Data Transformations

Overview: Use the Data Transformations dialog to create a new, calculated assessment data column from one or more existing ARM data columns.

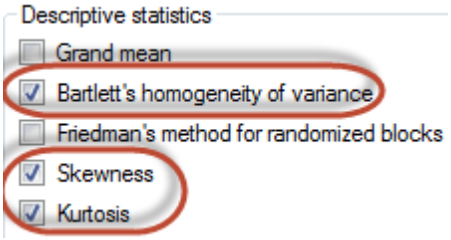
To display the Data Transformations dialog: open a trial, on Navigation Bar choose Assessment Data, then click Tools on menu bar and choose Transform.

Following are the different **categories** of data transformations. The information below with additional background and descriptions are included in ARM Help.

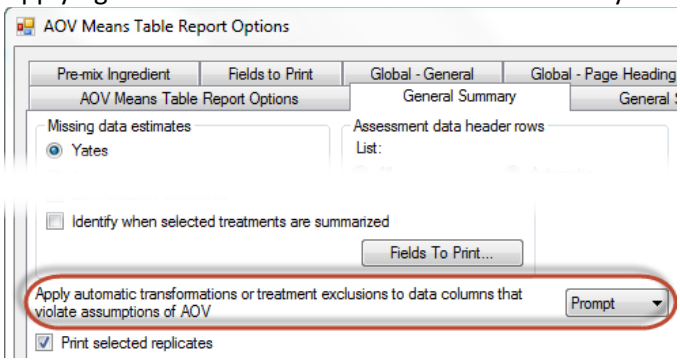
All data transformations are identified and linked to the calculated data column by a **transformation code** in the **ARM Action Codes** data header field of the **calculated column**.

1) **Data correction** - apply a data correction transformation only **when data violates assumptions of AOV**, either **non-normality** (Skewness, Kurtosis) or **heterogeneity** of treatment variance. One statistics reference describes data correction transformations as "What to Do When Data Breaks the Rules".

   Note: The **most efficient and statistically valid way to apply data correction transformations** is when ARM identifies there is evidence that a data column violation AOV assumptions during Analysis of Variance, typically while printing AOV Means Table report. The violation test statistics can also be printed on this report:



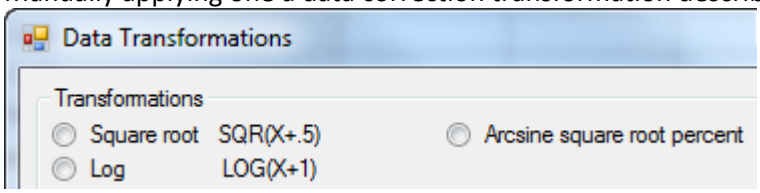
   Applying automatic data corrections is controlled by this General Summary option:



   The **benefits of** using **automatic data corrections** are:

   - Applied only when justified by AOV (it is generally considered ***inappropriate*** to apply a data correction transformation for a data column for which there is **no evidence of a violation**)
   - Applied when analyzing the original data column, so does not add transformed data columns that are only for data analysis.

   Manually applying one a data correction transformation described below is seldom necessary:



   a) Square root - for counts of a "rare event", or percentages ranging either from 0-20 or 80-100 SQR(X+0.5) accepts 0 and avoids over-correcting for values less than 10.

   b) Arcsine square root percent - for percentages or proportions, especially when range of percentages between treatments exceeds 40.

c) Log - for positive numbers that cover a wide range and all values are > 10
LOG(X+1) acts like square root for small values and log for large values.

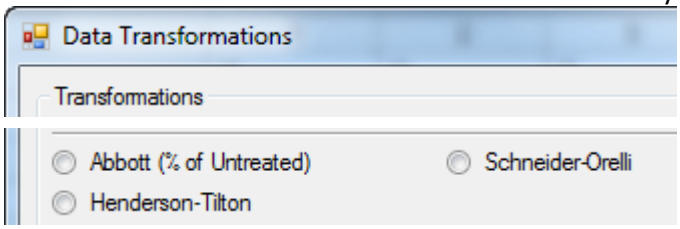
2) **Subsamples** - all calculate one value per 'plot' experimental unit based on subsample values in the original data column:

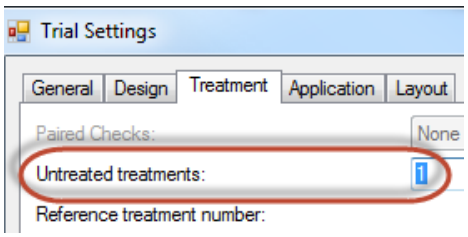
Transformation	ID	When to Use
Average subsamples	TAS[n]	Calculate one mean per experimental unit
Sum subsamples	TSS[n]	Calculate one total per experimental unit
% Incidence (0=none)	TIO[n]	Calculate the percentage of subsample values per experimental unit that are greater than 0
% Incidence (1=none)	TII[n]	Calculate the percentage of subsample values per experimental unit that are greater than 1
Count within range	TIN[n,r1,r2]	Calculate one count per 'plot' experimental unit of how many subsample entries in that experimental unit have assessed values within a specified range.

Count within range - calculated value is the count "per plot" of how many subsamples in the original data column are within the specified value range.
Note: this transformation must be entered into the ARM Action Code field; select it from the ARM Action Codes validation list, where it is listed as "TIN[n,r1,r2]"
For example, "TIN[1,76,100]" counts the number of subsamples in data column 1 that are between 76 and 100.

To calculate the **percentage** of subsamples per plot within the range of 76 to 100, enter this **user formula**:
&@IN([1],76,100)/@IN([1],0,100)*100

3) **Control correction** - use to calculate a corrected efficacy % relative to untreated treatments.





a) Abbott - for infestation/live individual measurements on a uniform population.
Note: Identify on **Treatment tab of Settings** dialog the list of check/untreated treatments in trial.

Four different Abbott transformations are offered:

i) TAB - on Tools - Transform dialog: calculates Abbott (% of check/untreated) per 'plot' experimental unit from check/untreated plot in same replicate, for when check assessments are **reasonably consistent** across replicates

ii) @UTAB - on ARM Action Codes validation list: calculates Abbott per plot from check/untreated treatment mean, for use when check treatment assessments **vary greatly across replicates**

iii) APC - on ARM Action Codes validation list: calculates on AOV Means Table report an Abbott per treatment from check/untreated treatment mean, listed in parentheses below treatment mean of assessed values.

iv) @TTAB - on ARM Action Codes validation list: calculates Abbott per treatment from check/untreated treatment mean. This approach is **not recommended**, because it **produces a non-analyzable data column.** Use APC instead (above).

b) Schneider-Orelli - for mortality measurements on a uniform population

c) Henderson-Tilton - for infestation/live individual measurements on non-uniform population

d) Sun-Shepard - for mortality measurements on a non-uniform population (not offered in ARM because this transformation has never been requested by a client)

4) **AUDPC** - Area Under Disease (or **any repeated assessment**) Progress Curve, traditionally used to summarize a plant disease infestation over time.
Can also be considered as Area Under Assessment Progress Curve (AUAPC), because the transformation is also appropriate for other assessments over time such as insect count or damage, crop response, growing degree days. Use AUDPC to summarize season-long performance of repeated assessments, especially when differences are less evident from individual data columns. (See example AUDPC_1 trial in Tutorial.)

For the AUDPC transformation, **data column selection is different than other transformations**, because you identify how ARM can automatically choose appropriate data columns to include in the AUDPC calculation. By using this approach, **new data columns are automatically added** to the AUDPC calculation as additional assessments are made during the season.

Steps to select data columns are:

a) Use "Convert data column number" on the Data Transformation dialog to identify **one data column** to include in the AUDPC. (By default, the current data column is selected.) ARM automatically selects all other data columns to include in the AUDPC calculation based on the selection rule defined in step c).

b) Select the OK button to close the Data Transformation dialog.

c) Use the "Define Data Column Matches for AUDPC" editor to identify data header field entries that are present in all data columns to include in this AUDPC calculation. Match entries almost always include Pest (e.g. **ERYSGT**), Rating Data Type (e.g. **COUDIS**), and Rating Unit (e.g. **%**). Other match fields may be included as needed to select the desired data columns to include in the calculated AUDPC.
Note: Define enough different data header field entries to uniquely identify columns across dates, to avoid including different types of measurements. For example, if two different leaf positions are assessed on each date, then include Part Rated to select only one leaf position per calculated AUDPC column.

d) Finally, confirm the selection from step c) by verifying that correct data header rows are defined for matching entries in the "Convert data column number" column with other identical data columns in this trial.

Starting date used for the AUDPC formula is the earliest date for the automatically selected data columns.

5) **SAUDPC** - Standardized Area Under Disease Progress Curve = AUDPC / Days, so AUDPC can be compared or summarized across trials. Functionally, SAUDPC is the **average AUDPC per day**.

6) **User Formula** - for entering ARM formulas that offer Excel-like automatic recalculation features. ARM formula operators are described in the "Data Transformation Formula" Help topic (see item "b)" below for an easy way to find this topic).

a) Click into User Formula, then press F9 to see some simple examples:

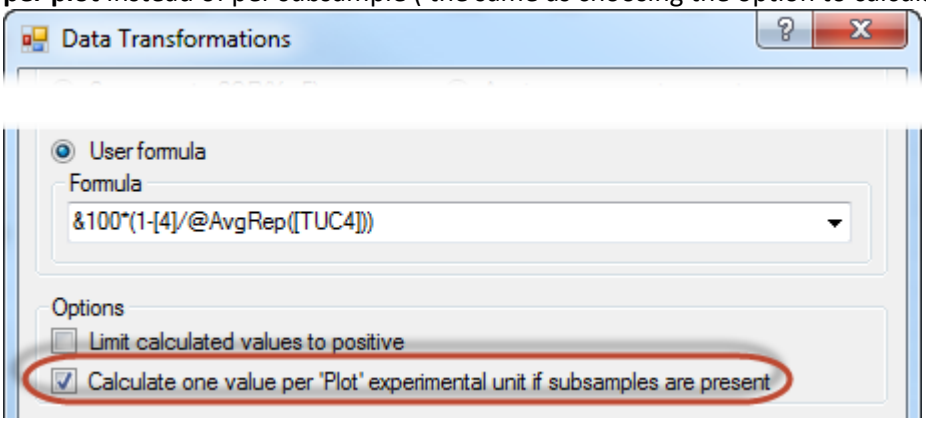
([1]+[2])/2	Average two data columns

@AVG([1].[3])	Average 3 data columns
100*(1-([3]/[TUC3]))	Percent control, TU is untreated treatment number
(1-([2]*[TUC1])/([TUC2]*[1]))*100	Henderson-Tilton, [1] is pre-count, [2] is post-count, TU is untreated treatment number
(1-@AvgRep([?])/@AvgRep([C?TU]))*100	Abbott's (% of check) based on treatment means.
& ([3]*2+[4]*7+[5]*18+[6]*38+[7]*75)/10	Index for mite classes 0-5

b) Click Help button on Data Transformation dialog, then scroll down and click User Formula link. The most **important points to remember when defining a user formula** are:

i) Enclose data column numbers in brackets, such as in the formula to add columns 1 and 2: ([1]+[2])/2 Relative column numbers such as [-1] are also supported, where "-1" is 1 column left of the destination column. This allows using the same saved formula for multiple data columns.

ii) Put **&** at beginning of formula to average any subsamples to a single plot mean, calculating only **1 value per plot** instead of per subsample (the same as choosing the option to calculate only 1 value per 'Plot'):



(The formula above is @UTAB, to calculate Abbott per plot from average of untreated treatment.)

c) Click into **ARM Action Code** data header field and:

i) Display validation list for many pre-defined user formulas.

ii) To use, select desired transformation from list, and then

iii) Enter original data column numbers for any [n] or [m] data column number references.

Some examples are:

TIN[n,r1,r2]	Count subsamples per plot within the range r1 to r2 (n=column)
TST[n]	Standardize data column according to collection basis and sample size data header entries in data column 'n' by converting to data header entries in the current data column
TCW[n]	Check Weighted Means conversion for multi-check trials (n=column, uses multiplier to adjust)
TCC[n]	Check Corrected Means conversion for multi-check trials (n=column, uses additive method to adjust)
@EC14[n]	EPPO Rating Scale 1 to 4 (n=scale 1 total column)
@EC14R	EPPO Rating Scale 1 to 4 (scale totals are immediately left of this index column)
@DS05[n]	Disease Severity from subsamples for 0-5 scale (n=column)
@DI05[n]	Percent Incidence from subsamples for 0-1, 0-2, ..., up to 0-5 scales (n=column)
@DIP16[n]	Percent Incidence for EPPO 1-6 disease scale from ratings summarized at plot level as separate data columns (n=scale 1 total column)
@SG06[n]	Stover-Gauhl disease severity from subsamples for 0-6 scale (n=column)
@TH07[n]	Townsend-Heuberger 0-7 disease scale from ratings summarized at plot level (n=scale 0 total column, 0=no attack)
@UTAB[n]	Abbott per plot from mean of untreated treatment (n=column)
@TTAB[n]	Abbott per treatments calculated from treatment means (n=column)

@PUAB[n,m]	Abbott per plot from paired untreated plots (n=data column of treatment assessment, m=data column of paired untreated plot assessment)
@APMAB[n,m]	Adjusted percent mortality using Abbott adjustment for natural check mortality (n=observed mortality data column, m=number treated data column)
@UPOC[n]	Percent of control (like APOC) relative to untreated treatment (untreated is 100%, change 'n' to data column to transform)
@POR[n]	Percent of reference treatment (reference is 100%, change 'n' to data column to transform)
@GMHA[n]	Gross Margin per hectare of crop (change 'n' to yield column, requires cost entries in treatment editor)
@TTHT[n,m]	Henderson-Tilton per treatment (n=pre-treatment column, m=post-treatment column)
@HB011[n]	Percent using Horsfall-Barrett 0 to 11 rating scale (0=0%, 11=100%, n=column)
@AVGNOT0[n]	Average per plot of all subsample values greater than 0 (n=original subsample data column number)

The following examples all use "@IN([column],r1,r2)", which returns count of data points within subsamples per experimental unit that are within the range from r1 to r2:

Stover-Gauhl calculated on a 0-6 scale (the formula ARM uses for SG06[n]):
&((@IN([?],1,1)*1+@IN([?],2,2)*2+@IN([?],3,3)*3+@IN([?],4,4)*4+@IN([?],5,5)*5+@IN([?],6,6)*6)*100)/(6*@IN([?],0,6))

Horsfall-Barrett 0-11 scale (the formula ARM uses for HB011B[n], where 0=1.17%, 11=98.2%):
&(@IN([?],0,0)*1.17+@IN([?],1,1)*2.34+@IN([?],2,2)*4.68+@IN([?],3,3)*9.37+@IN([?],4,4)*18.75+@IN([?],5,5)*37.5+@IN([?],6,6)*62.5+@IN([?],7,7)*81.25+@IN([?],8,8)*90.63+@IN([?],9,9)*95.31+@IN([?],10,10)*97.66+@IN([?],11,11)*98.82)/@IN([?],0,11)