

アンバランسد・データに対する重みづけしない平均を用いた分散分析と SAS の GLM プロシジャーの比較

田 中 敏*

(平成4年10月30日受理)

要 旨

アンバランسد・データに対して、重みづけしない平均を用いた分散分析 (UMA) の F 比と、SAS (Statistical Analysis System, (c) SAS Institute Inc.) の GLM (General Linear Model) プロシジャーの F 比とを比較するため、コンピュータ・シミュレーションを実施した。2×3の実験計画において各セルの N の初期値を20または50に設定し、総データ数の10%、30%、50%を欠損させて、UMA において主効果 (df=2) または交互作用 (df=2) が10%水準で有意であったケースを生成した。各効果の有意性について1000ケースを収集した結果、UMA と GLM (Type III) プロシジャーの F 比の差は、10%と30%の欠損条件では比較的小さかったが、50%の欠損で無視できない不一致が見られた。

KEY WORDS

unbalanced data アンバランسد・データ analysis of variance 分散分析
unweighted-means ANOVA 重みづけしない平均を用いた分散分析
GLM (Type III) procedure in SAS SAS 一般線型モデル (Type III) プロシジャー

問 題

実験計画における各条件のデータの個数 (N) がふぞろいである場合、それをアンバランسد・データ (unbalanced data) と呼ぶ。アンバランسد・データでは各条件の平均の重みが異なるので、そのまま分散分析をおこなうと、不当に実験要因の分散値を大きくしてしまう。このため、アンバランスの程度に応じた重みづけを平均に施さなくてはならないが、重みの付け方が複雑である。したがって簡便法として、各条件の N を等しいものと仮定して (各条件の N の調和平均をとる)、「重みづけしない平均」 (unweighted means) を用いた分散分析も一般に多用される。これはアンバランスの程度が著しくない場合に限っての近似法としておこなわれるが、どの程度のアンバランスまで、この近似が有効であるかは定かではない。そこで、本研究は、この重みづけしない平均を用いた分散分析 (unweighted-means ANOVAs : 以下 UMA と略記) の近似の有効性をコンピュータ・シミュレーションによって推定しようとする。

UMA は、STAR (Statistical Analysis Rescuers, (c) 1992 田中敏) 所収の PUMA (Package of Unweighted-Means ANOVA) version 2.1 を使用した。また、これと対照するため SAS

* 教育方法講座

(Statistical Analysis System) version 6.04((c)1988 SAS Institute Inc.) 所収の GLM (General Linear Model) プロシジャーを使用した。GLM プロシジャー Type III モデルは、データのアンバランスに応じて平均を重みづけした分析結果をあたえてくれる。

アンバランスド・データは 1 要因計画の分散分析、および、2 要因計画における自由度 1 の効果の分散分析に関しては問題を生じない。たとえば、 2×2 配置の分散分析では、SAS の ANOVA プロシジャーと GLM プロシジャーの結果はデータのアンバランスに関係なく完全に一致する。そこで、本研究では、2 要因計画の 2×3 配置の分散分析をシミュレートする。最初に、モデル・ケースを分析し、次にランダム・サンプリングによるシミュレーションを実施する。

研 究 I

目 的

2×3 配置 6 条件の平均と標準偏差を固定して、主効果のみ有意である場合と交互作用のみ有意である場合の 2 つのモデル・ケースを構成する。そして、1 つの条件の N を欠損させてゆく。これらのケースに対して、バランスド・データに用いる通常の分散分析 (ここでは誤用となる) と、UMA, GLM の計算結果を比較対照する。

方 法

計 画 要因 A (2 水準) と要因 B (3 水準) の被験者間計画。

設 定 主効果のモデル・ケースでは、要因 B の 2 水準における要因 A の 3 平均を、それぞれ 50, 55, 60 と 60, 65, 70 に固定した。交互作用のモデル・ケースでは、同様の順番で各平均を 50, 55, 60 と 60, 55, 50 (逆方向の交互作用 disordinal interaction となる) に固定した。標準偏差はすべて 10 である。N の初期値は 20 とした。以上の設定において、データの欠損を 6 セルのうち 1 セルだけに 3 段階で実施した。各段階の欠損数は、4, 10, 16 とした (比率でいえば 20%, 50%, 80% の欠損である)。たとえば欠損数 16 の場合の 6 セルの N は、20, 20, 20, 20, 20, 4 となる。

結 果

Table1 は、主効果と交互作用の各モデル・ケースにおいて、データを欠損させた場合の、通常の分散分析 (誤用) と、UMA, GLM プロシジャーの計算結果を比較したものである。それによれば通常の分散分析 (誤用) は、データのアンバランスが交互作用 ($df=2$) に算入されることがわかる。たとえば主効果モデルのケースをみると、データ欠損数 16 の交互作用は有意性さえ示すまでになっている ($F(2,48)=2.99, p=.055$)。その前段階のデータ欠損数 10 でも、交互作用の F 比 2.16 は $p=.120$ である。これに対して、UMA と GLM プロシジャーの偏差平方和 (SS) と F 比は、いずれもゼロであり、主効果モデルでは、まったく交互作用を検出しない。UMA と GLM プロシジャーは主効果 A ($df=1$) についても完全に一致する。しかし、主効果 B ($df=2$) については、0.3~0.4 程度のズレを生じている。

交互作用モデルのケースでは、通常の分散分析 (誤用) の偏向は主効果に反映されている。

Table 1
Sum of squares (SS) and F-ratios yielded by UMA and GLM (Type III) procedure
in the 2×3 design cases with varied number of missing values
in one cell, given that N of each cell is 20.

	Misused ANOVA		UMA ¹		GLM (Type III) ²	
	SS	F	SS	F	SS	F
[Main effect model]						
# of missing values=4						
Main effect A (df=1)	2692.98	25.23	2879.66	27.30	2879.66	27.30
Main effect B (df=2)	1697.41	8.05	1920.14	9.10	1884.74	8.94
Interaction A×B	195.55	0.93	0.00	0.00	0.00	0.00
# of missing values=10						
Main effect A (df=1)	2209.29	20.89	2571.60	24.31	2571.60	24.31
Main effect B (df=2)	1241.99	5.87	1713.86	8.10	1624.56	7.68
Interaction A×B	457.55	2.16	0.00	0.00	0.00	0.00
# of missing values=16						
Main effect A (df=1)	1698.92	16.01	1799.52	16.96	1799.52	16.96
Main effect B (df=2)	820.49	3.87	1199.82	5.65	1142.67	5.38
Interaction A×B	634.01	2.99	0.00	0.00	0.00	0.00
[Interaction model]						
# of missing values=4						
Main effect A (df=1)	3.68	0.03	0.00	0.00	0.00	0.00
Main effect B (df=2)	7.66	0.04	0.00	0.00	0.00	0.00
Interaction A×B	1885.02	8.94	1919.83	9.10	1884.44	8.93
# of missing values=10						
Main effect A (df=1)	27.26	0.36	0.00	0.00	0.00	0.00
Main effect B (df=2)	60.60	0.29	0.00	0.00	0.00	0.00
Interaction A×B	1639.24	7.75	1714.11	8.10	1624.85	7.68
# of missing values=16						
Main effect A (df=1)	83.95	0.79	0.00	0.00	0.00	0.00
Main effect B (df=2)	205.05	0.97	0.00	0.00	0.00	0.00
Interaction A×B	1249.19	5.89	1199.64	5.65	1142.63	5.38

¹ Unweighted-means ANOVA.

² General Linear Model (Type III) Procedure in Statistical Analysis System.

当然、UMA と GLM プロシジャーは、まったく主効果を検出していない。しかしながら、交互作用の計算結果は、やはり0.3~0.4程度の F 比のズレを生じている。

したがって、次後のシミュレーションでは、主効果モデルのケースにおいては自由度2の主効果(B)について、また交互作用モデルのケースにおいても自由度2の交互作用(A×B)について UMA と GLM プロシジャーを比較することにする。

研 究 II

目 的

シミュレーションによって、UMA と GLM プロシジャーの F 比の差の分布をつくる。そして、アンバランスド・データに対する UMA の近似の程度を検討する。

方 法

計 画 要因 A (2 水準) と要因 B (3 水準) の被験者間計画。

設 定 パーソナル・コンピュータ (PC-9800, NEC) の疑似正規乱数を用いてデータを発生する。N の初期値は 1 セル当たり 20 と 50 の 2 つの場合をもうける。全 6 セルのうち自由度分の 2 セルを N=20 に固定して、残り 4 セルのデータを無作為に欠損させる。欠損率は 10%、30%、50% の 3 段階とする。ただし、セルの N が初期値の 80% まで欠損したら、そのセルからの欠損を停止する。たとえば、N の初期値 20 の場合は N=4、また N の初期値 50 の場合は N=10 となった時点で、それ以後、そのセルからは欠損しない。

シミュレーションの達成目標として、1 設定につき UMA と GLM プロシジャーとの 1000 個の F 比の差を収集する。このため、UMA において当該の主効果または交互作用が 10% 水準の有意性を示したアンバランスド・データを 1000 ケース収集した (有意な主効果のケースは交互作用が有意でないときに限った)。そして、これらを GLM プロシジャーによって処理することにした。

結 果

Table2 は、各設定における UMA と GLM プロシジャーの F 比の差について平均と標準偏差を示したものである。N の初期値 20 の場合、主効果モデルも交互作用モデルも、F 比の差は欠損率の大小に関係なく小数点第二位以下の変化に止まっている。ただし、標準偏差は欠損率に比例して大きくなる傾向が見られる点に注意を要する。

これに対して、N の初期値 50 の場合、欠損率 30% までは近似は良好であるといえるが、欠損率 50% では F 比の差は 0.2~0.3 前後の小数点第一位にまで及ぶ。また、標準偏差も、やはり欠損率に比例して大きくなるようにみえる。

Table3 は、実際に、各設定において生成したアンバランスド・データ 1000 ケースの有意性を検定した結果である。5% 水準の有意性に限ってみても、傾向はやや読み取りにくいところがあるが、概して 30% の欠損率までは、そう大きな不一致は起こっていないように思われる。

Table 2
Difference between F-ratios yielded by UMA and GLM (Type III) procedure
for one thousand cases in which the main effect B (df=2)
or the interaction is significant at .10 level by UMA.

N per cell % of missing values	20			50		
	10	30	50	10	30	50
[Main effect]						
Mean	0.027 ¹	0.026	-0.045	0.031	0.056	0.310
S.D.	0.342	0.439	0.884	0.046	0.279	1.342
Minimum	-3.54	-1.22	-1.80	-0.15	-0.65	-2.93
Maximum	3.02	1.03	1.88	0.14	0.78	5.50
[Interaction]						
Mean	0.003	0.070	0.037	0.036	0.031	0.241
S.D.	0.117	0.399	0.604	0.036	0.275	0.492
Minimum	-0.33	-1.32	-1.14	-0.22	-0.73	-0.88
Maximum	0.27	1.38	0.99	0.38	0.69	1.27

¹ Subtracting F-ratio of General Linear Model (Type III) procedure
from F-ratio of unweighted-means ANOVA for the same sample.

Table 3
Number of cases in which the main effect B (df=2) or the interaction
is significant at .05 or .01 level.

% of missing val.	Main Effect B (df=2)					Interaction				
	UMA ¹		GLM(Type III) ²		Total	UMA		GLM(Type III)		Total
	p<.05	p<.01	p<.05	p<.01		p<.05	p<.01	p<.05	p<.01	
[N=20]										
10	494	81	510	81	9972	470	97	497	97	9929
30	534	70	524	98	11130	498	68	464	88	9155
50	450	100	501	250	16079	565	66	631	66	10715
[N=50]										
10	500	125	624	125	9962	442	7	442	6	11093
30	677	77	604	75	7796	538	78	461	78	7637
50	629	184	481	110	14339	453	120	453	48	9267

¹ Unweighted-means ANOVA.

² General Linear Model (Type III) Procedure in Statistical Analysis System.

Comparison between ANOVA Using Unweighted Means and GLM Procedure in SAS for Unbalanced Data

Satoshi TANAKA*

ABSTRACT

A computer-simulation was conducted to compare F-ratios of analysis of variance (ANOVA) using unweighted means and F-ratios of general linear model (GLM) procedure in Statistical Analysis System ((c) SAS Institute Inc.) for unbalanced data. Given that initial N of each cell is 20 or 50 in 2×3 experimental design and percentage of missing values for total sample size is varied from 10%, 30%, to 50 %, those cases of unbalanced data in which the main effect ($df=2$) or the interaction was significant at .10 level by ANOVA using unweighted means were generated. As results from one thousand cases for each significance, differences between F-ratios yielded by ANOVA using unweighted means and GLM (Type III) procedure were relatively small in those conditions of 10 % and 30 % missing. However, in the condition of 50 % missing, there was unignorable discrepancy between ANOVA using unweighted means and GLM (Type III) procedure.

* Division of Method and Evaluation.