AnoGen: A Program for Generating ANOVA Data Sets

Version 1.3

Jeff Miller
Department of Psychology
University of Otago
Dunedin, New Zealand

August, 1998

Contents

1 Introduction

This program was designed for use in teaching the statistical procedure known as *Analysis of Variance* (ANOVA). In brief, it generates appropriate data sets for use as examples or practice problems, and it computes the correct ANOVA for each data set. It handles between-subjects, within-subjects, and mixed designs, and can go up to six factors, with some restrictions on the numbers of levels, subjects, and model terms.

The program can be run in either of two modes: one designed for use by students; the other, by teachers.

The student mode is simpler: The student simply specifies the experimental design, and AnoGen generates an appropriate random data set. The student can then view the data set and answers, and save them to a file. Thus, students can fairly simply get as much computational practice as they want.

The teacher mode is more complicated: The teacher not only specifies the experimental design, but also controls the the cell means and error variance to obtain whatever F values are desired for the example. Considerable familiarity with ANOVA is needed to use this mode.

2 Step-by-Step Instructions: Student Mode

1. Start the program as is appropriate on your computer system (e.g., by typing AnoGen at a DOS prompt).

2. Once AnoGen is running, type S to enter the student mode.

3. Specify the design:

   (a) To set the number of within-subjects factors, type W, and then type the number you want, followed by enter.

   (b) Similarly, type B to set the number of between-subjects factors,

   (c) Similarly, type S to set the number of subjects per group. A "group" is defined by one combination of levels of the between-subjects factors. For example, if you have between-subjects factors of Male/Female and Young/Old, then there are four groups corresonding to the four combinations.

   Note that you can set these numbers in any order, and you can change each one as often as you like. After you have the settings you want, type ctrl-Q to move on to the next step.

4. Now specify the number of levels of each factor. Type the letter corresponding to the factor you want to change (A, B, ...), and then enter the number of levels you want. Again, after you have the settings as you want them, type ctrl-Q to move on to the next step.

5. Type P to display the problem (i.e., the data set). Ideally, you would now do the computations by hand, for practice. (The information given in the problem display is intended to be self-explanatory, but some explanation is given in Section 3.)

Table 1: An example of a problem display. This design has two between-subjects factors (A and B) with two levels each, and three subjects per group.

Group A1B1:
Sub 1: 95
Sub 2: 78
Sub 3: 97
Group A2B1:
Sub 1: -19
Sub 2: -37
Sub 3: -10
Group A1B2:
Sub 1: 55
Sub 2: 64
Sub 3: 73
Group A2B2:
Sub 1: 58
Sub 2: 63
Sub 3: 71

6. Type S to display the solution (i.e., cell means, ANOVA table, etc). This is where you check your solution and make sure you've done it correctly. The solution contains the various parts that I use in teaching ANOVA using the general linear model. (More explanation of the information given in the solution display is given in Section 4.)

7. If you want, type F to save the problem and solution to a file. (The main reason to for doing this is to get a printed version of the problem and solution.) Enter the name of the file to which you want the information saved.

8. Type ctrl-Q to quit when you are done with this problem. AnoGen will then ask if you want to start over again: Type Y if you want to do another problem, or N to quit.

3 Explanation of Problem Display

Table 1 shows an example of a problem display. There is one line per subject, and the different groups correspond to the different levels of the between-subjects factor(s). For this example, the problem display fits on a single screen; with larger designs (i.e., more groups or more subjects per group), the problem display may be split across several screens.

Table 2 shows an example of a problem display for a more complex experimental design. Note that the different conditions tested within-subjects are listed across the line, and the different subjects and groups organized as in the between-subjects design.

Table 2: An example of a problem display. This design has a within-subjects factor (A) with two levels, two between-subjects factors (B and C) with two levels each, and three subjects per group.

Group B1C1:

	A1	A2
Sub 1:	77	53
Sub 2:	84	56
Sub 3:	103	41

Group B2C1:

	A1	A2
Sub 1:	77	65
Sub 2:	54	64
Sub 3:	73	69

Group B1C2:

	A1	A2
Sub 1:	103	75
Sub 2:	100	78
Sub 3:	97	57

Group B2C2:

	A1	A2
Sub 1:	72	10
Sub 2:	74	18
Sub 3:	58	2

4 Explanation of Solution Display

The solution display has several components, as described below. Some of these components may be omitted if they do not fit well with the way your instructor teaches the material.

4.1 Design

This shows a list of factors with the number of levels per factor. Also shown is the number of subjects per group.

4.2 Cell Means

The cell means are given in a table of this form (these are the means for the problem in Table 2):

Cell:	Mean
u	65
A1	81
A2	49
B1	77
B2	53
A1 B1	94
A1 B2	68
A2 B1	60
A2 B2	38
C1	68
C2	62
A1 C1	78
...	

The first line (u) shows the overall mean across all conditions. The next two lines (A1 and A2) show the means of all scores at levels 1 and 2 of factor A, respectively. The next two lines (B1 and B2) show the means of all scores at levels 1 and 2 of factor B, respectively. The next line (A1 B1) shows the mean of all scores at level 1 of factor A and level 1 of factor B, and then the next three lines show means for the other combinations of levels on these two factors. And so on.

4.3 Model

The model section shows the form of the general linear model appropriate for this design. The main effect and interaction terms are denoted by capital letters (A, B, AB, etc), S is for subjects, and the subscripts are denoted by lower-case letters (i, j, k, etc).

4.4 Estimation Equations

The estimation equations section shows the equation used to estimate each term in the linear model. The period subscript is used to denote averaging across levels of the factor corresponding to that subscript.

4.5 Decomposition Matrix

The decomposition matrix shows the breakdown of all data values (numbers of the left sides of the equals signs) into the estimated values corresponding to each term in the linear model. The order of the numbers on the line is the same as the order of the terms in the model.

4.6 ANOVA Table

The ANOVA table is in a relatively standard format. The F value is marked with one asterisk if it is significant at the level of $p < .05$ and two asterisks if significant at $p < .01$. The error term used to compute each F is shown at the far right side of the table.

5 Background for Teachers

This section provides a brief review of the general linear model for ANOVA, intended for teachers who already have some background in this area. Besides refreshing the relevant concepts, this section is also intended to give some hints on how to generate desired patterns of data and to clarify the notation used in the program.

The model underlying ANOVA assumes that each data value (Y) is a sum of additive components due to main effects, interactions, and error. For example, the model for a two-factor between-subjects design is often written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where α ("alpha") and β ("beta") are the main effects of the two factors (rows and columns), γ ("gamma") is the interaction effect (associated with each cell), and e is a normally distributed random error associated with data value.

From a given set of data, ANOVA (implicitly or explicitly) estimates the numerical values of all the terms in the model. For example, these cell means for a 2×2 design:

	Factor A	
Factor B	Level 1	Level 2
Level 1	33	21
Level 2	35	31

yield these estimates for the model's parameters (notation: estimates are written with "hats" above them).

$$\begin{aligned}
\hat{\mu} &= \frac{33+35+21+31}{4} = 30 \\
\hat{\alpha_1} &= \frac{33+35}{2} - 30 = 4 \\
\hat{\alpha_2} &= \frac{21+31}{2} - 30 = -4 \\
\hat{\beta_1} &= \frac{33+21}{2} - 30 = -3 \\
\hat{\beta_2} &= \frac{35+31}{2} - 30 = 3 \\
\hat{\gamma_{11}} &= 33 - 30 - 4 - (-3) = 2 \\
\hat{\gamma_{12}} &= 35 - 30 - 4 - 3 = -2 \\
\hat{\gamma_{21}} &= 21 - 30 - (-4) - (-3) = -2 \\
\hat{\gamma_{22}} &= 31 - 30 - (-4) - 3 = 2
\end{aligned}$$

Remember, however, that these parameters are constrained relative to one another. Specifically,

$$\begin{aligned}
\hat{\alpha_1} &= -\hat{\alpha_2} \\
\hat{\beta_1} &= -\hat{\beta_2}
\end{aligned}$$

$$\begin{aligned} \hat{\gamma_{11}} &= -\hat{\gamma_{12}} \\ &= -\hat{\gamma_{21}} \\ &= \hat{\gamma_{22}} \end{aligned}$$

In essence, the linear model decomposes each cell mean into its components like this:

	Factor A	
Factor B	Level 1	Level 2
Level 1	33 = 30 + 4 - 3 + 2	21 = 30 - 4 - 3 - 2
Level 2	35 = 30 + 4 + 3 - 2	31 = 30 - 4 + 3 + 2

Individual observations are formed by these same sums, plus the random error term.

5.1 Generating Data

The program generates data by letting the user specify the values of the parameters of the general linear model. (These values are generated randomly in student mode.) The data values are simply computed by summing up the specified values and then adding in appropriate random error components. The error components are not completely random, however, but are constrained to add to zero, so that the cell means will come out exactly as desired. (This is not intended to be a tool for simulation, but only for teaching.)

It takes a bit of trial and error to generate a desired pattern of cell means by specifying the terms in the model, but this gets easier with a little practice. The basic strategy is to think not in terms of individual cell means but rather in terms of the parameters of the model. Looking at the four cell means for the example above, you must think in terms like these:

- I'd like an overall average of about 30, so I'll set $\mu = 30$.
- I'd like the average at A_1 to be about 4 points higher than than the overall average, so I'll set $A_1 = 4$. This will automatically make A_2's average 4 points lower than the overall average, and the difference between A_1 and A_2 will thus be 8 points.
- I'd like the average at B_1 to be about 3 points lower than than the overall average, so I'll set $B_1 = -3$. This will automatically make B_2's average 3 points higher than the overall average, and the difference between B_1 and B_2 will thus be 6 points.
- I'd like a small interaction to raise scores at AB_{11}, so I'll set $AB_{11} = 2$. Because of the constraints, this will also automatically raise the average for AB_{22} and reduce the averages at AB_{12} and AB_{21}.

5.2 Notation

The notation used in this program is suitable for a variety of multi-factor designs. The main effect of each factor is denoted by a capital letter (A, B, C, ...), and the factor's levels are denoted by a lower-case subscript (i, j, k, ...). Thus, for the previous example the program uses A_i and

B_j instead of α_i and β_j, respectively. Interactions are denoted by concatenating the letters (and subscripts) representing the interacting factors. For example, the program uses AB_{ij} instead of γ_{ij} to represent the interaction in a two-factor design. In a three-factor design, there would be three two-way interactions (AB_{ij}, AC_{ik}, and BC_{jk}) and one three-way interaction (ABC_{ijk}).

The notation for error terms always involves the letter S (for subject). By convention, the program lists the between-subjects factors in parentheses after the S as part of the subjects term. For example, the error term for the example two-factor design would be $S(AB)_{ijk}$, because factors A and B are both between-subjects factors. In repeated-measures designs, there are also various "subject by treatment" interaction error terms, and these are denoted as interactions with the treatment factor(s) listed before the S. For example, an AS_{ij} term would be a subject by treatment (A) interaction term for a one-factor repeated measures design. In mixed designs the between-subjects factors are always carried along as part of the S term. For example, with factor A a within-subjects factor and factor B a between-subjects factor, the error term for factor A is the $AS(B)_{ijk}$ interaction term.

6 Instructions for Teacher Mode

Teacher mode is more complicated to use, and these instructions are not step-by-step. It is assumed that you have already used and understand the Student mode, and that you have some familiarity with the general linear model approach to ANOVA. (Section 5 provides a brief review for those who wish a refresher on this approach to ANOVA.)

To begin, start the program by typing AnoGen, and type T to enter Teacher mode. Next specify the design and the number of levels of each factor just as in the student mode.

6.1 Generating the Desired Data Pattern

Next you have to specify the values of the terms in the linear model, using a display like the following, which is the display for a 2-factor repeated measures design (both factors A and B are within-subects; A has 2 levels and B has 3 levels):

```
Specification of Terms in Linear Model:            Page 1 of 1

 Num   Source        MS       F        ET     Estimates
----   ------    -------- -------     -----   --------------------
----->
  0:   u              0.0 1000.00***  S        0
  1:   A              0.0 1000.00***  AS       0
  2:   B              0.0 1000.00***  BS       0 0
  3:   AB             0.0 1000.00***  ABS      0 0
  4:   S              0.0                      random
  5:   AS             0.0                      random
  6:   BS             0.0                      random
  7:   ABS            0.0                      random
```

```
Type green number to change estimate(s) for corresponding source, or
type C, N, or P to view Cell means, Next page, or Previous page, or
type ^Q to proceed:
```

Note: The numbers in the "Num" column at the far left should appear in green.

Each line numbered 0–7 corresponds to one ANOVA source (i.e., one term in the linear model), as identified in the column labelled "Source." The current numerical value(s) corresponding to that term in the model are listed in the "Estimates" column, and the current mean square and F for that term are listed in the MS and F columns. The F value is marked with one asterisk if it is significant at the level of $p < .05$ and two asterisks if significant at $p < .01$. The error term used in computing each F is listed in the ET column.

To change the numerical value of a model term, you type the green one-digit number next to it (in the "Num" column). For example, if you want to change the overall mean (u), type 0, and you can then enter a new value for u at the bottom of the screen. The new value you enter will then appear in the "Estimates" column, and the MS and F for that source will be updated using the new value.

Note that there are two numbers listed in the "Estimates" column for the sources B and AB. This is because factor B had 2 degrees of freedom; the first estimate is the value for B1, and the second is the value for B2.

Estimates are handled slightly differently for the error terms (indicated by the word "random" under "Estimates"). When you type its "Num" to change one of these, the program will ask you for a "Maximum", and then it will generate random numbers from -Maximum to +Maximum to use as values for these random terms (one per df). The MS is recomputed using the randomly generated terms, and F's are updated as appropriate.

Note that you can inspect the cell means that would be obtained with the current estimates by typing C at any point in this process.

"Next page" and "Previous page" come into play when the model has more than 10 sources, so the terms are listed across multiple screens.

When you are happy with the cell means and F's, type ctrl-Q to continue on to the next phase.

6.2 Viewing the Generated Data

After the data have been generated, you can look at the results and save them to disk. (As in student mode, printing is accomplished by saving output to a file and then printing that file from outside AnoGen.) Here, you have more flexibility than the student. The following display lists your possible actions at each point:

```
which of the following actions do you want?
-------------------------------------------
    write Raw data
    write Cell means
    write Anova table
    write Factor list
    write Model
    write Estimation equations
```

```
    write Decomposition matrix
    write effecTs codes
    write Notepad line
    write Input file
    set Options

to select an action, type its capitalized letter, or ^Q to proceed :
```

Select the desired action by typing the capital letter in its description (e.g., "F" to display the factor list).

Seven of the write actions (i.e., *Raw data, Cell means, Anova table, Factor list, Model, Estimation equations, Decomposition matrix,* simply display the same sorts of information described in connection with the student mode.

Three rarely used actions are:

write effecTs codes This writes out a matrix of dummy variables that could be used to code the factor effects and interactions. It is useful if you want to illustrate how to do ANOVA with a regression program.

write Notepad line When the disk file is open (see "set Options", below), this action allows you to type comments into the file. If you type N, AnoGen will simply copy whatever you type directly into the output disk file, ending when you type escape *as the first character on a new line*. This is useful if you make up several problems in one run, so that you can note in the output file when you are starting a new problem.

write Input file This action writes out the raw data for analysis with another program. You are given the option of saving the data in a tab-delimited file (select by typing T) or in the format of an input file for the ANOVA program MrF (select by typing M). MRF is a freely available program that does equal cell-size ANOVAs for between-subjects, within-subjects, and mixed designs. Where appropriate, it provides both the traditional p level and the Geisser-Greenhouse corrected p level for the computed F's. As of February 1998, the latest version was 1.0, available as the file MRF1_0.zip from simtel and garbo (and perhaps some other ftp sites) as:

```
    ftp://garbo.uwasa.fi/pc/stat/Mrf1_0.zip
    http://www.simtel.net/pub/simtelnet/msdos/statstcs/Mrf1_0.zip
    ftp://ftp.simtel.net/pub/simtelnet/msdos/statstcs/Mrf1_0.zip
```

6.3 set Options

In teacher mode, you can also set several program options, as listed here:

```
OUTPUT OPTION:          CURRENT SETTING:
--------------          ----------------
P = decimal Places      :      0
F = disk File           :    Closed
```

```
D = output Destination  :    Screen
W = Widths of terms     :    variable
O = Output format       :    ASCII

Type capital letter to select option to be changed, or ^Q to proceed:
```

P = decimal Places AnoGen only accepts integer values for model terms set by the user, and it only generates integer values for random model terms. Nonetheless, in some situations the data are inherently real numbers with decimal places, and this option is provided to handle those cases. Here is how it works:

When you choose the P option, you can type in a number 0, 1, 2, 3, ... to indicate how many decimal places your data values should have. Then, the decimal point is shifted that many places to the left in all of the data values previously generated. For example, if you specified a mean of $\mu = 150$ when you were constructing the data and then use the P option to indicate that you want two decimal places, then the new mean after decimal point adjustment will be $\mu = 1.50$. The sums of squares will be adjusted accordingly.

F = disk File This allows you to open and close a disk file. When the file is open, everything that is displayed on the screen is also echoed into the file, usually so that you can later print it. (The next option tells how to turn this echoing off.) If you select "F" when the disk file is already open, it will be closed.

D = output Destination When the disk file is open, this is a toggle between "Screen" and "Screen + File". This option is provided so that you can avoid writing stuff into an open output file until you have seen it and are sure you are happy with it. **In summary, print by: (1) select the disk file option to make an output file; (2) display to the screen that which you would like to print, and it will also be written to the output file; (3) quit AnoGen and read the output file into your word processor for printing.**

W = Widths of terms This option allows you to set the number of columns (spaces) used in writing out each term in the model. By default, AnoGen computes the number of columns for each term from the width of that term (e.g., ABij is four characters wide), but you can adjust these widths if you want (usually, to pack more on a line).

O = Output format The default format for output is ASCII. You can set this to LaTex instead; if you use that word processor, this will allow you to print nicer equations and tables. If you do generate LaTex output, you will need the following commands that are used:

```
\newcommand{\anovaheader}{Source & $df$ & $SS$ & $MS$ & $F$ & Er-
ror Term \\ }

\newcommand{\startanovatable}{
      \begin{center}
      \begin{tabular}{|lrrrrc|}
      \hline
      \anovaheader \hline   }
```

```
\newcommand{\stopanovatable}{
      \hline
      \end{tabular}
      \end{center}  }
```

6.4 Restarting

When you are done viewing the generated data, you have several options described in this screen:

```
Restart from:
  1:  Numbers of factors and subjects.
  2:  Numbers of levels of each factor.
  3:  Estimates of variance sources.
  4:  More output of same data.

       or ^Q to exit program :
```

You can actually quit the program with control-Q, or you can restart it from various points if you want to do some more work, as follows:

1 This restarts from the beginning of the program.

2 This preserves the current numbers of within- and between-Ss factors and subjects, but allows you to change the number of levels and construct new data from there.

3 This preserves the numbers of levels as well as the number of factors and subjects, but allows you to generate a new data set with the same design.

4 This allows you to go back and view further the same data you just generated. It is only useful decide to quit but then change your mind.

7 Run-time Options

There are a few AnoGen options that I expect any given user will *always* want to set the same way. To facilitate this, these options are set on the command line or in an environment variable. As illustrated below, once you determine how you would like the options to be set, you can set them and forget them in either of two ways as described in detail below.

7.1 Total df and SS option

By default, AnoGen defines the total degrees of freedom as the total number of observations minus one ($N - 1$) and the total sum of squares as the sum of square differences from the grand mean. Some teachers and computer programs, on the other hand, use uncorrected totals, defining the total df as N and the total sum of squares as the sum of the squared data values. To make AnoGen conform to the latter style, you can invoke AnoGen with the optional command-line parameter CORRECTED-, like this:

```
C:\STATS> AnoGen CORRECTED-
```

7.2 Optional Output for Student Solutions

By default, AnoGen writes some formidable output when a student asks for the solution: in addition to the cell means and summary ANOVA table, it writes the appropriate version of the general linear model, the estimation equations, and the decomposition matrix. Depending on how you teach ANOVA, this may be more information than your students will be comfortable seeing (one instructor said that all these equations would throw his undergraduate students "into a state of terminal anxiety"!).

To simplify the output for your students, you can turn off any combination of these three types of solution output by specifying the desired combination of these three parameters on the command line:

```
C:\STATS> AnoGen MODEL- ESTEQ- DECOMP-
```

These parameters can be specified in any combination and any order, with or without the option CORRECTED-, and they are not case sensitive.

7.3 Setting Options with Batch Files or Environment Variables

Once you have decided which of the options you want for yourself and/or your students, one easy way to make sure they are always selected is to set up a batch file to invoke AnoGen. For example, you could create a file MyAnoGen.Bat with just the line

```
AnoGen corrected- esteq- decomp-
```

Then, at your command prompt you would type MyAnoGen instead of AnoGen, and AnoGen would be started with the appropriate settings.

Alternatively, you can use an environment variable to invoke your preferred options. Inside your autoexec.bat file, include a line like

```
SET ANOGEN= corrected- esteq- decomp-
```

to select the options you like. Then, you can just invoke anogen at the command line without any parameters, and these parameters will be taken as if they had been typed at the command line.

If both environment variables and command-line parameters are used, the latter take precedence.

8 Sample Acknowledgement Letter

I don't want much, just some feedback on who is using AnoGen and what they are using it for. Something like the letter shown below would be fabulous. I'd prefer a real signed letter on paper, but acknowledgements by email would be better than nothing (email address: miller@otago.ac.nz). Of course I would also welcome bug reports and suggestions for improvement, too, although I can't promise any fast action on those. Don't forget what you paid for this!

Prof Jeff Miller
Department of Psychology
Univ of Otago
Dunedin, New Zealand

Dear Prof Miller,

This is to acknowledge that I have used the computer program AnoGen for teaching/studying the analysis of variance during the past year.

Include all of the following that apply, and any other uses that I haven't thought of: I have used it for generating practice problems, homework assignments, exam questions, for myself, for my class of 50 upper-division Psychology students at the University of …

Sincerely,
etc