Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture   2001 - 13th Annual Conference Proceedings

DEVELOPING SYSTEM FOR AUDITING THE DATABASE OF AGRICULTURAL TRIAL

Olga Susana Filippini de Delfino

Ana AguUa

Hugo Delfino

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

Delfino, Olga Susana Filippini de; AguUa, Ana; and Delfino, Hugo (2001). "DEVELOPING SYSTEM FOR AUDITING THE DATABASE OF AGRICULTURAL TRIAL," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1231

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

DEVELOPING SYSTEM FOR AUDITING THE DATABASE OF AGRICULTURAL TRIAL

Olga Susana Filippini de Delfino
Agronomy School
University of Buenos Aires

Ana Agulla
Basic Sciences Department
University of Luján.

Hugo Delfino
Statistician
Former Operations Director ACNielsen Argentina

Abstract

Many steps are involved in getting data from an experimental unit of an agricultural trial into a final report. Each step may introduce a great variety of errors. Building quality into systems is much more productive than building checks onto the end. Poor quality database have effects on final study results in terms of estimation, significance testing and power; but auditing agricultural trial is a complex process designed to ensure that it will provide a reliable answer to the question being posed. By introducing digit errors into database in a tomato assay, with small sample size, we demonstrate that simple ranges checks allows to detect and therefore correct, the main errors that impact the final study results and conclusions. For investigating significance level and power, two groups of data were simulated, having identical distributions and variances, but different population means. T-tests were carried out and relative frequencies of rejecting Null Hypotheses were determined. We have demonstrate that simple random errors in data affect the conclusions and that some form of data checking is required. Two different methods are analyzed and recommended, exploratory data analysis with and without a second data entry. On the other hand, not all errors that are found by exploratory data analysis are detectable by double data entry.

Key Words: Quality, error rate, double data entry, exploratory data analysis.

1.Introduction

Monitoring a trial refers to oversight of all aspects of its conduct, many issues must be considered during the design, conduct, and analysis stages of trials. In contrast, auditing a trial refers to the process of ensuring that the paper record and the electronic record of the trial correspond to what actually happened to the experimental units [1].

In general, many steps are involved in getting data from an experimental unit into a final report.[2] Each stage may introduce a variety of error. Data entry may be the only one, where quite routinely, we process everything twice.[3]

We suggest studying the effect of the different errors in databases, especially in small samples, on the final conclusion of the investigation, and discussing a system of auditing thus providing a reliable answer to the question being posed.

New Prairie Press
https://newprairiepress.org/agstatconference/2001/proceedings/21

2.The model

We consider a small sample size of tomato crops, where the concentration of titratable acidity of two different varieties were analyzed (each for different groups of 18 homogenized preparations) in order to establish if there were any differences in the concentration depending on variety and considering each group of measurements as identical and independently distributed as normal.

1000 similar trials with $n_1= n_2=18$ were simulated. A digit error was introduced at random in a figure generating two data bases; a correct one and an incorrect one.

Results were previously shown when a typical error was introduced in data base, such as transfer of digits or digit errors that could affect final conclusions.

We suggest three specific methods to check the quality of the data base:

1) Simple entry in addition to the use of statistical methods.
2) Simple entry with auditing of the information collected (record versus data entered) to determinate error rate.
3) Double data entry (by different people).

3.Methods

We used SAS [4] for all the simulations. We simulated 1000 random trials of size $n_1= n_2=18$ with properties similar to initial random samples $X_{i\,j}$ generated as independent observations from a Normal distribution.

Each observation was assigned a random digit R, uniformly distributed on (0,1)

We introduced errors into the values X_i, where $R_i < \gamma$ with γ=%erroneous data
X_i* was set equal to X_i

For those cases where $R_i < \gamma$ each of the digits was separated into three variables:

$X_{i1}* = \text{integer}\ (X_i)$
$X_{i2}* = \text{integer}\ (X_i - X_{i1}) *10$
$X_{i3}* = \text{integer}\ (X_i *100 - X_{i1} *100 - X_{i2} *10)$

The appropriate digit in X_i was then replaced by another digit randomly distributed on the integers 0 to 9. Then it was reconstructed.

One time in ten the random digit is expected to coincide with the original digit, and thus X_i* would remain unchanged.

For investigating significance levels, two groups of data were simulated, each having normal distributions, with the same variance and different means. The results obtained in each case were analyzed and compared to draw final conclusions. Summary statistics and "t" test were calculated for each replicate trial and for type I error rates.

4.Results

The Tables 1 and 2 show a typical small dataset illustrating the kind of errors that may be introduced and the possible problems. If no checks were applied, the errors would distort the means and standard deviations. This error impacts upon the analysis and changed the final conclusions. The table III shows the analysis of this dataset with true and erroneous data.

4.1.Simulation results

We consider the effect of errors on much smaller samples, in which one or two errors could contribute substantially to any summary statistics.

We simulated 1000 random trials and their errors. We established range checks in accordance with the original trial, considering as outlier value the one which exceeds ± 3 standard deviation of the average of the two samples of the original trial. In this case we must verify that the error results from data entry and is not a record error. Then the database is inspected and corrected.

The Table 4 gives the distribution of two groups for the true data, erroneous data and inspected data.

When only a small proportion of data is corrupted, the estimates of μ and σ^2 were incorrect and were very unstable. Similarly, the proportion of simulated trials does not reject incorrectly null hypothesis ($\mu_1 = \mu_2$) increase with the erroneous data. The impact on summary statistics and P- values is substantial, for (α= 5% and α=1%). Introducing errors into the data compromises final study results. (See Table 5 and Figure 1)

5. Summary and Conclusions

It has been suggested that double data entry is enough and necessary to ensure good quality essays.[5]

Nevertheless, double data entry without exploratory analysis assumes that records were correct and all errors result exclusively from data entry. This assumption would seem little reliable for, from this point of view, double entry would not trap errors made by researchers at every stage of the development of the study. The simplicity of exploratory analysis of data through the introduction of check ranges detected more than half of the errors introduced at random. It has also been demonstrated that simple errors introduced at random in the database affects the possibility of drawing valid conclusions and therefore we need to find a form of auditing the whole research data systematically. That is to say, there could be an impact on scientific development as a result of the bad design of the experiments, random problems, sample size, unrealistic projects or problems in the compilation of information. With regard to the conduction of a study, the research team should also be carefully selected.

To minimize problem in the detection of errors, it would be convenient to carry out histograms, range checks, crossed validation between variables that were expected to be correlated and validation between the same variables measured in repeated occasions and other methods that can be relevant to special situations.

With regard to the proposals for data entry, the ones suggested below should be analyzed according to the complexity of the research and the size of the sample.

We suggest, in concordance with Gibson et al [5]

Classification of error

In order to interpret the results of auditing, a method for classifying error or inconsistencies in collecting and recording data must be made. Errors in essential variables for interpretation of the trial are far more serious than errors in a secondary item .

For this reason:

Creating a hierarchical list of essential variables (primary endpoints, experimental units, etc.)

Critical variables (factors, outcome, variables, etc)

Less important variables

This, provides a system for prioritizing the auditing of the database.

Alternative methods for the auditing process of the database

Use of statistical methods (Exploratory analysis) **in addition to the simple data entry**

Simple entry with auditing of the information collected (record versus data entered to determine error rate)[6]

Double data entry (by different people)

New Prairie Press
https://newprairiepress.org/agstatconference/2001/proceedings/21

References

[1] Califf, R; Karnash S.and Woodlief L.. Developing Systems for Cost-Effective Auditing of Clinical Trials. Controlled Clinical Trials (1997) 18: 651-660

[2]Wyatt JC Clinical datasystem, part 1: data and medical records. Lancet 1994; 344:1543-1547.

[3]Byar D.P. Discussion following Neaton et al. *Stat. Med* 1990; 9:124

[4]SAS Institute Inc. SAS Language Reference. Version 8. Cary N.C. SAS Institute Inc. 1999.

[5] Gibson D., Harvey A.J., Everett V., et al. Is double data entry necessary? The CHART Trials. *Controlled Clinical Trials.* 1992;13:156-169

[6]Desgouilles R., Auger J.M., Tisserand B., et al. A quantified *internal* audit system allowing quality rating of clinical research. *Drug Inf J.* 1992;26:379-388

New Prairie Press
https://newprairiepress.org/agstatconference/2001/proceedings/21

Conference on Applied Statistics in Agriculture
Kansas State University

Table 1: Values of titratable acidity from samples of homogenized hybrid commercial tomato (*Lycopersicon esculentum Mill*, the traditional type) and another new material (True data)

Titratable acidity	**Genetic Material**	
Observation	Traditional	New
1	3,40	2.75
2	3.75	2.65
3	2.95	2.60
4	3.10	2.95
5	3.15	3.05
6	3.80	3.10
7	3.85	2.25
8	3.40	2.85
9	3.75	2.75
10	3.50	2.80
11	3.40	3.00
12	2.85	3.10
13	3.30	2.95
14	3.85	3.45
15	3.20	3.90
16	3.60	2.60
17	3.95	2.55
18	3.35	2.55

Note: The tomatoes were harvest randomly for two groups (traditional and new) on the same date. Each experimental unit consists of three tomatoes, divided in four parts and processed in homogenized preparations.

New Prairie Press
https://newprairiepress.org/agstatconference/2001/proceedings/21

Table 2: Values of titratable acidity from samples of homogenized hybrid commercial tomato (*Lycopersicon esculentum Mill*, the traditional type) and another new material (Erroneous data)

Titratable acidity	**Genetic Material**	
Observation	Traditional	New
1	3,40	2.75
2	3.75	2.65
3	2.95	2.60
4	3.10	9.5**
5	3.15	3.05
6	3.80	3.10
7	3.85	2.25
8	3.40	2.85
9	3.75	2.75
10	3.50	2.80
11	3.40	3.00
12	2.85	3.10
13	3.30	2.95
14	3.85	3.45
15	3.20	3.90
16	3.60	2.60
17	3.95	2.55
18	3.35	2.55

Note: The tomatoes were harvest randomly for two groups (traditional and new) on the same date. Each experimental unit consists of three tomatoes, divided in four parts and processed in homogenized preparations.

****: Erroneous data, true data 2.95**

Table 3: Analysis of small dataset with true and erroneous data_

	TRUE DATA		*ERRONEOUS DATA*	
	Group 1	**Group 2**	**Group 1**	**Group 2**
Mean	3.45277	2.88055	3.45277	3.24444
Standard dev.	0.32741	0.37422	0.32741	1.605
P (T≤t) one tail	0.000012		0.29653135	
P (T≤t) two tail	0.000024		0.5930627	
Difference (G1 vs. G2)	SIG		NS	

Standard dev.: Standard deviation

Table 4: Small sample (n_1= n_2=18). Results: Distribution of Summary statistics under the Null Hypothesis (μ_1=μ_2) when observation were true, corrupted or audited.

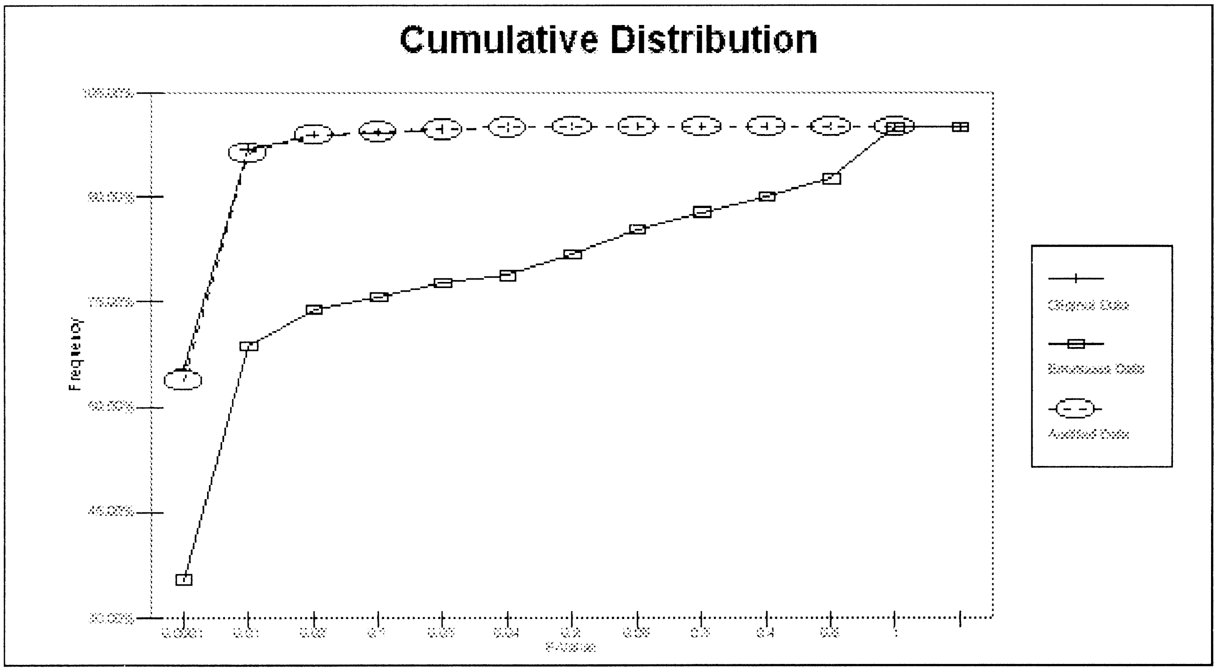
PARAMETER ESTIMATES				
	μ		σ	
	GROUP I	**GROUP II**	**GROUP I**	**GROUP II**
TRUE VALUES	3.452	2.880	0.327	0.374
TRUE DATA *	3.402	2.876	0.297	0.367
ERRONEOUS DATA *	3.535	2.925	0.460	0.557
AUDITED DATA *	3.403	2.877	0.299	0.372

(Standard deviation)
*Note: sample dwastributions considering 1000 simulations

Table 5: Small sample (n_1= n_2=18). Results: effect on P- values when α= 5% and α= 1%. Comparwasons of relative frequency of no reject.

RELATIVE FRECUENCY		
	α=0.05	α=0.01
TRUE DATA	0.004	0.033
ERRONEOUS DATA	0.221	0.315
AUDITED DATA	0.006	0.038

Figure 1: Cumulative distribution of relative frequency of no reject, for original, erroneous and audited


Cumulative Distribution
Frequency
P-Value
Original Data
Erroneous Data
Audited Data