝固時間の平均は等しいという帰無仮説を棄却する根拠はない．これは，未知の対象間の変動に由来する，データに見られるばらつきが大きすぎるからである．それゆえブロック化は，処理法の間の差を検出する力を大幅に強めたと言える．

Question 2

(i) 標本内の不良品の個数 X はパラメータ 20 と p の二項分布 B(20, p) に従う．よって

$$P(\text{バッチが合格となる}\mid p)=P(X=0\ or\ 1\mid p)=(1-p)^{20}+20p(1-p)^{19}$$
$$=(1-p)^{19}(1+19p)$$

である．また各 p におけるバッチが合格となる確率は以下のようになる．

$p = 0.01$: P(バッチが合格となる) = 0.9831

0.05 : 0.7358

0.1 : 0.3917

(ii) バッチが合格となるのは以下の場合である．

Number of defectives in first sample	
0	(Second sample not taken)
1	(Second sample not taken)
2	Second sample has 0 defectives

したがって，P(バッチが合格となる)

$$=(1-p)^{19}(1+19p)+P(\text{始めの標本内に不良品が 2 個}\underline{\text{かつ}}\text{次の標本内に不良品が 0 個})$$
$$=(1-p)^{19}(1+19p)+\frac{20\times19}{2}p^2(1-p)^{18}\times(1-p)^{20}$$

となり，各 p における確率は以下のようになる．

$p = 0.01$: $0.9831+0.01586\times0.81791=0.9831+0.01297=0.9961$

0.05 : $0.7358+0.18868\times0.35849=0.7358+0.06764=0.8034$

0.1 : $0.3917+0.28518\times0.12158=0.3917+0.03467=0.4264$

(iii) <u>検査手順 (i)</u> P(バッチが不合格となる) = 0.0169 for $p = 0.01$

0.2642 for $p = 0.05$

0.6083 for $p = 0.1$

検査される全装置数を S とする．

$E(S)$ = 20P(バッチが合格となる) + 1000P(バッチが不合格となる) の値は以下のとおり．

$p = 0.01$: $20\times0.9831+1000\times0.0169=36.6$

$p = 0.05$: $20\times0.7358+1000\times0.2642=278.9$

$p = 0.1$: $20\times0.3917+1000\times0.6083=616.1$

<u>検査手順 (ii)</u>

$E(S)$ = 20P(1 つ目の標本でバッチが合格となる)

$+ 40P$(2 つ目の標本でバッチが合格となる) $+ 1000P$(バッチが不合格となる)
の値は以下のとおり.

$$p = 0.01 \quad : \quad 20 \times 0.9831 + 40 \times 0.01297 + 1000 \times 0.0039 = 24.1$$
$$p = 0.05 \quad : \quad 20 \times 0.7358 + 40 \times 0.06764 + 1000 \times 0.1966 = 214.0$$
$$p = 0.1 \quad : \quad 20 \times 0.3917 + 40 \times 0.03467 + 1000 \times 0.5736 = 582.8$$

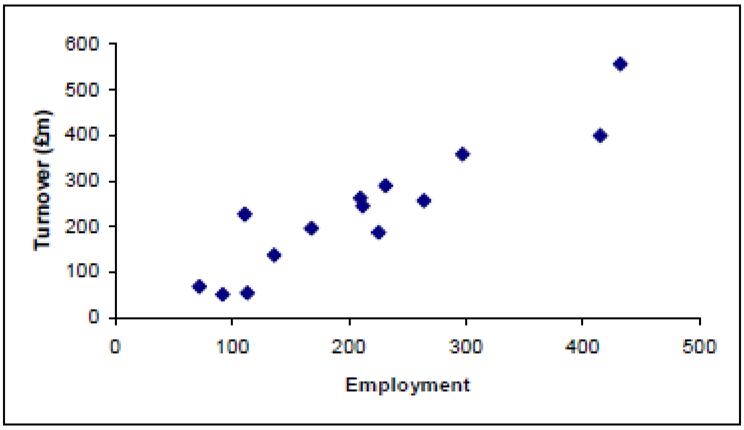
p が上記のような値であるならば，検査手順 (i) より (ii) の方が低い検査コストで済むだろう.

Question 3

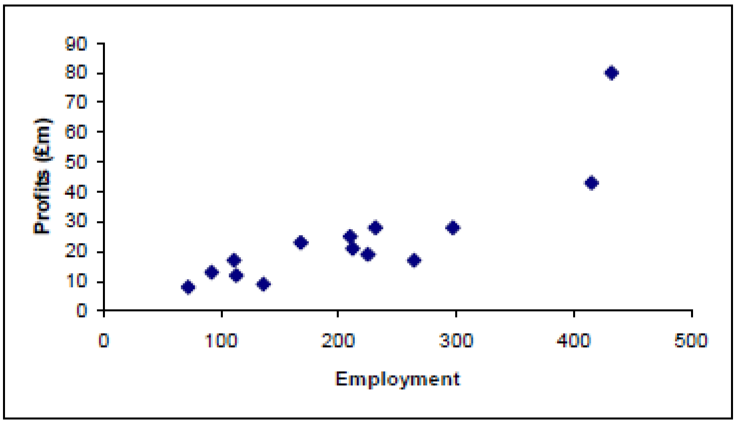
(i) 両方の 1 変数モデルにおいて，標準誤差から直ちに導けるのは，説明変数の係数は 0 との差が有意だが定数は有意ではないことである．これらのモデルは R^2 の値で比較され，2 つ目の「取引高のみ」のモデルの方がよいとわかる．このモデルの R^2 は完全（2 変数）モデルの R^2 と事実上等しいので，2 変数モデルを用いて何かが得られることはほとんどないようだ．さらに，取引高のみのモデルは解釈にほぼ困らない．というのは，標準誤差を考えると，定数も従業員数の係数も 0 との差が有意でないが取引高の係数の有意性はほんの少しだが存在する，と導けるからである．

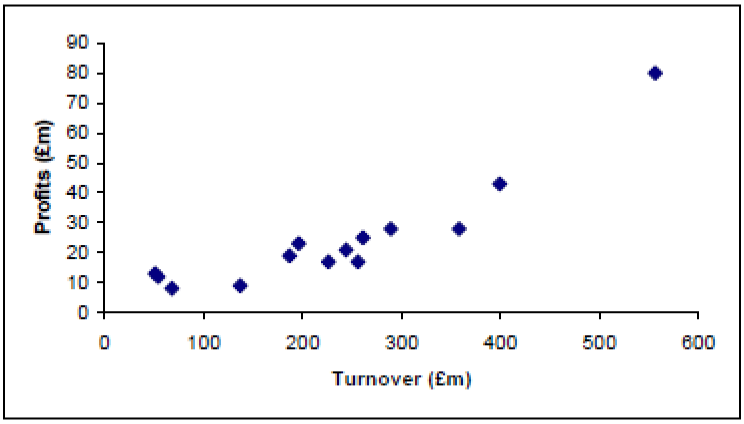
この結果から判断すると，取引高を用いた 1 変数モデルが最適だろう．

(ii) 全ての変数の組の散布図をつくると有益であろう．従業員数と取引高の散布図では，仮定した 2 つの説明変数の間にかなり強い線形関係があると読み取れるので，いずれか一方のみがモデルに必要であることを示している．



利益と従業員数の関係は，利益と取引高の関係よりもいくぶんか線形ではない（だが，視覚的な印象に基づいてどちらかを選ぶのはよくない）．





よってまとめると，最も適切なモデルは，y を利益(£m)，x_2 を取引高(£m)として $y = -3.15 + 0.118x_2$ となる．これは y の変動全体の 81.2%を説明するので，妥当なモデルだとみられる．取引高の代わりに従業員数を使って得られるあてはめはよくない．また，取引高だけでなく 2 つ目の説明変数として従業員数を加えるのも価値はないようだ．回帰係数のより大きな標準誤差で示されるように，それら両方を含むことはより正確でない回帰に至る．しかしながら，これまでに使われなかった変数で利用できる他のデータがあるなら，それをより大きなモデルで追加可能な説明係数として検討してもよい．

Question 4

(a) (i) 基本原理は平均値管理図の場合と同じである．製品ひとつひとつが検査され「良品」か「不良品」に分類される（「合格」か「欠陥」のような他の用語も用いられる）．個々のの製品がほとんど情報を与えないため，標本サイズは平均値管理図の場合よりかなり大きくなければいけない．

$\hat{p} = r/n$ を不良品である割合とする．通常，標本サイズ n は十分大きく，この分布は正規分布 $N(\hat{p},\, \hat{p}(1-\hat{p})/n)$ で近似できる．
この近似がよいためには，n が 50 以上である必要がある．π を不良品の割合の真の値とすると，π が非常に小さいときは条件 $n\pi \ge 10$ が満たされれば，近似における深刻な問題は避けられる．こうして，各標本サイズ n は 200 くらいかそれ以上だとしてよい．

管理図では，π の目標値として横線が描かれ，r/n の連続観測がプロットされている．上下方の処置限界線（action line）と警戒限界線（warning line）はそれぞれ

99%限界点での $\pi \pm 3.29\sqrt{\dfrac{\pi(1-\pi)}{n}}$ （3.29 の近似として 3 を使ってもよい），

95%限界点での $\pi \pm 1.96\sqrt{\dfrac{\pi(1-\pi)}{n}}$ （1.96 の近似として 2 を使ってもよい）

に引かれる．ただし π は目標値を意味している．（他の管理限界もときどき使われる．）下方限界が負になる場合は 0 に直す．

ある状況下では π の目標値がないだろう. その場合は過去の大量なデータに基づく $\hat{p}$ の値が目標値として用いられる.

プロットされた r/n の値が上方処置限界線を上回ったり, 連続した 2 つの値が上方警戒限界線を上回ったりしているときは，工程を止めて点検するというのが典型的なルールであろう．下方線を下回る値は必ずしも深刻というわけではないが，そのプロセスでの極端なばらつきの証拠として取り上げられるかもしれない.

(ii) プロセスの平均値で起こりうる変化を検出するのが特に重要な場合に，累積和管理図は役立つ．累積和管理図は変化を起こした時点も示している．

繰り返し標本の平均 $\bar{x}_1, \bar{x}_2, \ldots$ を用いて「累積和」の組

$$s_1 = \bar{x}_1 - k,\quad s_2 = \sum_{i=1}^{2}(\bar{x}_i - k),\quad \ldots,\quad s_r = \sum_{i=1}^{r}(\bar{x}_i - k),\quad \ldots$$

が得られる．k は母平均に対応する参考値であり，具体的に指定されるかあるいは過去のデータから計算される．これらは管理図にプロットされ，プロセスが正常に稼働しているなら, 標本平均は k より大きい値も小さい値もとりながら, k に近い値を取り続け,

累積和管理図はおおよそ横ばいになるだろう. プロセス平均で増加があれば, 差 ($\bar{x}_i - k$) のほとんどは正になり累積和管理図は常に増加を見せるだろう．同様に，プロセス平均が減少すれば管理図は常に減少を見せるだろう．累積和管理図はこのような変化の検出に敏感である．累積和管理図は過去全てのプロセスの情報を含んでいるので，たいていの場合は変化の時刻を十分正確に特定できる．しかし一方では，一貫して，適切な基準を開発し管理図を解釈すべきであるという主張がなされ，テキストブックにはいくつかの方法が提案されている.

大きな変化が起こったかどうかを決定するために V マスク法は有用で，シューハート管理図の限界線の目的と同様である．これらの詳細はテキストブックで与えられ，システムがうまく稼働しないときにどんな変化が予想されるかに応じて他の方法を用いることもできる.

(b) (i) モデル $Y = \alpha + \beta x$ を $Y = \alpha + \beta x + \gamma x^2$ に拡張することを考える. x の範囲のどこかで Y が最大値か最小値をとるようであれば，2 次関数は有用で，より一般的に，データが反映される領域内での曲線的な関係の説明に役立つ. 適合度がまだよくないときは, 3 次, 4 次というようにもっと項を加え続けたくなる——が，高次の係数の意味を解釈するのは非常に難しく確信できない．ただし，残差の正規性は依然として仮定してよい.

(ii) 直線関係とみなせる，より「自然な」モデルがある．たとえば $Y = ae^{bX}$ は $\log Y = \log a + bx$ とみなせる．このようなモデルはたとえば，反応が「加法的」ではなく「乗法的」な生物学的な研究でよく見られる．他のべき乗則の関係は生物学や農学その他でも見られる．たとえば $Y = cx^k$ は $\log Y = \log c + k \log x$ と変形できる．ここでの $\log c$ や k は推定されるべきパラメータである．指数モデルやべき乗モデルでのパラメータの説明は難しすぎてはいけない．データが導かれるプロセスをよりよく理解することが可能であるだろう．しかし，残差の正規性は保持されるだろうか？ 仮に変数変換前の正規性を仮定すると，変換後に正規性は保持されるだろうか？