CHARACTERIZATION OF VARIANCE IN MEDICAL DIAGNOSTICS: ANALYSIS OF "USUAL" DATA

John A. Wass
Abbott Laboratories
Abbott Park, IL

Introduction

A variety of diverse data sets are offered to the industrial statistician in order to characterize variability in reagents and instrument systems. Many times this data contains putative outliers, missing data, augmented data or so-called "equivalent data." To generate estimates of variability the analyst must first deal with these minor problems before proceeding to the major analytic functions. For the examples contained herein we will address outliers, missing data and equivalent data. In the medical diagnostics industry addressing the above may be as simple as calculating variances for complete data sets, or as non-trivial as characterizing the variance components from a variety of sources. The SAS system offers a number of ways of accomplishing this and the analyst must be aware how the design and source characterizations affect the choice of statistical tools.

A Sample Data Set

Our sample data set (Table 1.) comes from a single control level of a standard clinical chemistry analyte. These data are collected on automated clinical analyzers that can repetitively sample and randomly access any given sample cup in a multi-sample carousel. Any given repetitive sample of a given cup is called a "rep" while the sum of all reps on that sample constitutes a run.

The analyte is aliquoted into three separate sample cups, all of which are repetitively sampled so there are three runs per instrument. Alternatively, 3 separate preparations (of the same reagent) are made and each aliquoted into 3 cups. Two instruments are run in this manner for a period of 5 days. Therefore we have, 3 reps per run, 3 runs per instrument, 2 instruments per day for five days for a total of 3 x 3 x 2 x 5 or 90 data points. How we treat these data points will depend on what we would like to define as our experimental unit.

We will first examine the data for transcription errors and then examine the resultant set for outliers. This can be done graphically by:

```
data analyte;
input day instru run rep concen;
time = run + day*10;
datalines;
(data here)
proc gplot data = analyte;
plot y*time = instru;
run;
```

```
 proc sort data = analyte;
 by instru;
 run;

 proc gplot data = analyte;
 by instru;
 plot y*time / vaxis = 150 to 210
by 10  vminor = 1;
run;
```

By visualizing the data graphically we reveal several potential outliers as well as trends, i.e. an increase in means with time and a decrease in variance with the second instrument. The putative outliers may be further investigated for possible removal by checking experimental logs (justifiable cause) applying standard statistical methodologies (3 and 4 sigma rules, Dixon's test, Shapiro-Wilk's W and generalized ESD tests) or re-running portions of the experiment, sample permitting. In this case we choose to retain all of the data points.

Missing data may be handled by generating points statistically based on extant data or using those SAS procedures (such as Mixed) that can handle such problems. "Equivalent" data requires consultation with the personnel who generated the data and calls for some judgements on the part of the statistician.

With a "clean" data set we proceed to the next phase. One of the more important aspects in studies such as these are to identify and estimate the relative con tributions to total variability from the various sources. This is done to identify and isolate those candidates for variance reductions by scientific and engineering teams. SAS/STAT$^{(TM)}$ offers a variety of procedures to do this including Proc GLM, Proc Mixed, Proc Nested and Proc Varcomp. Each method offers unique advantages under different circumstances.

Table 1.

```
data analyte;
input day instru run rep concen;
datalines;

1       1       1       1       176
1       1       1       2       204
1       1       1       3       196
1       1       2       1       197
1       1       2       2       169
1       1       2       3       178
1       1       3       1       173
1       1       3       2       175
1       1       3       3       177
1       2       1       1       179
1       2       1       2       182
1       2       1       3       170
1       2       2       1       159
1       2       2       2       179
1       2       2       3       179
1       2       3       1       182
1       2       3       2       184
1       2       3       3       183
2       1       1       1       185
2       1       1       2       189
2       1       1       3       192
2       1       2       1       190
2       1       2       2       193
2       1       2       3       192
2       1       3       1       190
2       1       3       2       200
2       1       3       3       205
2       2       1       1       186
2       2       1       2       190
2       2       1       3       191
2       2       2       1       182
2       2       2       2       184
2       2       2       3       183
3       1       2       1       183
3       1       2       2       175
3       1       2       3       185
3       1       3       1       193
3       1       3       2       198
3       1       3       3       194
3       2       1       1       182
3       2       1       2       183
3       2       1       3       188
3       2       2       1       189
3       2       2       2       191
3       2       2       3       193
```

3	2	3	1	187
3	2	3	2	191
3	2	3	3	194
4	1	1	1	195
4	1	1	2	193
4	1	1	3	199
4	1	2	1	199
5	1	1	1	192
5	1	1	2	193
5	1	1	3	193
5	1	2	1	189
5	1	2	2	191
5	1	2	3	194
5	1	3	1	192
5	1	3	2	197
5	1	3	3	199
5	2	1	1	192
5	2	1	2	194
5	2	1	3	197
5	2	2	1	192
5	2	2	2	195
5	2	2	3	197
5	2	3	1	193
5	2	3	2	190
5	2	3	3	197

```
proc mixed method = REML;
class day instru run rep;
model concen =  instru;
random day instru*day run(instru*day);
lsmeans instru/pdiff;
run;
```

Analysis

Proc Varcomp

The Varcomp procedure assumes that, unless otherwise specified in the model statement, all input variables represent random effects. The model statement will specify both the dependent and independent (effects) variables. The effects utilized may be main effects, interactions or nested effects but not continuous functions. The procedure offers four computational methodologies. For data distributions occuring in our sytems, biochemical/electronic, we find that the Restricted Maximum-Liklihood Method (REML) most useful. This method segregates the liklihood into two parts; one containing the fixed effects and the other containing no fixed effects. The procedure iterates to convergence the log-liklihood objective function for that liklihood function not containing the fixed effects. The method gives both the variance component estimates and the asymptotic covariance matrix, both useful in characterizing variability. In addition, we are never troubled by the negative variance estimates that sometimes occur with Type I and MIVQUEO. This by itself however, is no reason to select Varcomp and during this procedure we have to do assumption checking and error calculations .

The following code runs this data:

```
data analyte;
input day instru run rep concen;
datalines;
 (data here)
proc Varcomp method = reml;
class day instru run rep;
model concen = day
               instru
               day*instru
               run(day instru)
               rep (run);
                run;
```

The importance of determining the exact identity of the experimental unit may be illustrated by arranging the same data as in Table 2. Here we see that by considering each repitition to be unique, as opposed to essentially similar samples, the error is ascribed to repitition rather than to baseline error. This is a non-trivial distinction as the replicates are usually treated as non-unique.

Table 2.

Day	Instrument	Run	rep	Data
1	1	1	1	176
1	1	1	2	204
1	1	1	3	196
1	1	2	4	197
1	1	2	5	169
1	1	2	6	178
1	1	3	7	173
1	1	3	8	175
1	1	3	9	177
1	2	1	10	179
1	2	1	11	182
1	2	1	12	170
1	2	2	13	159
1	2	2	14	179
1	2	2	15	179
1	2	3	16	182
1	2	3	17	184
1	2	3	18	183
2	1	1	19	185
2	1	1	20	189
2	1	1	21	192
2	1	2	22	190
2	1	2	23	193
2	1	2	24	192
2	1	3	25	190
2	1	3	26	200
2	1	3	27	205
2	2	1	28	186
2	2	1	29	190
2	2	1	30	191
2	2	2	31	182
2	2	2	32	184
2	2	2	33	183
2	2	3	34	186
2	2	3	35	195
2	2	3	36	193
3	1	1	37	162
3	1	1	38	161
3	1	1	39	195
3	1	2	40	183
3	1	2	41	175
3	1	2	42	185
3	1	3	43	193
3	1	3	44	198
3	1	3	45	194
3	2	1	46	182
3	2	1	47	183
3	2	1	48	188
3	2	2	49	189
3	2	2	50	191
3	2	2	51	193
3	2	3	52	187
3	2	3	53	191
3	2	3	54	194
4	1	1	55	195
4	1	1	56	193
4	1	1	57	199
4	1	2	58	199
4	1	2	59	209
4	1	2	60	209
4	1	3	61	191
4	1	3	62	195
4	1	3	63	198
4	2	1	64	197
4	2	1	65	199
4	2	1	66	199
4	2	2	67	195
4	2	2	68	201
4	2	2	69	201
4	2	3	70	192
4	2	3	71	197
4	2	3	72	196
5	1	1	73	192
5	1	1	74	193
5	1	1	75	193
5	1	2	76	189
5	1	2	77	191
5	1	2	78	194
5	1	3	79	192
5	1	3	80	197
5	1	3	81	199
5	2	1	82	192
5	2	1	83	194
5	2	1	84	197
5	2	2	85	192
5	2	2	86	195
5	2	2	87	197
5	2	3	88	193
5	2	3	89	190
5	2	3	90	197

Proc Mixed

This linear model is a generalization of the standard model used in GLM and has the advantages of permitting data collection and heteroscadasticity. Variance/Covariance parameters (and therefore variance components) can be obtained from the output of the Covariance Parameter Estimates. One of the most useful features of Mixed is the calculation of appropriate standard errors for all estimable linear combinations of fixed and random effects as well as for the cooresponding F and t tests. The calculation of appropriate error terms for the least squares means represents a major advance over Varcomp, where those needed to be hand-calculated for verification. The ability to handle unbalanced data is a definite plus in clinical diagnostics where data may be lost due to a variety of factors.

In the model that follows (Table 3.) the REML method is again found useful. It is immediately noticed that the covariance estimates are quite close to those calculated by Varcomp. One difference is that here instru(ment) is declared to be a fixed effect, a truer picture or our purpose. Many times we will either use the instruments available or choose those that we feel are most representative of that population of instruments in the field.

We also note that a wide choice of methodologies are available within Proc Mixed to test assumptions. For example, we may test for heterogeneity of variance across the independent variables or request confidence limits on the estimates. This may be implemented for instru by the following:

```
proc mixed;
class statement;
model statement;
random statement / grp = instru;
lsmeans instru/diff cl;
```

This is significant where variability itself may vary across levels of a factor. In any event, these are highly interesting and important methodologies that yield accurate estimates of process variability sources.

```
data analyte;
input day instru run rep concen;
cards;
(data here)
proc proc mixed  method=reml;
class day instru run rep;
model concen = inst :
random day instru*day run(instru*day)
lsmeans instru/pdiff;

run;
```

Parameter Estimates

	Proc Mixed	
	unique	non-unique
Day	42.89	42.89
Day*Instru	0	0
Run (Day*Instru)	19.31	19.31
Residual	41.64	41.64

	Proc Varcomp	
	unique	non-unique
Day	43.06	43.07
Instru	0	0
Rep(Run)	40.71	4.16
Day*Instru	0	0
Run (Day*Instru)	18.30	19.81
Residual	0.92	36.84

John A. Wass
Abbott Laboratories
Ph: (847) 938-3675
Fax: (847) 937-2486
EMail: John.Wass@add.ssw.abbott.com

***Table 3.**