

PLABSTAT

A computer program
for statistical analysis
of plant breeding experiments

Version 3A of 2011-06-14

H.F. Utz

Institute of Plant Breeding,
Seed Science and Population Genetics
University of Hohenheim, D-70593 Stuttgart, Germany
Email: utzf@uni-hohenheim.de
<http://www.uni-hohenheim.de/~ipspwww/soft.html>

The most important changes of version3 compared to version2

1. Omit primes enclosing statement names.
2. Text after ! is interpreted as comment
3. Input files of version2 are transformed internally to version3 and than analysed.

Contents

1	What PLABSTAT can do	3
2	A Simple Example	4
3	General Usage	6
3.1	Basic Rules	6
3.2	The General Statements	7
3.3	The Input of Data	12
4	Calculation of Basic and Regression Statistics	12
4.1	Statements	12
4.2	Output and Methods	14
5	Analysis of Randomized Complete Block and Lattice Designs	15
5.1	Statements	15
5.2	Output	17
5.3	Methods	18
6	Analysis of Multi-factorial Experiments	19
6.1	Statements	19
6.2	Output	25
6.3	Methods	26
7	Unreplicated Experiments	27
7.1	Statements	28
7.2	Output	29
7.3	Methods	30
8	Supplementary Statements	31
9	Hints, Faults, and Limits	35
10	References	37
APPENDIX A	List of Statements	40
APPENDIX B	Examples	43
APPENDIX C	MODEL Lines for Several Designs	44

1 What PLABSTAT can do

PLABSTAT may be read as PLAnt Breeding STATistical program or alternatively as PLAnned Block experiments and their STATistical analysis.

In the age of wide-spread statistical computer packages the question arises if special programs for statistical analyses of agronomy and plant breeding experiments should be written.

The present program is concise and includes only the most important routines for the analysis of plant experiments. It is easy to handle, generates compact output, checks on outliers, and estimates genetic variances and correlations, or heritability routinely.

As a program is in continuous progress and can also never comprise all statistical methods, PLABSTAT is designed such a way, that changes to more sophisticated software packages, e.g. GENSTAT are easy. PLABSTAT does not include a routine for the data input. Therefore an editor program, e.g. KEDIT, THE, EMACS, or one of the group SCITE, Crimson Editor, UltraEdit, ConTEXT, or WORD and EXCEL or a database program, must be used.

The analyzing capabilities of PLABSTAT are summarized in the following.

Basic Statistics and Multiple Regression

PLABSTAT estimates basic statistics, e.g. variance, skewness, kurtosis, and coefficients of correlation. Simple or multiple linear regression including path coefficients are also calculated.

Lattice Experiments

PLABSTAT analyses lattice designs, i.e. square, rectangular, or generalized lattices (PATTERSON and WILLIAMS, 1976) in each case without repetitions. For lattice designs it calculates adjusted mean values or effects, and gives a list of the incomplete block effects. The residuals are checked on outliers. Augmented partially replicated designs are analysable.

Multi-factorial Experiments

PLABSTAT performs analyses of variance for balanced data sets. Missing data are allowed. In case of very unbalanced data sets only least squares means are estimated. The program processes factorial and nested classifications, with random, fixed or mixed models. Partially balanced data, like latin squares, can be handled with the help of a "rearrangement" (see HARTLEY, 1956).

ANOVA tables, including components of variance, their respective standard errors and F tests as well as the standard errors and least significant differences for mean values are computed. Tables of mean values or effects are printed. Residuals and effects are checked on extreme values.

Heritabilities, coefficients of genetic correlations, or covariance analyses can be performed. The subdivision of a two-factor interaction table according to a regression

model is possible.

Unreplicated Experiments

Experiments without replications can be analysed if standard treatments exist or using moving averages.

Other Procedures

Scatter plots may be printed. Means (adjusted in lat- tices) are written on a file for further analyses.

For creating new variables or modifying variables, the program works with an easy-to-use code. Observations can be proofed, wether they exceed certain limits.

2 A Simple Example

The input of PLABSTAT is demonstrated in a simple example. In a recurrent selection program the yield of six consecutive cycles of maize may be recorded. To get an idea about the yield improvement due to selection you may want to sketch a scatter plot. The statements and the data arrangement are as follows:

```
!!      Yields of maize in 6 cycles
!      Input of grain yield in kg/ha
BASIC  6  0 2 2
VARIABLE_NAMES  CYCLE YIELD
CALCULATE  X2=X2/100    ! Yield in q/ha
PLOT/2  X2.X1
RUN
1 4758
2 4970
3 5053
4 5376
5 5429
6 5607
EOD
STOP
```

In the first part you find the PLABSTAT statements, the statement names are given in capitals. In the second part, the data are written in six rows with the cycle number and the yield in each cycle. All statements and data are in free format. Thus the numerical values as well as the parameters in statements must be separated by at least one blank.

After a ! a comment is given until end of line. The first comment line is interpreted as the title of the experiment to be analyzed. After the statements comments can be added, as you can see in the CALCULATE line.

The statement BASIC tells the computer that data rows have to be read in. The size of the data set is specified with BASIC: in this example 6 rows consisting of 2 columns or variables each. The second "2" in the BASIC line indicates that two variables should be analyzed. If new variables shall be created by means of the CALCULATE statement

the last number must be increased accordingly. The second number, i.e. 0, will be discussed later in 4.1.1.

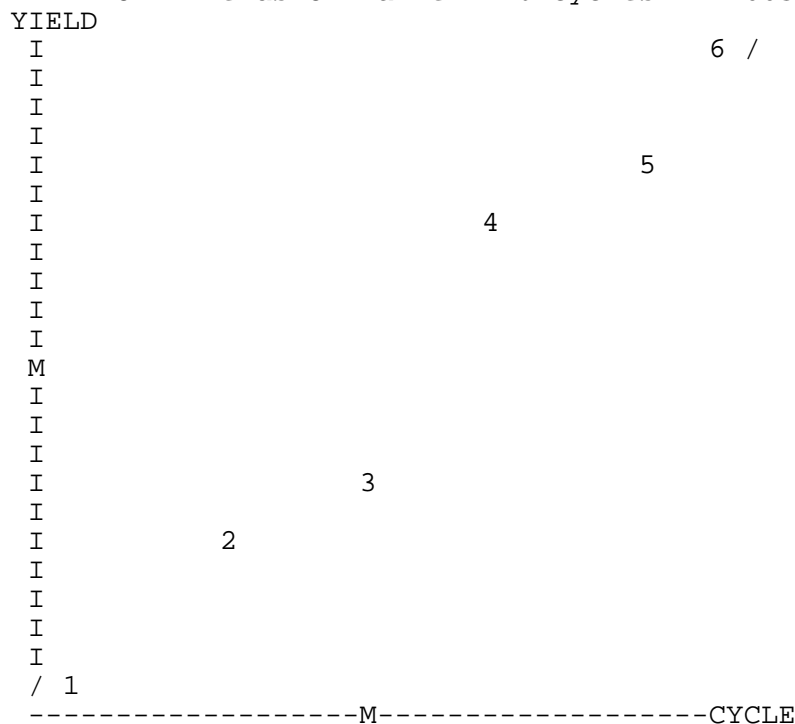
With the statement `VARIABLE_NAMES` the names for the two variables are defined. In the `CALCULATE` statement the second variable is divided by 100, to get the yield in quintels per hectare.

```
PLABSTAT   VERSION 3A           of 2003-08-16
INSTITUT FUER PFLANZENZUECHTUNG, UNIVERSITAET HOHENHEIM
At 2003/08/19   17h 54m 10s
```

```
!!      Yields of maize in 6 cycles
!      Input of grain yield in kg/ha
BASIC  6  0 2 2
VARIABLE_NAMES  CYCLE YIELD
CALCULATE  X2=X2/100    ! Yield in q/ha
PLOT/2  X2.X1
RUN
```

```
      CYCLE  YIELD
MIN    1.00  47.58
MAX    6.00  56.07
```

```
SCATTER PLOT  Yields of maize in 6 cycles      2003/08/16 12h 09m
```



```
YIELD: Min 47.6  Max 56.1  Mean 52.0  s.d. 3.2  b(Y.X) 1.6986
CYCLE:   1.0    6.0      3.5    1.9  b(X.Y) 0.5728
                                Corr. 0.986
```

```
EOD
STOP
```

Using the statement `PLOT` a scatter plot of the second variable against the first one is printed. In more detail, the statement is `PLOT/2` with the qualifier 2 after a slash. Such qualifiers are optional. In this case it causes that the diagram is printed on half a page and consequently two diagrams will fit on one page. If the qualifier is omitted the scatter diagram will use a full page.

With the RUN line, the PLABSTAT statements are finished transitorily, and the data follow. The end of data is announced by EOD, the abbreviation of End Of Data. The STOP statement indicates that the job is finished and no other statement follows.

Now the output for the small example may be discussed:

The statement lines are listed first. The program reads the statements until a RUN statement comes. After RUN the statements are executed.

If the computer detects errors in reading a statement line he gives a message out, e.g. if you have written the BASIC line like this BASIC 6 0 K 2 then the corresponding lines in the output are:

```

      BASIC      6      0      K      2
      ***      WRONG  PARAMETER  VALUE

```

So you can alter the line and set the correct value "2" instead of K.

If the data are read in, the two lines labeled with MIN and MAX are printed out. For each variable the minimum and maximum value are given.

In the diagram the means are symbolized with the letter M on both axis. The two extreme points of the regression line, YIELD on CYCLE number are marked with slashes /. Therefore, the regression line can easily be drawn by hand. The points are numbered consecutively, in the same order as the data. So the experimenter can identify his points without any complication.

Below the scatter diagram you can find some other estimates of the two variables: the mean, the standard deviation (s.d.), the regression coefficients of X2 on X1 and of X1 on X2, and the correlation coefficient between the two variables.

3 General Usage

PLABSTAT works with about thirty statements. This chapter tells you how to write PLABSTAT statements, begins with a description of the more general statements, and comments how the data sets must be constituted.

3.1 Basic Rules

3.1.1 PLABSTAT statement names must be given in capitals. A statement line can obtain up to 255 characters.

3.1.2 PLABSTAT uses only the first three characters of a statement name. The other letters in a statement name are not relevant. For readability it is recommended not to abbreviate the statement names too much. So VARIABLE_NAMES can be written as VAR or VARIABLES or VAR_LABELS.

3.1.3 Each statement must begin on the first column of a line. If more than one line is used, the following lines must start with at least one blank, e.g.

```

      VARIABLE_NAMES      CYCLE
                          YIELD

```

- 3.1.4** Statements can mostly be used in any order. A restriction is, e.g., that PLOT can be used only after a BASIC, LATTICE, ANOVA, or UNREP, statement, since data must be read in first. It is recommended to write the statements down in a logical order.
- 3.1.5** At least one of the statement BASIC, LATTICE, or ANOVA must be given in each PLABSTAT run.
- 3.1.6** Qualifiers are given after a slash, e.g. ANOVA/12. With a qualifier you can modify default values in the analysis or give other specifications. If the qualifier which is read consists of less than defined digits, only the given first are altered compared with the default.
- 3.1.7** If a statement is written by mistake, more than once, only the last statement will be used.

3.2 The General Statements

3.2.1 Comments after !

Comments start with an exclamation mark or the character ! and end with the end of line. The title of an experiment or other specifications may be given.

! any TEXT

It is recommended to start the first comment line with !! since this line contains mostly the name of the experiment and for !! can be searched easily. The first comment is used also for title in mean tables and diagrams.

Comments can be given after the end of statements, e.g.

```
CALC X10 = X2/X3*100  ! Standability in percents
                      ! whereby X2 = number of erect plants per plot
                      !           X3 = number of existing plants per plot
```

3.2.2 The INPUT Statement

With the INPUT statement variables can be defined which are read in and afterwards addressed by names in CALCULATE, PLOT, RESTRICT, FIT, or TBT_TABLE statements:

INPUT list of variables

If INPUT is not used variables are addressed by X1, X2 X3,

Names of variables must start with a letter, more exactly with a character whose ASCII decimal code is greater or equal 64. Upper-case and lower-case letters were treated as different. Only the first five characters are of importance. Further characters can be appended as a form of commenting but are not considered. Thus Yield and Yield_g/plot is equivalent.

Reserved names which have meanings within CALCULATE, i.e. NOTE, EXP, LN, SQRT, ARCSINST, can not be used. The arithmetic special characters like + - * / : () and some others \$! ? > < = . ; are not allowed in names. Exactly all characters with ASCII decimal code 40-47 and 58-63 are not allowed. But other special characters like % ^ [] { } can be used in names.

3.2.3 The CALCULATE Statement

With the CALCULATE statement you can create new variables, modify existing variables or obtain observations which exceed a certain limit. With respect to CALCULATE statement the following restrictions have to be considered:

- (1) Variables are addressed by `x1`, `x2`, `x3`, ... or by the names declared by INPUT. Furtherhin auxiliary variables can be defined, e.g. `x50`, `Dummy1`, or `plot_size`.

All operations are conducted in real mode. An example for a CALCULATE statement is:

```
CALCULATE X2=X1+X2+X3-100
CALCULATE plot_size=3.5      yield=weightt/plot_size
CALC  \qqquad DMC\%=dry_weight/wet_weight*100
```

- (2) A CALCULATE line might contain several assignments separated by blanks. Hence blanks within an assignment are not allowed. Various assignments on the same CALC line are carried out from left to right.

```
CALC X2=X6/X7*100.0      X3=X8*X9/X10
```

- (3) The fundamental operations of arithmetic, i.e. addition, subtraction, multiplication, and division `+`, `-`, `*`, `/` are feasible. Moreover the integer division (`:`) is possible, too, truncating the decimal places; the statement `CALC X1=5:3` e.g. results in 1. Assignments in parentheses are possible whereas powers are not allowed.

```
CALC X5=X1/(X2-X3)*(-X4)      X5=(( -X3+X4)/X1)-X2.
```

- (4) The possible functions are

```
LN          = natural logarithm
SQRT        = square root
EXP         = exponential function
ARCSINST    = arc sine of square root of X (with 0 < X < 1)
```

Using these functions the variables must be given in parentheses:

```
CALC X3=LN(X2)  X5=ARCSINST(X5)  X3=X1+EXP(-1/X2)
```

- (5) Using the NOTE operator the observations for the individual variables are checked on values exceeding a certain limit. In this connection only boolean expressions with "less than" or "greater than" are allowed, e.g.:

```
CALCULATE NOTE X6<X7      NOTE X7>99.9
```

- (6) Comments can be given after the end, e.g.:

```
CALC X10=X2/X3*100 ! percent erect plants
```

- (7) Powers are not allowed whearas terms in parentheses are possible. More complex formulas shall be dissect in pieces, see [psdlatt.dat](#) and [psdanov.dat](#).

```
CALC Z0=factor*(Sol1p-VPF)+VPF  Z1=LN(2*X1*X1*X1)
```


(8) Conditional statement can be used like:

```
CALC X15=X1>0.5 X16=X3==100
```

whereby the following relational operators are allowed:

```
< less
> greater
== equal
<= less or equal
>= greater or equal
<> unequal
```

A value of 1 result during execution if the comparison is true, or 0, if false. In the case of missing data a missing value result.

This operators can be used in expressions, e.g.:

```
CALC X21=X14+(X1>=0.5)*10
```

i.e. it is added always 1 to X14 if X1 is greater or equal 0.5.

```
CALC X22=X14+X1>=0.5*10
```

i.e. it is added always 1 to X14 if X1 is greater or equal 0.5*10 or 5.

Notify the sequence in working out of the expressions whereby first * and : , than + and - , finally the comparisons above are performed.

Note to calculations with zero:

A division by zero or a natural logarithm of zero results in zero again. Using the qualifier MISS=0, i.e. zeros are not considered as missing values then $100 + 0 = 100$. However, setting the qualifier MISS to 1, i.e. zeros are regarded as missing values, then all arithmetic operations including a zero result in zero, e.g. $100 + 0 = 0$, what means a missing value.

3.2.4 The VARIABLE NAMES Statement

With this statement you can specify the names for the variables. These names are printed in the head of tables.

```
VARIABLE_NAMES list of variables
```

The number of names must correspond to NVAR, the number of variables to be analyzed. Only the first five characters of each name are printed in the table heads. However, for documentation purposes more characters can be used. As usually the names are separated by one or more blanks. Signs of equality, blanks, or exclamation marks must not be used within a variable name.

```
VARIABLES Yield Dry_matter Height
```

If the statement INPUT was set the VARIABLE NAMES gives the variables to be analysed and the order in the analysis and tables. In this case the variable names must be congruent with those of INPUT or CALCULATE with certain restrictions in names see [3.2.2.](#)

3.2.5 The NAMES_OF_TREATMENTS Statement

This statement is used for naming the treatments. These names are printed in the heading of the mean tables.

```
NAMES_OF_TREATMENTS/Q list of texts
```

The number of names must correspond to the number of treatments. Omitting the qualifier, only the first five characters of each name are used. Characters exceeding the number of five are skipped while with less than five characters the names are filled with blanks at the end. Only the first four NAMES lines are listed in the report later on. Signs of equality, blanks, or exclamation marks must not be used within a treatment name.

If you desire more than five characters for the names the number can be specified using the qualifier Q. Note, that the number of characters has to be specified with the first NAMES statement. For indicating the factor, for which names are meant the character symbolizing the respective factor has to be included in the qualifier too if ANOVA is called. Examples may be:

```

NAMES_OF_TREATMENTS      D1    D2    D3
NAMES      E1*D2_(EMMA.LENA)*(DORA.ROSA)
              M1*O2_(MARY.LENA)*(ORA_.ROSA)
NAMES/T9      Manchuria    Svansota    Velvet

```

Of course, in the second example only E1*D2 and M1*O2 are used by PLABSTAT.

3.2.6 The CHECKS Statement

With the CHECKS statement the check or control treatments are defined and the treatment values are given as percent of the check mean in an additional table:

```
CHECKS/Q list of treatment identifiers
```

Identifiers are the numbers or names of checks. If a star * is given, the treatment values are expressed as percent of the general mean. If CHECKS is used in connection with an ANOVA call the character symbolizing the factor to which the check entries belong to as treatments has to be written as qualifier (see FACTOR statement 6.1.2). If the LATTICE statement is used, the qualifier Q is omitted. For example:

```

CHECKS      10  20  30          (with LATTICE)
CHECKS/T      24  25          (with ANOVA)

```

To analyse series of experiments with often additional local checks, a qualifier +n can be used. Only the first n check treatments are taken for calculating relative values, e.g.:

```
CHECKS/+2 TASSILO SYMPHONY DUMMY1 DUMMY2 DUMMY3
```

The later local checks DUMMY1, DUMMY2, DUMMY3 may be missing in certain experiments. The first n checks are shown in the scatter diagrams by their numbers, afterwards the treatments to be tested, and at the end the local added checks. In multi-factorial experiments, e.g., in series of tester×lines, checks introduced with CHECK/+ or CHECK/+n are not included in ANOVA, those were added only in the scatter diagrams.

3.2.7 Der Befehl csv

With the CSV statement, csv (comma separated values) files can be produced which can be read easily from data bases. Four csv files are written in total: *_mean.csv with the treatment means, *_perc.csv with the relative values as percents from check mean, *_anov.csv with parameters from the analyses of variance (e.g., degrees of freedom, sums of squares, mean squares, components of variance, F values), and *_corr.csv with the correlation matrix of treatment means. Besides the *.mnv file is also a sort of a csv file with a blank as delimiter. Using the ANOVA statement one and two way table of means can be written in the *_mean.csv and *_perc.csv files calling the TBT_TAB statement. English or German format of csv files can be chosen.

```
CSV/Q      list of five identifiers
```

or

```
CSV/Q      Calculation_ID,Crop,Trial,Year,Location
```

With the qualifier E or without qualifier, the English format with the decimal point and comma as separator is used in the output. With the qualifier G, the decimal comma and the semicolon as separator is used.

The five identifiers are serving for characterising the experiment, e.g.,

```
CSV/E      LP1,wheat,17,2009,serie
```

All produced csv files start the records with this five columns.

The five identifiers may be chosen freely but if the first identifier is a number, e.g.,

```
CSV/G      123,,,,,
```

the names of the csv files are built by appending to this number. In the above case the files 123_mean.csv, 123_corr.csv, 123_anov.csv, 123_perc.csv are produced. On the other hand the data file name is extended correspondingly.

3.2.8 The RUN Statement

After the statement RUN, PLABSTAT executes the statements given earlier. It may be possible that additional errors in the statements are detected now, which were not identified earlier during reading in the statement lines. For instance, it will be checked if the number of variable names and the number of variables correspond. RUN is an essential statement.

3.2.9 The EOD Statement

After the RUN statement, PLABSTAT expects a set of data, whose structure has been defined in the foregoing BASIC, LATTICE, ANOVA, or UNREP statement. The data set has to be closed by a separate EOD line.

3.2.10 The STOP Statement

With the STOP line a PLABSTAT job is terminated. If necessary more PLABSTAT jobs with data and statement lines can be added before the STOP statement.

3.3 The Input of Data

Most data are of numerical type and can be fed in with free format, separated by at least one blank character. A separate data row must be used for each plot or experimental unit.

Sometimes, a formatted data input is desirable, e.g. if data are imported from a data base. In such cases, a FORTRAN like format can be appended to the statements BASIC, LATTICE, or ANOVA starting with \$F. For further details see Chapters 4.1.1, 5.1.1, and 6.1.1.

If desired the data for LATTICE and ANOVA statements can be sorted using any sorting routine (5.1.2 and 6.1.11).

Missing data are fed in as stars * or as zeros if a free format is used. With a formatted input, blanks or zeros can be used for missing values. PLABSTAT replaces them by estimates if LATTICE or ANOVA is called. Some more details are given in 5.1.2 and 6.1.11.

4 Calculation of Basic and Regression Statistics

The BASIC statement enables the estimation of some basic statistics, i.e. mean, variance, standard deviation, coefficient of variation, skewness, curtosis for the individual variables as well as the matrix of correlation coefficients for a set of variables. With the FIT statement a simple or multiple regression analysis can be conducted.

4.1 Statements

4.1.1 The BASIC Statement

With the BASIC statement data can be read in and basic statistical parameters can be estimated:

```
BASIC/Q NROW NIDT NINPT NVAR
```

whereby NROW = number of units or rows in the data set

NIDT = number of alphanumeric fields identifying the data rows

NINPT = number of variables to be read

NVAR = number of variables to be analysed.

The qualifier Q is used for controlling input and output. Often, the qualifier including the slash can be omitted. In this case the default value 1110 is assumed, see 3.1.6. The qualifier Q consists of 4 digits MISS, EXTR, PRIN and NEWF,

whereby MISS = 0 do not interpret zeros as missing values

1 interpret zeros as missing values (default)

EXTR = 1 unused (default)

PRIN = 1 normal output (default)

2 output of Spearman's coefficients of correlation

NEWF = 0 plot values not saved on secondary file (default)

1 or 2 (transformed) plot values saved on secondary file.

If a formatted reading of data is desired, e.g. if no blanks have been used for separating the data, if more than one data line needs to be read in per plot, or if the sequence of characters shall be changed append \$F followed by a FORTRAN like format to the BASIC statement as follows:

```
BASIC 16 0 6 6 $F X5, F2.1, 3F3.0, 1X, 2F.2
```

The actual observations must be read in with the F- or E-descriptor. Moreover other descriptors, e.g. X for skipping individual characters or / for skipping the rest of line can be used. For further details see a FORTRAN description.

With CALCULATE one can modify variables (traits) or create new ones (3.2.3). Using the VARIABLE_NAMES statement variables can be named (3.2.4). If you wish scatter plots use the PLOT statement (8.3).

4.1.2 The FIT Statement

With the FIT statement simple or multiple linear regression analyses are executed after having used the BASIC statement before for reading in of the data:

```
FIT/Q Y X1 X2 . . . .
```

whereby Y = variable identifier for the dependent variable Y

X1 = variable identifier of the first independent variable

X2 = variable identifier of the second independent variable

etc.

Each variable set must be written on a separate line if several regression analyses should be conducted, e.g.,

```
FIT X5 X1 X2 X3
      Yield Density kernels_per_spike TKW
```

The qualifier Q is used for controlling. Often, the qualifier including the slash can be omitted. In this case the default value 111 is assumed, see 3.1.6. The qualifier Q consists of 3 digits PRIN, RSDL, and ICPT:

whereby PRIN	= 0	only regression analyses are printed
	1	covariance matrix of regression coefficients is printed additionally (default)
	2	covariance matrix of the variables is printed additionally
RSDL	= 0	no residuals are printed
	1	test on outliers (default)
	2	fitted values and residuals are printed on the secondary file
ICPT	= 0	regression without intercept
	1	regression with intercept (default)

4.1.3 The Data Input

For every plot or experimental unit a separate data line should be fed in, which is structured as follows:

```
ID1 ID2 . . . X1 X2 . . .
```

ID1, ID2, ... are alphanumeric fields, coding the experiment, year, location, type of unit, etc. In total NIDT such fields - separated by blanks - are read in, whereby NIDT may be zero. These codes are not essential for BASIC and hence any character may be chosen.

X1, X2, ... are the NINPT observations per plot to be read in. The fields or numbers are separated by one or more blank characters. Examples for the data input are located in chapter 2 and on the file [psdbasc.dat](#).

Missing data must be represented by * or zeros in free format, and by blanks or zeros in formatted reading. If the qualifier of BASIC is MISS=0, then zeros are not treated as missing values. Note: Data pairs including a missing value are ignored in estimating covariances in the FIT statement.

4.2 Output and Methods

For each variable to be analyzed, one line is printed out including the number of observations (Units) - missing values are subtracted -, the mean value (Mean), the variance (Variance), the standard deviation (Std.dev.), the coefficients of variance (Coef.var.), the skewness (Skewness), and the kurtosis (Kurtosis).

The coefficient of skewness, g_1 , which is zero for a normal distribution is printed out. Positive values of skewness indicate a right-skewed, negative ones a left-skewed distribution. If it deviates significantly from zero, the corresponding significance at $P=10\%$, 5% , and 1% is notified by means of +, *, and **, respectively. The same applies for the kurtosis value, g_2 , which is also zero for a normal distribution. Negative values of kurtosis are caused by a top-flatted distribution, positive ones indicate a peaked distribution. For further details see SNEDECOR and COCHRAN (1980, p. 79).

Calculating coefficients of correlation between two variables those pairs of observations comprising a missing value are excluded.

In the multiple regression table the intercept and the coefficients of regression are given under the heading `Regr.coeff.` followed by two columns comprising their respective standard errors (`Std.error`) and t values (t -value). Under `Std.regr.coeff.` the path coefficients or standardized regression coefficients are listed as a measure for the relative importance of the regression variables, see SNEDECOR and COCHRAN (1980, p. 357). Under `part_r` the partial coefficients of correlation of Y on X_i holding X_j (all $j \neq i$) constant, under `part_SS` the partial deviation sums of squares, and under `SSXY**2/SSX` the reduction in sums of squares due to simple regression of Y on X_i can be found.

For each regression model the coefficient of determination in % ($R^2\%$), the multiple coefficient of correlation (R), and the corrected Akaike Information Criterion (AIC_c) are given for comparing several models, see BURNHAM and ANDERSON (1998).

Outliers are tested during the multiple regression calculations, see DRAPER und SMITH (1998). Thereby suspicious residuals and data points influencing the regression strongly are listed.

In multiple regression it should be considered that the numerical calculations may become instable if very high correlations among variables, e.g. $r = 0.998$, are present in the matrix of correlation coefficients. This risk can be avoided by subtracting a

value, A , which is close or equal to the variable mean, X_i , from all observations, X_i , of this variable by means of the CALCULATE statement (see SNEDECOR and COCHRAN 1980, p. 400).

5 Analysis of Randomized Complete Block and Lattice Designs

If a formatted reading of data is desired, e.g. if no blanks have been used for separating the data, if more than one data line needs to be read in per plot, or if the sequence of characters shall be changed append \$F followed by a FORTRAN like format to the LATTICE statement as follows:

```
LATTICE  7 1 1 2  6 3  $F I6,2I2, 3F3.0,1X,3F5.0
```

The first field with the identifier must be read in with A6 descriptor, the next two fields (the block and treatment number, see 5.1.2) with an I descriptor, and the data fields with F- or E-descriptors. Moreover other descriptors, e.g. X for skipping individual characters or / for skipping the rest of line can be used. For further details see a FORTRAN description.

The effects of the incomplete blocks (based on the adjusted values and with the corresponding replication effects added) can be saved on the secondary output file, if \$An is appended to the LATTICE-statement, e.g.

```
LATT  7 1 1 2  6 3  $An
```

Augmented designs can be analysed this way. First, the check treatments are read in and subjected to an analysis of variance (lattice analysis). Second the number of augmented treatments, i.e. the number of appended plots n, are read in and adjusted for the replication and incomplete block effects. The appended observation data must be given in the same format as the check or replicated treatments with block and treatment numbers (T+1, T+2, .. T+n).

It is possible to use LATTICE with one replication (R=1) only.

With CALCULATE you can modify variables or create new ones (3.2.3). The VARIABLE_NAMES (3.2.4.) and NAMES_OF_TREATMENTS (3.2.5.) statements can be used for feeding in variable and treatment names, respectively. These names will then appear in the heading of the treatment by trait tables.

More information can be obtained using two additional statements: With CHECKS (see 3.2.6.) the treatment values are given relative to the check-mean. With INDEX (see 8.2) selection index criteria of treatments are calculated and printed as last column in the mean table or sorted treatment tables are generated.

If you wish scatter plots of the treatment means or of individual replications use the PLOT statement (8.3) and with the TBT_TAB statement (8.1) you obtain formatted tables of treatment means.

5.1.2 The Data Input

For each plot or experimental unit a separate data line is fed in, which is structured as follows:

```
IDENTIFIER  BN  TN  X1  X2  ...
```

IDENTIFIER is an integer number, specifying the experiment, year, location, etc.

For writing the adjusted mean values on secondary file, it should not exceed 6 characters.

BN in lattice designs specifying the number of the incomplete block, in which the plot is located with the numbers 1, 2, ... R*S; in RCB designs specifying the replication number, with the numbers 1, 2, ... R

TN is the treatment number, with the numbers 1, 2, ... T

x1, x2, ... are the NINPT observations per plot to be read. These numbers are separated by one or more blank characters. If not so, a correct data input might possibly be still achieved using the above mentioned \$F-format.

The data may be sorted or unsorted. Examples are given on the file [psdlatt.dat](#).

Missing data must be identified by * or zeros in free format and by blanks or zeros in formatted reading. If MISS=0 in the qualifier of the LATTICE statement, zeros are not treated as missing values. It is also possible that entire treatments can be considered as missing. In this case PLABSTAT sets the treatment effect to zero and estimates the missing data as usually.

If the LATTICE qualifier PRIN is not set to 3, missing treatments are printed out as zeros in the mean tables and the secondary output file. PLABSTAT handles also cases in which the values of individual variables from one or more replications are missing.

5.2 Output

Unless PRIN is set to 3 in the qualifier, estimates for missing values are only printed out if their number does not exceed 15.

The test on outliers is executed in a modified way following the method of ANSCOMBE and TUKEY (1963), see also SNEDECOR and COCHRAN (1967, Chap. 11.11). It detects extreme residuals (plot errors). To identify multiple outliers as good as possible at first a robust estimate for the error variance, i.e. the MAD (median of absolute residuals) is calculated. Using this parameter the standardized (studentized) residuals are calculated, i.e. the residuals divided by the standard deviation. An observation will be identified as outlier, if a certain limit, which depends on the error variance and its degrees of freedom is exceeded. This test on outliers is based on the assumption of normal distributed residuals. If the distribution is different, however, sometimes too many outliers may be identified.

In any case the experimentator should decide whether he/she will judge a certain extreme value as a possible observation or as an outlier handling it as a missing value. To support your proving you may find the replication means (Replic.means) for each variable on the first line followed by a list of suspected outliers comprising their studentized residuals (St.Resid.) and respective observations in each replication. It is recommended to check the lines with the highest studentized residual first. If you had to correct gross errors, e.g. misspelling, in your data it may happen that in the next run further outliers are identified by PLABSTAT.

Moreover you find a table of the incomplete block effects in case of lattice analysis. These effects estimated in an adjusted mode reflect the block variability, e.g. due to soil differences, in the experiment. It may be possible that you detect a gradient or some other strong influence in your data by that.

For each varying trait an ANOVA table is printed out. In the case of lattice analysis, the components of variance are calculated by means of an "intra-block analysis". Concerning this see FEDERER (1955) or UTZ (1969, Bad Nauheim). The F test of treatments is conducted on this basis, too. The significance test have the following meaning:

- + significant at the 10% level
- * significant at the 5% level
- ** significant at the 1% level

Some other parameters in the analysis of variance table are:

<code>SE(tr)</code>	= standard error of a single (adjusted) treatment mean
<code>SD(tr)</code>	= standard deviation among (adj.) treatment means
<code>C.V.%</code>	= coefficient of variation, i.e. the square root of = the effective error variance divided by the experimental mean and multiplied by 100
<code>Repeat1%</code>	= repeatability of one plot value in % = treatment variance comp. / (treatment var. comp. + eff.err.var.)
<code>Effic.</code>	= efficiency of the lattice analysis relative to a RCB analysis in %
<code>my</code>	= weighting factor of the lattice, see COCHRAN and COX (1957)
<code>Eff.Err.MS</code>	= effective error mean square in lattice or RCB analysis

With the parameter `Repeat1%` or repeatability of one plot you can compare experiments regarding their accuracy independent from the respective number of replications. This parameter is also quite interesting, if error and treatment variance are positively correlated. However, if the error variance is correlated with the experimental mean, the `C.V.` may be more appropriate for judging the accuracy of an experiment.

With lattice analysis treatment means are adjusted for those traits, for which the efficiency exceeds 100 %. This approach is supposed to supply the best information and is also recommended in most text books.

At the top of the table of the treatment means you will find two to three additional lines:

<code>C.V.</code>	= obiger Variationskoeffizient
<code>REP1</code>	= obige Wiederholbarkeit <code>Repeat1%</code>
<code>LSD5</code>	= least significant difference at the 5% level
<code>CHEC</code>	= mean of the check treatments (if CHECKS is given)
<code>MEAN</code>	= experimental mean alternatively
<code>MwCh</code>	= mean without checks (= mean of the non-check treatments)

If you use the qualifier `NEWF = 1`, some of these parameters are also written in the secondary output file. If the `CHECK` statement is used, the adjusted treatment means are also given as percents to the mean of the checks in a separate table.

Moreover, a table of coefficients of correlation between all traits of the treatment by trait table is printed out including an approximative significance test. The calculation of the coefficients is based on the adjusted treatment means.

5.3 Methods

The analysis of generalized lattices is done after WILLIAMS (1977) in an iterative way. Integrated are the special cases of a square lattice or a randomized complete block design.

Missing data are calculated iteratively minimizing the residual mean square (YATES, 1933; HEALY and WESTMACOTT, 1956). The iteration is terminated if the fourth decimal of the residual mean square is not changed anymore after repeated iteration. Thus the decimals given in the table of the analysis of variance may be valid also in the case of missing data. The number of iterations is reduced using a trick of PREECE (1971).

6 Analysis of Multi-factorial Experiments

The statement ANOVA analyses balanced multi-factorial experiments and subdivides two-factorial interactions according to a regression model, e.g. to conduct an analysis on phenotypic stability.

6.1 Statements

6.1.1 The ANOVA Statement

With the ANOVA statement you can read data and conduct an analysis of variance for balanced data:

```
ANOVA/Q NIDT NINPT NVAR
```

whereby NIDT = number of alphanumeric fields identifying the data rows (0,1,2,...)
 NINPT = number of variables to be read
 NVAR = number of variables to be analysed.

The qualifier Q is used for controlling input and output. Often, the qualifier including the slash can be omitted. In this case the default value 1110 is assumed, see 3.1.6. The qualifier consists of 4 digits MISS, EXTR, PRIN, and NEWF,

whereby MISS = 0 do not interpret zeros as missing values
 1 interpret zeros as missing values (default)
 EXTR = 0 no test on extreme values or outliers
 1 test of residuals on extreme values (default)
 2 test of residuals and effects on extreme values
 PRIN = 1 normal printing (default)
 2 instead of means printing of effects ??
 3 extended printing of missing data,
 i.e. output of estimates for missing values,
 including these estimates in the mean tables,
 and output of convergence information
 NEWF = 0 mean values not saved on secondary file (default)
 1 mean values saved on secon. file (use TBT_TAB to define the means)
 2 (transformed) plot values saved on secondary file.

If a formatted reading of data is desired, e.g. if no blanks have been used for separating the data, if more than one line needs to be read in per plot, or if the sequence

of characters shall be altered append \$F followed by a FORTRAN like format to the ANOVA statement as follows:

```
ANOVA 0 6 6 $F X5,F2.1,3F3.0,1X,2F5.2
```

The actual observations must be read in with the F- or E-descriptor. Apart from these other descriptors, e.g. X for skipping individual characters or / for skipping the rest of the line can be used, too. Further details see a FORTRAN description.

The statements FACTORS, MODEL, and, if necessary, RANDOM must precede the ANOVA statement, to define the structure of the experiment. In doing so the FACTOR statement has to be the very first one.

With the CALCULATE statement (3.2.3) you can modify variables or create new ones (traits). The VARIABLE_NAMES (3.2.4) and NAMES_OF_TREATMENTS (3.2.5) statements can be used for feeding in variable and treatment names, respectively. The respective names will then appear in the heading of the treatment by trait tables.

More information can still be obtained from these tables using the two following statements: With CHECKS (see 3.2.6) the treatment values are given relative to the check mean. With INDEX (see 8.2) selection index criteria of treatments are calculated and printed as last column in the mean table or sorted treatment tables are generated.

If you wish scatter plots of the treatment means or of the individual replications use the PLOT statement (8.3) and with the TBT_TAB statement (8.1) you may obtain formatted tables of treatment means.

For conducting combined analyses across series of experiments the ERROR statement may be used to take over the estimates for error variances, their degrees of freedom and other relevant parameters from the single experiments (8.5).

6.1.2 The FACTORS Statement

This statement lays down the factors and the number of levels for each factor:

```
FACTORS TEXT1=N1 TEXT2=N2 TEXT3=N3 ....
```

The first 10 characters of each TEXT are used as factor name. After the sign of equality the number of factor levels is given. Note: signs of equality, blanks, or exclamation marks are not allowed within factor names (TEXT).

The first character of each factor name is used as symbol in tables and in other statements like MODEL or MEANS. Therefore each factor name must start with a different character.

For example, the following statements are valid:

```
FACTORS T_variety = 35 B_Replications = 3
FACT MALES = 178
      FEMALES_IN_MALES = 9
      PLACES = 3
```

The FACTORS statement must precede the MODEL or RANDOM statements specifying the analysis of variance. For feeding in sorted data, the sequence of factor names in the FACTORS statement must correspond to the order of data (see 6.1.11). Or, if \$ specification in ANOVA is used the sequence there must be in the same sequence as the factor names.

It is possible to perform analysis of variance with one factor only.

6.1.3 The MODEL Statement

This statement defines the biometrical model for the analysis of variance. It consists of effects which are described by the factor symbols and linked by plus signs.

```
MODEL EFFECT1 + EFFECT2 + EFFECT3 + ...
```

Main effects are given with the single factor symbol, and interaction effects with a combination of several symbols. For example the model for a three-factorial classification has to be set up as follows:

```
MODEL A + B + C + AB + AC + BC + ABC
```

Nested effects are described by a colon, e.g. if C is nested within A and B, than the biometrical model reads as:

```
MODEL A + B + AB + C:AB
```

The last effect is always interpreted as error or residuum. If further effects shall be considered as errors, e.g. the error of main plots in split plot designs, append a slash to the particular effect as in the example below:

```
MODEL R + M + RM/ + S + SM + SMR
```

where R symbolizes the replication effect, M the effect of the main plot treatment, RM the main plot error, S the effect of the subplot treatment, SM the SxM interaction and SMR the subplot error.

The lines in the analysis of variance table are arranged according to the sequence of the factor in the MODEL statement.

The MODEL statement must be elaborated carefully, as the correct breakdown of total variability and the correctness of the significance tests depends on it. One should clearly distinguish between nested and cross-classified (factorial) factors and set up the model accordingly. Furthermore you must prove, how randomization was done in reality before you can be sure of having set up a correct model. Therefore no short cuts can be provided. Some guide-lines you may find in the books of SNEDECOR and COCHRAN (1980) or SEARLE (1971), in the appendix C, and in the manuscript of the lecture "Datenverarbeitung in der Angewandten Genetik".

6.1.4 The RANDOM Statement

With this statement the random factors are defined. All factors not listed are assumed to be fixed. Therefore all factors are considered as fixed if the statement is omitted.

```
RANDOM list of factor symbols
```

In the above 3-factorial example the RANDOM statement must be written as follows, if A is considered to be fixed while B and C are random factors.

```
RANDOM B C
```

Mostly you may define blocks, locations, and years as random to obtain the significance tests common in agronomy or breeding, i.e. the test of treatment variability against interaction with environments. The aspects of whether a factor is to be considered as fixed or random is discussed in the text book of SEARLE (1971, pp 376).

6.1.5 The MEANS, EFFECTS, PERCENTS, or RANKS Statement

Two-way mean tables, included the marginal means, are generated by

```
MEANS list of two-factor EFFECTs
```

e.g.,

```
MEANS AB AC
```

The construction of the mean table is determined by the order of the factor symbols. The first factor symbol is the row factor and the second factor symbol the column factor of the table.

Similar, a two-way table of effects is printed by

```
EFFECTS list of two-factor EFFECTs
```

or a two-way table of percents (relativ to the mean of the checks) by

```
PERCENTS list of two-factor EFFECTs
```

or a two-way table of ranks by

```
RANKS list of two-factor EFFECTs
```

Such tables are only produced if the corresponding two-way effects are included in the MODEL line.

Note:

RANKS must be abbreviated by RANK. Otherwise it can not be differed from RANDOM.

PERCENTS produce the relative values which are defined by CHECKS. Hereby CHECKS must be the means of checks of the row factor. If a table PERCENTS TP, e.g., should be generated then CHECK/T must be defined, but not CHECK/P. The import of check means by the CHEC rows is only possible for TBT.TAB tables.

RESTRICT (6.1.10) restrict the output of such two-way tables to the variables defined with RESTRICT. So the output can be concentrated to the most important traits, e.g. for a Line×Tester×Places situation:

```
RESTRICT yield
EFFECTS LT      (i.e. the gca-sca table)
RANKS LP        (i.e. the ranks of lines at each place)
```

6.1.6 The SUBINT Statement

The SUBdivision of the INTeractions (=SUBINT) in a two-way table can be achieved through the SUBINT statement.

```
SUBINT list of two-factor EFFECTs
```

e.g.

```
SUBINT AB
```

SUBINT produces also a correlation matrix of the AB means and calculates an analysis of phenotypic stability which is interesting for series of experiments especially, see chapter 6.2.

As subdivision makes only sense with at least three levels per factor, only such data sets can be analyzed with the SUBINT statement.

Data with missing values are not optimally treated.

6.1.7 The TBT_TABLE Statement

A Treatment By Trait Table (TBT_TAB), i.e. a table of treatment means across traits, is caused by this statement.

```
TBT_TABLE list of EFFECTs
```

Examples may be:

```
TBT_TABLE AB
```

```
TBT_TAB ABC AB BC B
```

Moreover, the means specified in this statement are written to a secondary output file too, if the corresponding ANOVA qualifier is used. Using the PLOT statement they are printed in scatter plots. How to obtain formatted tables and at the same time absolute as well as relative values from the TBT_TAB statement is described in chapter 8.1.

6.1.8 The HERITABILITY Statement

The heritability for certain treatment means can be calculated by

```
HERIT factor symbol
```

e.g.

```
HERIT T
```

```
HERIT T:C
```

Simultaneously the 95% confidence interval is printed in brackets.

6.1.9 The GENOTYPIC_CORRELATION_MATRIX Statement

The matrix of phenotypic and genotypic correlation coefficients and the extended variance-covariance analyses if necessary can be calculated:

```
GENOT EFFECT
```

The two correlation matrices are estimated for the treatment effect which is represented by an effect in the MODEL line, e.g.

```
GENOT T
```

```
GENOT T:BC
```

The variance-covariance analyses for each pair of variables are printed if the qualifier is set to 1:

```
GENOT/1 EFFECT
```

6.1.10 The RESTRICT Statement

The output of two-way tables which are generated by MEANS, EFFECTS, PERCENTS, RANKS or SUBINT can be restricted on certain variables:

RESTRICT list of variable identifiers

e.g.

```
RESTRICT X1 X5
RESTRICT Yield TGW
```

6.1.11 The Data Input

For each plot or experimental unit a separate data line should be fed in, which is structured as follows:

```
ID1 ID2 ... X1 X2 ...
```

ID1, ID2, ... are alphanumeric fields, coding the experiment, year, location, type of unit etc. for identification. The fields are separated by one or more blanks. In total NIDT such fields are read, whereby NIDT may be zero. If the data input is sorted, these codes are not essential for ANOVA, and hence any character may be chosen.

X1, X2, ... are the NINPT observations per plot to be read in. The fields or numbers are separated by one or more blank characters. If the observations are not separated by blanks, it might be still possible to obtain a correct data input using the above mentioned \$F-format.

If data set is read in sorted already the sequence of the factors in the FACTORS statement must correspond with the respective sorting order. If the data are sorted in the following manner,

A	B	X11		B	A	X11
1	1	12.3		1	1	12.3
1	2	11.7		2	1	11.7
2	1	13.5	or	1	2	13.5
2	2	12.8		2	2	12.8
3	1	10.4		1	3	10.4
3	2	12.9		2	3	12.9

the sequence of the factors in the FACTORS statement reads in both cases as follows:

```
FACTORS      A_Factor=3      B_Factor=2
```

since those is the lexical order of the above data based on sorting the A names and then the B names. Examples for data input can be found on the file [psdanov.dat](#).

In the case of unsorted data or missing data lines, the ANOVA statement must be supplemented by a \$ character and subsequent integers indicating where to find the factor levels on the data lines, e.g.:

```
ANOVA NIDT NINPT NVAR $ K1(KS1) K2(KS2) K3(KS3) ...
```

The integers K1, K2, K3 ... state the beginning of those columns in the data line, in which the names for the levels of individual factors are given. The integers KS1, KS2, KS3 ... indicate the number of characters of the respective names.

The order of these keys has to correspond with the sequence of factor names in the FACTORS statement. In this case the NAMES statement is unnecessary. If the last statement is used the factor names in tables are given in this order. The following example may elucidate this approach:


```

FACTOR  variety = 15      years = 2      places = 6
ANOVA   3 8 8    $ 12(4)  3(2)    8(3)

```

The corresponding data lines are arranged in the following way:

```

9987 XXP30 LEO 70.3 552 1.3 29.3 43.6 96.0 514 0
--   --- ---

```

In this line the year is represented by "87", the place by "P30" and the variety by "LEO". The number of identifiers is still optional and is laid down as usual by NIDT in the ANOVA statement.

Missing data must be represented by * or zeros in free format, and by blanks or zeros in formatted reading. In case of ANOVA with the \$ extension, lines with totally missing data can be omitted. If the qualifier of ANOVA is MISS = 0, then zeros are not treated as missing values. Note, if certain factor levels or combinations are completely missing, PLABSTAT assumes the corresponding effects to be zero and estimates the missing values following the least square method (method of fitting constants).

6.2 Output

The output of ANOVA starts with a table of the expectations of mean squares which is the basis for calculating the components of variance and F values. The table is valid for a balanced data set only.

Besides the degrees of freedom DF, sums of squares SS, and mean squares MS, components of variance Var.cp and their respective standard errors s(V.cp.) are given in the table of analysis of variance. The variance components are also given for fixed effects, though these are rarely of interest. However, they can be interpreted, see SNEDECOR and COCHRAN (1980, chap. 13.9).

Moreover, the F values with significance tests, see also 5.2, the corresponding degrees of freedom of the nominator DF-NM, and of denominator DF-DN are printed out. The degrees of freedom need not to be integers since SATTERTHWAIT's approximation is used in more complicated cases.

The heritability, i.e. the ratio of genetic to phenotypic variance of certain means, and the 95% confidence interval are determined after calling HERITAB. For calculation see KNAPP and BRIDGES (1987).

For every source of variation or rather for the respective means, the standard error s.e. and the least significant difference LSD5 at a probability level of 5% are given in the last two columns.

The outlier test is performed according to ANSCOMBE and TUKEY (1963), see SNEDECOR and COCHRAN (1967, Chap. 11.11.) and discovers extreme residuals (plot errors). This test requires normal distributed residuals. For a suspected extreme value the residual is given in percent of the test statistic. Additionally, you can also test effects. In this case the suspected effect is given out together with its respective level numbers and its estimate.

Unless PRIN is set to 3 in the ANOVA qualifier, estimates of missing values are only printed if there are less than 15 missing values. In the mean tables and the secondary output files, missing values are printed out as zeros, if PRIN is not set to 3.

Using TBT_TBLE in combination with a two-factor table AB the correlation coefficients of the AB means on the marginal A means are computed for each level of B. With less than 20 levels for the factor A this table is missing. These correlation coefficients may particularly be of interest in analyzing series of experiments. If the correlation coefficients are different the interaction effects might be correlated differently with the main effects. Single coefficients of about zero might indicate, that the observations on the respective level are contradictory to those on the other levels. These estimates can help in identifying associations while interpreting a two-way table. A more detailed analysis results from a SUBINT call.

The subdivision of interactions in a two-way table by SUBINT results at first in a combined analysis of variance. For interpretation see BLISS (1967), WRIGHT (1971) or UTZ (1972). The two subanalyses of variance or SubANOVA's are calculated according to YATES and COCHRAN (1938) or PERKINS and JINKS (1968).

For the alternative levels of both factors, several estimates are given. If the values in a two-way table may be designated as x_{ik} , in which i symbolizes the index for the row and k the one for column the estimates for the i th level read as follows:

Mean	= mean $x_{i.}$, in which the dot indicates that averaging was performed across index k
Corr.	= correlation coefficient between x_{ik} and $x_{.k}$
Regr.	= corresponding regression coefficient of x_{ik} on $x_{.k}$
MSdev	= deviation mean squares of the x_{ik} from the above regression
MSentry	= simple variance for the i th level, i.e. $var(x_{ik})_i$
MSinteract.	= variance of the interaction effects $var(x_{ik} - x_{i.} - x_{.k} + x_{..})$, corresponding to the ecovalence of WRICKE, defined as SS
MSdevXHY	= variance of deviations from "site highest yield", as described by JENSEN (1976).

Accordingly the estimates for the k th level of the other factor can be defined.

In the analyses of covariance following estimates are printed additionally:

In the column v.cp.% of the ANOVA tables, you find the ratio of the variance component of the respective variation source to the variance of the appropriate means. In genetical situations this is identical with the ratio of genetic to phenotypic variance or heritability. Note, adding up this column does not result in 100%.

The phenotypic coefficients of correlation r -phen are estimated and tested on significance. The genotypic coefficients of correlation are found under r -gen, their standard error under $s(r$ -gen). It is appended + if the absolute value of genotypic coefficient of correlation is greater than the single standard error, or ++ if he is greater than the doubled standard error.

6.3 Methods

The analysis of variance is conducted by means of the algorithms AS1, ASR1, AS18, and AS19, published in Appl. Statist., vol. 17 and 19, see also UTZ (1978). These algorithms were revised to save computing time and converted in FORTRAN. Calculations are executed with double precision and the sums of squares are computed from the effects.

SATTERTHWAITE's approximation for the F value is based on the type

$$F = (MS1 + MS4) / (MS2 + MS3),$$

and not the one

$$F = MS1 / (MS2 + MS3 - MS4),$$

see COCHRAN (1951) or SNEDECOR and COCHRAN (1980, ch. 16.14).

Missing data are calculated iteratively following the least square method as described already in 5.3. Sometimes a factor level or a whole factor combination is missing entirely then PLABSTAT displays:

```
MISSING LEVEL(S) IN ...
```

If the iteration process does not converge - hopefully seldom - a notice is given and simple means are calculated.

With missing values the treatment mean squares are biased upwards and F values result in significant results too often. For a few missing data this can be neglected (COCHRAN and COX, 1957, ch. 3.7). A rough limit for an intolerable bias may be 10% missing observations according to BENNETT and FRANKLIN (1954). If there are more than 13% missing values the ANOVA table is not printed any more. In case of missing values the MS can be corrected as described by SNEDECOR and COCHRAN (1980, ch. 15.3), GOULDEN (1952, ch. 14), or SEARLE (1971). BERK (1987) has given a comparative discussion of possible methods. If necessary a statistical package for analyzing unbalanced data sets, e.g. SAS or GENSTAT has to be used.

The standard error of genotypic coefficients of correlation requested by GENOT (6.1.9) are calculated after MODE und ROBINSON (1959). The analogous analysis of variance method or method 1 after HENDERSON (1953) is used for the variance-covariance analyses, see also SEARLE (1971, p. 424ff). Components of variance, genotypic coefficients of correlation, and F tests are estimated assuming mean squares belong to balanced data sets. This procedure is regarded as more appropriate than other approximation at least for the F test, see TIETJEN (1974, Biometrics 30,573). Especially for mixed models the user may use an advanced procedure like REML with package SAS, GENSTAT, ASREML, or R.

Note:

In case of missing data, PLABSTAT may calculate different SS and MS in variance-covariance analyses requested by GENOT/1 compared to SS and MS from ANOVA. With the HENDERSON method 1 are negative SS (or SS analogous more exactly) possible. The two ANOVAs are performed newly for each covariance analysis eliminating the possible missing data in both variables. Therefore the printed coefficients of correlation can not be recalculated from the other printed values in certain cases.

7 Unreplicated Experiments

Unreplicated experiments are analysed with the UNREP statement whereby a certain gradient is eliminated with the aid of check treatments or by using moving averages.

7.1 Statements

7.1.1 The UNREPLICATED Statement

With the UNREP statement you can read and analyse data from unreplicated experiments:

```
UNREP/Q  NPARZ NROW NCOL NSTD NIDT NINPT NVAR
```

with NPARZ = total number of experimental units or number of plots to be analysed,
with $NPARZ \leq NROW \times NCOL$

NROW = number of rows or strips

NCOL = number of columns or plots per row

NSTD = 0 adjusting with moving averages

= n adjusting with check treatments

whereby n the number of check treatments which occur several times in the experiments. (Check treatments occurring once must be treated as regular treatments.)

NIDT = 4 the number of fields identifying the data row (since an alphanumeric identifier of the experiment, the row, column, and treatment number is needed.)

NINPT = number of variables to be read

NVAR = number of variables to be analysed.

The qualifier Q is used for controlling input and output. Often, the qualifier including the slash can be omitted. In this case the default value 1110 is assumed, see 3.1.6. The qualifier consists of four digits MISS, EXTR, PRIN, and NEWF,

whereby MISS = 0 do not interpret zeros as missing values

1 interpret zeros as missing values (default)

EXTR = 1 not used (default)

PRIN = 1 normal printing (default)

3 extended printing with unadj. and adj. observations,
listing side-by-side for each variable in *.prt

NEWF = 0 observations not saved on secondary file (default)

1 (or 2) adj. observations are saved on secondary file.

If a formatted reading of data is desired, e.g. if no blanks have been used for separating the data, if more than one data line needs to be read in per plot, or if the sequence of characters shall be changed append \$F followed by a FORTRAN like format to the UNREP statement as follows:

```
UNREP 27 3 9 3 4 3 4 $F 6X,2I2,I5,3F5.0
```

The actual observations must be read in with the F- or E-descriptor. Moreover other descriptors, e.g. X for skipping individual characters or / for skipping the rest of line can be used. For further details see a FORTRAN description.

With CALCULATE (3.2.3) you can modify variables (traits) or create new ones. The VARIABLE_NAMES (3.2.4) and NAMES_OF_TREATMENTS (3.2.5) statement can be used

for feeding in variable and treatment names, respectively. If you wish scatter plots use the PLOT statement (8.3). With TBT_TAB (8.1) formatted or sorted tables can be obtained.

In case of moving averages, the CHECKS (3.2.6) statement can be used with the treatment numbers of checks, e.g.,

```
CHECKS  1 2  11 12
```

7.1.2 The Data Input

For each plot or experimental unit a separate data line must be written, for example:

```
T299XX 01 01 1 5.0 1.0 ...
```

whereby the data are separated by one or more blanks, first the identifier of the experiment, then the row and column number (1,2 ... NROW and 1, 2 ... NCOL, respectively), the treatment number, and the observation values. The identifier should not exceed six characters. Check varieties used in the check plot method must be characterized by a treatment number greater 5000.

The order of data lines is free. Plots or data lines within the rectangle described by NROW×NCOL can miss, e.g., in an edge.

Single missing values must be characterized by star * as usually.

Pedigrees are introduced by the NAMES_OF_TREATMENTS statement whereby the names are in the same order as the treatment numbers (if NSTD=0) or in the order of data lines (if NSTD>0).

In the file [psdunrep.dat](#) some example data are found for both analyses, the check plot method and the moving average method.

7.2 Output

The output starts with a summary table of the unadjusted observations as it can be found with the BASIC statement. It follows the coefficient of correlation $r(\text{obs}, \text{concom})$ and the coefficient of regression $b(\text{obs}, \text{concom})$ to judge the efficiency of adjusting, see 7.3.3. The table of adjusted observation values is given in form of a TBT table. In case of moving averages mean values are calculated if few or all treatments occur manifold. Thus it can be proven if a soil gradient must be assumed in a block experiment and should be eliminated.

If PRIN=3 is chosen further parameters are given separately for each variable, in case of check plot method:

Mean of checks, Check means, No. of check plots and

NBETW = maximum distance between two check plots

NNMIN = minimum number of check plots for the adjustment of a test plot
(e.g. in the corner or with missing check plots)

NNMAX = maximum number of check plots for the adjustment of a test plot

Under CHECK PLOT VALUES the residuals of each check plot and the means of residuals of the neighboured check plots are listed. The association of the plot values and the concomitant residual means of the neighboured check plots are described under $r = r(\text{obs}, \text{concom})$ and $b = b(\text{obs}, \text{concom})$.

Under PLOT VALUES, the position in the experiment, i.e. the row and column number, the type (if test plot $\text{typ} = 0$, if check plot $\text{typ} = i$, i.e. the number of the check), the unadjusted and adjusted observation value, and the difference of the last two values are listed for each plot.

In case of moving average method, under PLOT VALUES the position, the treatment number, and the unadjusted and adjusted observation value, and the moving average (mov. ave.) used for adjusting are listed for each plot.

7.3 Methods

Gradients or spatially correlated errors are eliminated with the aid of the next check plots or by moving average of neighboured plots. The check plot method is less recommendable see 7.3.3 but is used often in experiments of practical plant breeding.

7.3.1 Check plot method

The differences of the check plots from their corresponding check means are calculated first to get estimates for the spacial effect of each check plot. The adjustment of the test plots is done with the average of this differences as covariable (concomitant variable) whereby over those check plot differences is averaged lying in a rectangle around the test plot. The rectangle is build by the next check on the left and right side and the stripe above and under the test plot. The corresponding regression coefficient $b(\text{obs. concom})$ can be estimated by for the check plots and is used for the test plot adjustment. Over fewer check plot differences can be averaged in the corners and borders of an experiment or if check plots are missing. The check plots can be positioned arbitrarily, e.g. in columns, in regular patterns (diamonds), or randomly. If multiple neighbored check plots exist, only the nearest neighbored check plots are used for adjustment.

7.3.2 Moving Average Method

The ten neighbored observations of each plot are averaged for adjusting, i.e., the four left and right lying plots and the plot above and below the test plot. This averages are used as covariable (concomitant variable) as before. Fewer plots are used in the corners for adjustment correspondingly. The trend is estimated for this procedure with the aid of more plots than with the check plot procedure. This is necessary since the observations of the neighbored plots include also genotypic effects.

7.3.3 Remarks to adjusting methods

The regression coefficient $b(\text{obs. concom})$ is mostly less than 1.0 and is used to avoid over-adjustment (see YATES, 1936). With the correlation coefficient $r(\text{obs, concom})$ the worth of the adjustment can be judged. The size should be 0.3 at least after COCHRAN (1957), otherwise a correction is not worthwhile taking in to account the estimation errors. The correlation coefficient should be positive, negative values are only expected if strong competition between neighbored treatments exists. Thus negative values indicate that soil heterogeneity can not be adjusted.

The check plot method can be seldom recommended (BAKER and McKENZIE, 1967). Since it is too less to choose every tenth plot a check to estimate a soil fertility gradient with sufficient accuracy. But if the part of checks is increased then the adjustment will be too expensive.

Moving averages can be recommended rather. The condition is that the treatments are randomly allocated, otherwise eventual family effects are eliminated. The ten plots in the moving average may be a compromise to estimate also small gradients and to get an accurate elimination. Important seems that also plots in the column are used for a correction since often vertical effects are observed, e.g. trucks of a machine, especially in small plots experiments. See also for literature to spatial or neighbor methods.

Further details and literature see CHANDRA (1991), KEMPTON (1984), KEMPTON und FOX (1997, chap. 5-7), or YATES (1936).

8 Supplementary Statements

In the following, some supplementary statements are described, which are especially interesting for LATTICE and ANOVA runs.

8.1 The TBT_TABLE Statement

With the TBT_TAB statement you may specify the mean tables wanted together with an ANOVA statement (already described in 6.1.7). Apart from this the output format of the mean value tables can be determined by this statement. It is very helpful in creating special output formats, e.g. a DIN A4-format. TBT_TAB tables can include 34 variables or 255 characters in width without new-line breaks.

Appending of \$F, followed by a FORTRAN formatting statement the mean values of a TBT_TAB table can be formatted, i.e. they can be equipped with a specified number of places in front or after the decimal point. All FORTRAN rules for formatting are also valid. For changing the title of a table you can feed in a line consisting of the supplement \$T followed by the new title. Up to 6 such \$T lines are allowed, which are placed directly above the observation means in the tables.

If the list of the relative values generated with the CHECKS statement shall get a different format, too, the \$F- and \$T supplements need to be repeated. The mean tables can be sorted according to the mean values of a certain variable appending \$S followed by the variable number. For example a output sorted for variable 2 or DMC is obtained using the supplement \$S X2 or \$S DMC .

Appending \$V and a sequence of variable identifiers associated with the characters a for absolute or p for percentage, e.g.:

```
$V X3p X3a X2a X1a
```

```
$V Yield.a Yield.p DMC.a Score.a
```

The point serves as separator of variable name and symbol a or p. The mean tables can be rearranged with respect to the traits, comprising simultaneously absolute and relative values (regarding the check mean).

In the following you will find an example for a TBT_TAB statement:

```
FACTORS      REPLICAT. = 2      GENOTYPES = 25
```

```

MODEL          R   +   G   +   RG
ANOVA/1211      0   7   6
CHECKS/G        *
! in case of LATTICE, G will be missed in the following line
TBTTABLE       G   $F   2(F5.0),3(F5.1),F4.1,F5.0
                $T
                $T PLHT COHT YIELD COB D.M. SCORE SEL
                $T cm   cm  dt/ha  %   %   1-9  IND
                $F  10(F6.1)
                $T PLHT COHT YIELD COB D.M. SCORE SEL
                $T cm   cm  dt/ha  %   %   1-9  IND
                $$ YIELD
! Sorting of TBT_tables according variable 3 YIELD
CHECKS/G        *
INDEX           -0.2  0   1   -0.2  1   0   50
VAR_NAMES       PLHIG COHIG YIELD  %COB  %D.M. SCORE

```

The mean values (absolute) of the first and second trait are printed with 5 places only in front of the decimal point (format F5.0 repeated twice). The 3rd, 4th and 5th trait is printed with 5 places including one after the decimal point (format F5.1 repeated 3 times). The relative values, however, are all printed with 6 places consisting of the sign, 3 digits in front of the decimal point, the decimal point, and one place after the decimal point (format F6.1).

8.2 The INDEX Statement

With the INDEX statement the weights for a selection index or for another function of the mean values in TBT_TAB tables are defined. With the aid of CALC analogous statements new variables can be generated for which now analysis of variance is performed and appended as last column in a TBT_TAB table. A selection index can be calculated for example as following:

```
INDEX 1 X4=-0.5*X1-1.33*X2+0.77*X3+100
```

or

```
INDEX 1 Index=2*DMC%+Yield+Stand
```

The statement needs as first parameter the number of the index variables to be appended. The parameter NVAR in the LATTICE or ANOVA statement must include the new index variables and also the VARIABLES statement the names of the index variables.

With the aid of the function MC calculating Means of Checks, relative values can be generated, e.g. the "Ertragswertzahl" of the "Bundessortenamt" with two different weights for dry matter content DMC%:

```

INDEX 2 Ch_Yield=MC(Yield,1,2,3)
Ch_DMC%=MC(DMC%,1,2,3)
Ch_Std=MC(Std,1,2,3)
INDX1=Yield/Ch_Yield*100+2.5*(DMC%-Ch_DMC%*100)+1.0*(Std-Ch_Std)
INDX2=Yield/Ch_Yield*100+1.5*(DMC%-Ch_DMC%*100)+1.0*(Std-Ch_Std)

```

Therein the means of the checks were first calculated by averaging the treatment means 1, 2, and 3 for the three variables Yield, DMC%, and Std. The new means are assigned to Ch_Yield, Ch_DMC%, and Ch_Std. Finally two index variables INDX1 and INDX2

are build with different weights. Since two additional index variables should be appended in the TBT table, INDEX must start with number 2.

Based on an index variable a sort of a TBT table can be performed with the \$S option, see statement TBT_TAB (8.1).

8.3 The PLOT Statement

With this statement scatter plots can be generated

```
PLOT/Q X1.X2 X2.X3 ...
```

e.g.

```
PLOT/2 X3.X2 Yield.DMC
```

The first identifier gives the variable number or variable name which represents the y axis or ordinate and the second one following the full stop the x axis or abscissa. You can request several plots by adding additional variable pairs, separated by at least one blank character.

If the first variable equals the second one, e.g. in

```
PLOT X1.X1 TKW.TKW
```

observations of individual replications are plotted. To be more precise, while with two replications the observations of the first one are plotted against those of the second one, the observations of the individual replication are plotted against the means of the observations across all replications if more than two replications exist. Such scatter plots may be helpful to detect outliers or other gradients in the experiment (only valid in combination with the LATTICE statement).

The maximum number of plots per run is 40.

The qualifier Q determines the number of scatter plots per page:

- Q = 1 one plot per page (default)
- = 2 two plots per page (appropriate for DIN A4)
- = 3 three plots per page.

If PLOT is used in combination with an ANOVA call scatter plots are generated for those mean values only, which have been specified in a preceding TBT_TABLE statement.

In the scatter plot (see ch. 2) the dots are consecutively numbered according to the numbering of the entries in the TBT_TAB or the mean value table, respectively. If a dot represents several entries, this is indicated by a star followed by a figure. For example, if a dot is denoted by *3 the corresponding entry numbers are listed at the right-hand margin following *3. Dots representing check treatments are underlined. The regression line of Y on X is marked by / on the y-axis and another / on the right-hand margin and thus the regression line can be drawn by hand.

The mean values are denoted by M on the X-and Y-axis. If the least significant difference (LSD) can be calculated, it will be visualized by a sequence of DD . . D on the axes. The number of dots not depicted in the scatter plot due to missing values is indicated under Miss.

8.4 The PLOT Statement

With this statement a single scatter plot can be generated in PostScript. The output is written on a separate file *.ps.

```
PLOT/n X1.X2
```

e.g.

```
PLOT/1 Yield.DMC
```

A scatter diagram is plotted for the first variable on the Y-axis and the second one on the X-axis. Check treatments, see CHECKS, are written as bold numbers. The regression line of Y on X, the trial means, and the least significant differences (LSD5%) are drawn.

With the qualifier *n* the *n*-th character in the treatment names (or pedigrees) is used to symbolize a scatter point instead of the default circle.

8.5 The ERROR Statement

The ERROR statement is used for analyzing series of experiments. This way it is possible to include *.mnv files without modifications.

```
ERROR factor symbol column number
```

or

```
ERROR number of trials to be pooled column number
```

The *factor symbol* refers to the factor, for which the single experiments differ (for example places). Of course it must appear in the corresponding FACTOR statement, too. Instead of the *factor symbol* the number of single experiments to be pooled can be given directly. The *column number* indicates the beginning column, in which the values for DFE, SE, LSD, CHEC, MEAN are written. Using these parameters the corresponding estimates from the single experiments can be included in the analysis. The default value for the *column number* is 12.

For example:

```
ERROR P 7
ERROR Q
```

Instead of the *factor symbol* the number of experiments to be pooled can be given directly if another number of errors exists in a trial series, for example:

```
ERROR 5
```

Such an import of error variances make sense only if the variables are not changed by a CALCULATE. A wrong ERROR statement often results in a "TOO MUCH LEVEL" error. Note the following recommendations:

1. VAR, DFE, SE, LSD, CHEC or MEAN lines are optional and can be positioned anywhere in the data set. Important are only the DFE and SE lines whereby DFE must be given before SE.
2. VAR . . . MEAN must therefore not appear in the treatment names; being more precise: not at the position indicated by the above number for the column.
3. VAR, . . . MEAN must be placed after the identifiers for the data. Their lines must be specified by the same number of identifiers as a normal data line.

4. If no CHECK statement is given relative values are still calculated on basis of the CHEC lines read in. If there is a CHECK statement, however, the check mean are newly calculated.

Example for the input of a series of experiments, see psdanov-6 of [psdanov.dat](#):

```
!!      SERIES OF RYE EXPERIMENTS
FACTORS  PLACES = 3      TREAT = 25
MODEL    P  +  T  +  TP
ANOVA    2   9   9
CHECKS   24  25
ERROR    P   8
TBT_TAB  TP  T
RUN
117001 DFE  12   16   0 16   16   23   0   0   0
117001 SE   4.79 2.17  0  0.86 .82   .40   0   0   0
117001 LSD 14.77 6.51  0  2.59 2.45  1.16   0   0   0
117001 CHEC 0    0    0  0    0    0    0   0   0
117001 001  79.77 130.36 .00 1.21 4.88 23.00 .00 .00 .00
117001 002  73.90 133.93 .00 1.32 6.00 23.00 .00 .00 .00
117001 003  74.78 132.86 .00 2.00 7.00 22.50 .00 .00 .00
      as the start of the .mnv file of the first experiment,
      .mnv files of further experiments follows subsequently
EOD
STOP
```

8.6 The OMISS Statement

Using this statement numbers can be specified for marking missing data in the secondary output file, which may be desirable if the means values are transfered to a data base program.

```
OMISS      list of real numbers
```

This list has to consist of NVAR real numbers, so that there is one number available for each variable to be analyzed. For example if NVAR = 4:

```
OMISS      0    -9    0    99.99
```

If OMISS is omitted, each missing value is marked by a star * in the secondary output file. Consequently the default is OMISS followed by a list of zeros.

9 Hints, Faults, and Limits

- (1) In biology observations are mostly measured with maximum 3 digits, thus the accuracy of the mean squares or components of variance cannot be higher than than 5 or 6 places.
- (2) It is not possible for PLABSTAT to detect every error with the input and produce a correct error message. For example, you may write

```
ANOVA  10 10 100
```

since you doubled the zero in the last number accidentally. However, the statement is legal, so that PLABSTAT starts to compute and may produce an error message at a different place, in this case e.g. that not sufficient space has been reserved.

Therefore in case of unexpected errors, it is recommended to check at first the correct writing of each statement.

Of course a program can never be perfect and therefore the author would be pleased to be informed about errors appearing while running the program.

(3) There are some limits with the program:

maximum number of factors in ANOVA	= 10
maximum number of variables (including the constants in CALCULATE)	= 350
maximum number of missing data in LATTICE	= 150

There are some some other arrays having a fixed size, too. If necessary, you must look for a version with greater array handling.

10 References

- ANDERSON, R.L. and T.A. BANCROFT, 1952. Statistical Theory in Research. McGraw-Hill, New York.
- ANSCOMBE, F.J. and J.W. TUKEY, 1963. The examination and analysis of residuals. *Technometrics* 5, 141-160.
- BAKER, R.J. and R.I.H. MCKENZIE, 1967. Use of control plots in yield trials. *Crop Sci.* 7, 335-337.
- BENETT, C.A. and N.L. FRANKLIN, 1954. Statistical Analysis in Chemistry and Chemical Industry. Wiley, New York.
- BERK, K., 1987. Computing for incomplete repeated measures. *Biometrics* 43, 385-398.
- BLISS, C.I., 1967. Statistics in Biology. Vol. 1. McGraw-Hill, New York.
- BURNHAM, K.P. and D.R. ANDERSON, 1998. Model Selection and Inference. A Practical Information-Theoretic Approach. Springer, New York.
- CHANDRA, S., 1991. Optimal planning of unreplicated field trials in recurrent selection. Dissertation Hohenheim.
- COCHRAN, W.G., 1951. Testing a linear relation among variances. *Biometrics* 7, 17-32.
- COCHRAN, W.G., 1957. Analysis of covariance: Its nature and uses. *Biometrics* 13, 261-281.
- COCHRAN, W.G. and G.M. COX, 1957. Experimental designs. Wiley, New York.
- DRAPER, N.R. and H. SMITH, 1998. Applied Regression Analysis. 3rd ed. Wiley, New York.
- FEDERER, W.T., 1955. Experimental Designs. Mcmillan Comp., New York.
- GOULDEN, C.H., 1952. Methods of Statistical Analysis. Wiley, New York.
- HARTLEY, H.O., 1956. A plan for programming analysis of variance for general purpose computers. *Biometrics* 12, 110-122.
- HEALY, M.J.R. and M.H. WESTMACOTT, 1956. Missing values in experiments analyzed on automatic computers. *Appl. Statist.* 5, 203-206.
- HENDERSON, C.R., 1953. Estimation of variance and covariance components. *Biometrics* 9, 226-252.
- JENSEN, N.F., 1976. Floating checks for plant breeding nurseries. *Cereal Research Communications* 4, 285-295.

- KEMPTON, R.A., 1984. The design and analysis of unreplicated field trials. Proc. of Fifth Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Hohenheim, Vortr. Pflanzenzüchtg. 7, 219-242.
- KEMPTON, R.A. and P.N. FOX (eds.), 1997. Statistical Methods for Plant Variety Evaluation. Chapman&Hall, London.
- KNAPP, S.J. and W.C. BRIDGES, 1987. Confidence interval estimators for heritability for several mating and experimental designs. Theor. Appl. Genet. 73, 759-763.
- LE ROY, H.L. und W. GLUCKOWSKI, 1961. Die Bestimmung der Varianzkomponenten im a.b.c-Faktorenversuch mit ungleichen Klassenfrequenzen. Biometr. Zeitschrift 3, 73-91.
- MODE, C.J. and H.F. ROBINSON, 1959. Pleiotropism and the genetic variance and covariance. Biometrics 15, 518-537.
- PATTERSON, H.D. and E.R. WILLIAMS, 1976. A new class of resolvable incomplete block designs. Biometrika 63, 83-92.
- PERKINS, J.M. and J.L. JINKS, 1968. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. Heredity 23, 339-356.
- PREECE, D.A., 1971. Iterative procedures for missing values in experiments. Technometrics 13, 743-753.
- SEARLE, S.R., 1971. Linear Models. Wiley, New York.
- SEARLE, S.R., 1982. Technical reports on variance component estimation. Cornell Univ. See <http://www.biom.cornell.edu/Publications/aco.html>.
- SNEDECOR, G.W. and W.G. COCHRAN, 1967, 6th ed.; 1980, 7th ed. Statistical Methods. Iowa State University Press, Ames.
- TIETJEN, G.L., 1974. Exact and approximate tests for unbalanced random effects designs. Biometrics 30, 573-581.
- UTZ, H.F., 1972. Die Zerlegung der GenotypxUmwelt-Interaktionen. EDV in Medizin und Biologie 3, 52-59.
- UTZ, H.F., 1978. Ein Algol-60-Programm fuer die Analyse von balancierten faktoriellen Versuchsanlagen. EDV in Medizin und Biologie 9, 34.
- WILLIAMS, E.R., 1977. Iterative analysis of generalized lattice designs. Austr. J. Statist. 19, 39-42.
- WILLIAMS, E., PIEPHO, H.-P., and WHITAKER, D., 2011. Augmented p-rep designs. Biom. J. 53, 19-27
- WRIGHT, A.J., 1971. The analysis and prediction of some two factor interactions in grass breeding. J.agric. Sci., Camb. 76, 301-306.
- YATES, F., 1933. The analysis of replicated experiments when the field results are incomplete. Emp.J.Exp.Agric. 1, 129-142.

- YATES, F. 1936. A new method of arranging variety trials involving a large number of varieties. *J. agric. Sci., Camb.* 24, 424-455.
- YATES, F. and W.G. COCHRAN, 1938. The analysis of groups of experiments. *J. agric. Sci., Camb.* 28, 556-580.

APPENDIX A List of Statements

! text until end of line [see 3.2.1]
 ! text after ! is a comment

BASIC/Q NROW NIDT NINPT NVAR [see 4.1.1]
LATTICE/Q T S K R NINPT NVAR [see 5.1.1]
ANOVA/Q NIDT NINPT NVAR [see 6.1.1]
UNREP/Q NPARZ NROW NCOL NSTD NIDT NINPT NVAR [see 7.1.1]

with NROW = number of units or rows in the data set
 NIDT = number of alphanumeric fields identifying the data rows
 NINPT = number of variables to be read
 NVAR = number of variables to be analysed

 T = number of treatments
 S = number of incomplete blocks per replication
 (if RCB, S=1)
 K = number of plots per incomplete block
 (if RCB, K=1)
 R = number of replications

with 4-digit qualifier Q (default = 1110):
 MISS = 0 zeros are not interpreted as missing values
 = 1 zeros are treated as missing values (default)
 EXTR = 0 no test on extreme values
 = 1 test of residuals on extremes (default)
 = 2 test of residuals and effects on extremes (ANOVA only)
 PRIN = 0 no printing of ANOVA table (in LATTICE only)
 = 1 normal printing (default)
 = 2 effects are printed instead of mean values (LATTICE only)
 Spearman's correlation coeff. are printed (BASIC only)
 = 3 extended printing of missing data (in ANOVA and LATTICE only)
 NEWF = 0 mean values not saved on the secondary output file (default)
 = 1 mean values saved on secondary file
 = 2 transformed and/or adjusted plot values saved
 = 3 adj. means and adj. plot values saved (LATTICE only)

Additionally,

for formatted reading append \$F with an FORTRAN like
 format, e.g.

ANOVA 6 0 5 5 \$F 15X, 2(F2.1), F5.2, 3X, F4.1

for the output of incomplete block effects on secondary file and for
 augmented designs with n augmented plots

LATTICE . . . \$An [see 5.1.1]

for unsorted and incomplete data, e.g.:

ANOVA . . . \$ 2(1) 5(3) [see 6.1.11]

INPUT list of variables (first 5 characters are used) [see 3.2.2]

CALCULATE X2=X6/X7*100 X3=X3*X4-100 ! some examples [see 3.2.3]

X2=LN(X2) X3=SQRT(X8) X4=EXP(X4) X5=ARCSINST(X10)

X15=X1>0.5 X15=X3==100 X15=X14+(X1>=0.5)*10

VARIABLE NAMES list of variables (first 5 characters are used) [see 3.2.4]

Statements connected with ANOVA (randomized complete block as example)

[see 6.1.2 and 6.1.7]

FACTORS TREAT=15 REPLICATION = 3 (must be first line) [see 6.1.2]

MODEL R + T + RT (use the first character of factor names) [see 6.1.3]

RANDOM R T [see 6.1.4]

MEANS RT TA [see 6.1.5]

EFFECTS TA [see 6.1.5]

RANKS TA [see 6.1.5]

SUBINT RT [see 6.1.6]

RESTRICT X1 [see 6.1.10]

TBT_TAB T (necessary, if qualifier NEWF=1 or [see 8.1]
PLOT is chosen to define the means to be handled)
\$F 2(F5.0), 3(F5.1), ... to formatize
\$T Text ... to entitle
\$S X2 to sort mean table in X2
\$V X3p X3a X1 Yield.p to choose variables

HERIT T or HERIT T:C [see 6.1.8]

GENOT/Q name e.g. GENOT T or GEN T:BC [see 6.1.9]
with name = effect symbol

Q = 1 variance-covariance analyses additionally

FIT/Q Y X1 X2 ... (only with BASIC, one line per regression)
e.g. FIT X5 X1 X2 [see 4.1.2]

with 3-digit qualifier Q (default 111):

PRIN = 0 only regression analyses are printed

= 1 covariance matrix of regression coefficients
is printed additionally (default)

= 2 covariance matrix of variables is printed additionally

RSDL = 0 no residuals are printed

= 1 test on outliers (default)

= 2 fitted values and residuals are printed on the secondary file

ICPT = 0 regression without intercept

= 1 regression with intercept (default)

NAMES_OF_TR/Q list of texts [see 3.2.5]
 (by default first 5 characters per name used)
 with Q = length of texts, e.g. NAMES/12
 or for ANOVA:
 Q = factor symbol and length of text, e.g. NAMES/T12

CHECKS list of treatment identifiers or * [see 3.2.6]
 for ANOVA with qualifier factor symbol, e.g. CHECKS/T 24

CSV/Q list of five identifiers [see 3.2.7]
 with Q = E for English csv format (decimal point and comma as separator)
 = G for German csv format (decimal comma and semicolon)

INDEX 1 $X4 = -0.5 \cdot X1 - 1.33 \cdot X2 + 0.77 \cdot X3 + 100$ [see 8.2]
 $X11 = MC(X2, 1, 2, 3)$

PLOT/Q X1.X2 X5.X2 X3.X3 ! as example [see 8.3]
 with Q = 1 one scatter plot per page (default)
 = 2 two scatter plots per page
 = 3 three scatter plots per page

PLOT/n Xi.Xj [see 8.4]
 PostScript scatter plots for Xi on Xj
 n = n-th character of treatment names plotted instead of numbers

ERROR P n [see 8.5]
 with P = factor symbol of the single experiments
 (or number of trials with DFE and SE lines)
 n = column which DFE, .. start in the data (default 12)

OMISS list of NVAR real numbers [see 8.6]

RUN [see 3.2.8]
 after RUN the data follow

EOD [see 3.2.9]

STOP [see 3.2.10]

APPENDIX B Examples

1. Examples with BASIC on file `psdbasc.dat`

- PSDBASC-1 = Simple example with PLOT
- PSDBASC-2 = With a correlation matrix
- PSDBASC-3 = Multiple regression test (Amer. Statist. 1984)
- PSDBASC-4 = Multiple linear regression (SNEDECOR and COCHRAN, 1980, p. 336)
- PSDBASC-5 = Polynomial regression (SNEDECOR and COCHRAN, 1980, p. 399)
- PSDBASC-6 = Orthogonal polynomials (SNEDECOR and COCHRAN, 1980, p. 405)
- PSDBASC-7 = Regression through origin (SNEDECOR and COCHRAN, 1980, p. 173)
- PSDBASC-8 = Spearman's correlation (SNEDECOR and COCHRAN, 1980, p. 192)

2. Examples with LATTICE on file `psdlatt.dat`

- PSDLATT-1 = Simple lattice (COCHRAN and COX, p. 406 and 413)
- PSDLATT-2 = Randomized complete block (RCB)
with use of CALCULATE, CHECKS, INDEX, and PLOT
- PSDLATT-3 = Triple lattice (COCHRAN and COX, p. 418)
- PSDLATT-4 = Triple rectangular lattice (FEDERER, 1955, p. 347)
with use of qualifier
- PSDLATT-5 = 10*8 Generalized lattice
- PSDLATT-6 = Triple lattice with several missing data types
- PSDLATT-7 = Randomized complete block (RCB) with several missing data types
- PSDLATT-8 = 5*5 Simple lattice (SCHNELL, 1957)
- PSDLATT-9 = 7*3 Generalized lattice with tree replications

3. Examples with ANOVA on file `psdanov.dat`

- PSDANOV-1 = Split plot design (SNEDECOR and COCHRAN, 1980, p. 327)
- PSDANOV-2 = Randomized complete block (RCB)
with use of CALCULATE, CHECKS, INDEX, and PLOT
- PSDANOV-3 = 3-factor experiment (SNEDECOR and COCHRAN, 1980, p. 318)
with TBT_TAB and MEANS
- PSDANOV-4 = Subdivision of a two-factor interaction, phenotypic stability
(YATES and COCHRAN, 1938)
- PSDANOV-5 = North Carolina Design I
- PSDANOV-6 = Series of rye experiments with use of ERROR and INDEX
- PSDANOV-7 = Calculation of expectation of mean squares
in a split plot over sites and years

4. Examples with UNREP on file `psdunrep.dat`

- PSDUNREP = Example data for designs with standards or moving averages

APPENDIX C MODEL Lines for Several Designs**1. Randomized complete block design with one factor**

with T_treatments and R_replications,
over P_places, Y_years, and E_regions, respectively

Single experiment:	MODEL R + T + RT
Series over places:	MODEL P + R:P + T + TP + RTP
Series over places and years:	MODEL Y + P + YP + R:YP + T + TY + TP + TYP + RTYP
Series over regions and places:	MODEL E + P:E + R:EP + T + TE + TP:E + RTEP
Series over regions, places, and years	MODEL Y + E + YE + P:E + YP:E + R:YEP + T + TY + TE + TYE + TP:E + TYP:E + RTYEP

2. Randomized complete block design with two factors

Factor A, Factor T, the AT combinations within R_replication completely randomized;
over P_places, Y_years, and E_regions, respectively

Single experiment:	MODEL R + A + T + AT + RAT
Series over places:	MODEL P + R:P + A + T + AT + AP + TP + ATP + RATP
Series over places and years:	MODEL Y + P + YP + R:YP + A + T + AT + AY + TY + ATY + AP + TP + ATP + AYP + TYP + ATYP + RATYP
Series over regions:	MODEL E + P:E + R:EP + A + T + AT + AE + TE + ATE + AP:E + TP:E + ATP:E + RATEP
Series over regions, places, and years:	MODEL Y + E + YE + P:E + YP:E + R:YEP + A + T + AT + AY + TY + ATY + AE + TE + ATE + AYE + TYE + ATYE + AP:E + TP:E + ATP:E + AYP:E + TYP:E + ATYP:E + RATYEP

3. Split-plot design with two factors

Factor A in main plots, factor T in sub plots within main plots,
with R_replications on P_places, Y_years, and E_regions

Single experiment: MODEL R + A + RA/ + T + AT + RAT
Series over places: MODEL P + R:P + A + AP + RA:P/
 + T + AT + TP + ATP + RATP
Series over places MODEL Y + P + YP + R:YP
and years: + A + AY + AP + AYP + RA:YP/
 + T + TY + TP + TYP
 + AT + ATY + ATP + ATYP + RATYP
Series over regions: MODEL E + P:E + R:EP
 + A + AE + AP:E + RA:EP/
 + T + TE + TP:E
 + AT + ATE + ATP:E + RATEP
Series over regions, MODEL Y + E + YE + P:E + YP:E + R:YEP
places, and years: + A + AY + AE + AYE + AP:E + AYP:E
 + RA:YEP/
 + T + TY + TE + TYE + TP:E + TYP:E
 +AT +ATY +ATE +ATYE +ATP:E +ATYP:E
 + RATYEP

4. Split-block design with two factors (stripe-design)

Factor A in main plots, factor T in crossed main plots,
with R_replications on P_places, Y_years, and E_regions

Single experiment: MODEL R + A + RA/ + T + RT/ + AT + RAT
Series over places: MODEL P + R:P + A + AP + RA:P/
 + T + TP + RT:P/
 + AT + ATP + RATP
Series over places MODEL Y + P + YP + R:YP
and years: + A + AY + AP + AYP + RA:YP/
 + T + TY + TP + TYP + RT:YP/
 + AT + ATY + ATP + ATYP + RATYP
Series over regions: MODEL E + P:E + R:EP
 + A + AE + AP:E + RA:EP/
 + T + TE + TP:E + RT:EP/
 + AT + ATE + ATP:E + RATEP
Series over regions, MODEL Y + E + YE + P:E + YP:E + R:YEP
places, and years: + A + AY + AE + AYE + AP:E + AYP:E
 + RA:YEP/
 + T + TY + TE + TYE + TP:E + TYP:E
 + RT:YEP/
 + AT +ATY +ATE +ATYE +ATP:E +ATYP:E
 + RATYEP