

RSS Higher Certificate in Statistics, Specimen B

Module 6 : Further Applications of Statistics

1. たんぱく質の種類と量が体重の増加に与える影響を調べるための実験が計画され、それぞれ 10 匹のオスのマウスからなる 6 つのグループの各マウスに対し、2 つの量 (High, Low) と 3 つの種類 (A, B, C) の異なる組み合わせのたんぱく質がえさに混ぜられて与えられた。体重の平均増加量 (単位: グラム) は以下のものである。

		Protein type		
		A	B	C
Protein level	High	100.0	89.9	99.5
	Low	79.2	83.9	78.7

- (i) 「交絡」とは何かを説明せよ。上記のデータに関する適切なグラフを描き、たんぱく質の量 (level) と種類 (type) の間に交絡があるかどうかについてコメントせよ。
- (ii) 各マウスそれぞれの体重に関する分散分析表は以下のものである。「*」の部分の埋めて表を完成させよ。また、 F 比の解釈を与えよ。さらに、ここでの結論と上問 (i) のプロットとの間の関係を論ぜよ。

Analysis of Variance				
Source	DF	SS	MS	F
Level	*	3776.3	*	*
Type	*	82.5	*	*
Level*Type	*	730.1	*	*
Error	*	*	*	*
Total	*	16174.9		

- (iii) 当てはめられたモデルによって説明される変動の全変動に対する比率を求め、また残差分散の推定値を計算せよ。そしてその結果に対しコメントせよ。

2. 鋼板の穴あけ機があり、この機械は 1 枚の鋼板に 3 つの穴を開けるよう設定されている。機械の状態をチェックするため 12 枚の鋼板が選ばれて穴の直径を計測した結果は右の表のものである (単位: mm)。穴の直径の目標値は 22.5mm であり、これまでの経験から直径の標準偏差は 0.3mm である。穴の直径を表す確率変数 X は正規分布に従うと仮定する。

Sample	Diameters			Mean	Range
1	22.7	22.8	23.1	22.87	0.4
2	23.4	23.6	23.0	23.33	0.6
3	23.1	22.7	22.6	22.80	0.5
4	22.4	22.5	22.9	22.60	0.5
5	22.4	22.8	23.0	22.73	0.6
6	22.1	22.5	22.3	22.30	0.4
7	22.6	22.8	22.1	22.50	0.7
8	22.0	22.2	22.5	22.23	0.5
9	22.2	22.0	22.1	22.10	0.2
10	22.0	21.9	22.3	22.07	0.4
11	22.3	22.4	21.9	22.20	0.5
12	21.6	21.8	22.2	21.87	0.6

- (i) 機械が正常に稼動しているとした場合の X の確率分布は何かを述べ、3 つの観測値の平均値 \bar{X} の従う分布は何であることを示せ。

- (ii) 3つの観測値の平均値 \bar{x} に関するシューハート管理図を作成せよ。その際、警戒限界および処置限界を図に描き入れ、何故そのように設定したか、およびどのようにそれらを用いるのかについて説明せよ。
- (iii) 値の変動の大きさは範囲によって表現される。 R_i を i 番目のサンプル（鋼板）における範囲とし、管理図に $R_i, i = 1, \dots, 12$ をプロットせよ。その際、これまでの経験から範囲の目標値は 0.508 であり、 R の 95%信頼区間は 0.09 から 1.46 であることを用いよ。
- (iv) 上述の2つの管理図に基づき、この機械の稼動状況についてコメントせよ。
3. ランダムに選ばれた 5×5 のラテン方格を用いて実験を構成するにはどのようにすればよいかを説明せよ。

5種類の適性検査 (A – E) が5日間に渡って5人の被験者 (subjects) に対して実施される。被験者は誰もその種の適性検査をこれまで受けたことがなく、検査に対する慣れによる効果を取り除く必要があることから、下記のような実験計画によって実験が行われ、スコアが得られた。

	<i>Day 1</i>	<i>Day 2</i>	<i>Day 3</i>	<i>Day 4</i>	<i>Day 5</i>	<i>Subject totals</i>	<i>Totals for Tests</i>
Subject 1	E 56	B 62	A 65	D 59	C 76	318	A: 331
2	C 74	A 65	D 60	E 61	B 70	330	B: 332
3	B 63	E 59	C 80	A 66	D 64	332	C: 393
4	A 64	D 63	B 67	C 81	E 64	339	D: 310
5	D 64	C 82	E 64	B 70	A 71	351	E: 304
Day totals	321	331	336	337	345	1670	1670

これらのデータに関する分散分析表を作成せよ。ただし以下の値を用いてもよい。

25個の観測値の平方和は 112754 である。

被験者に関する平方和は $318^2 + 330^2 + 332^2 + 339^2 + 351^2 = 558370$ である。

実施日に関する平方和は $321^2 + \dots + 345^2 = 558092$ である。

検査に関する平方和は $331^2 + \dots + 304^2 = 562750$ である。

適切な検定を用いることにより、検査は同じ能力を測定しているという仮説の真偽、ならびに被験者間およびと実施日間に差があるか否か、について調べ、結果を簡潔にレポートせよ。

4. 以下の各問に答えよ。
- (i) 目的変数 Y に対し r 個の説明変数 x_1, x_2, \dots, x_r があるとして、後進変数除去による適切な重回帰モデルを選択する手順について説明せよ。
- (ii) (Y, x_1, x_2, x_3, x_4) に関する $n = 13$ 組の測定値が得られた。目的変数 Y をなるべく少ない数の説明変数でモデル化して説明したいとする。説明変数のすべての異なる組み合わせによる回帰モデルに関する情報が以下のように与えられているとき、後進変数除去法により適切な回帰モデルを選択せよ。

Predictors used	df	Residual (error)		F value for fitted regression, with df	
		Mean square	R ² (%)		
x_1	11	115.062	53.39	12.60	1, 11
x_2	11	82.394	66.63	21.96	1, 11
x_3	11	176.309	28.58	4.40	1, 11
x_4	11	80.352	67.45	22.80	1, 11
$x_1 x_2$	10	5.790	97.87	229.50	2, 10
$x_1 x_3$	10	122.707	54.82	6.07	2, 10
$x_1 x_4$	10	7.476	97.25	176.63	2, 10
$x_2 x_3$	10	41.544	84.70	27.69	2, 10
$x_2 x_4$	10	86.888	68.01	10.63	2, 10
$x_3 x_4$	10	17.574	93.53	72.27	2, 10
$x_1 x_2 x_3$	9	5.346	98.23	166.35	3, 9
$x_1 x_2 x_4$	9	5.330	98.23	166.83	3, 9
$x_1 x_3 x_4$	9	5.648	98.13	157.27	3, 9
$x_2 x_3 x_4$	9	8.202	97.28	107.37	3, 9
$x_1 x_2 x_3 x_4$	8	4.983	98.53	134.25	4, 8

- (iii) 説明変数を1つだけ選ぶとしたら x_4 であることが上の表の情報からどのように帰結されるのかを述べよ。この結果と後進変数除去法との間に食い違いがあるとしたらそれについて簡潔に説明せよ。