Applying Statistical Sampling Plans To Data Entry Procedures To Increase Data Quality
Roderick Lashley, STATKING Consulting, Inc.

ABSTRACT
Double data entry is a widely used method to enhance the quality of clinical trials data entered into the computer from paper case report forms. King and Lashley [1] proposed an alternative procedure using single data entry and a procedure known in the statistical quality control literature as a continuous sampling plan. This paper will investigate applying this procedure to data where the entry errors are not randomly distributed, but occur in bursts. Empirical results for the cost to perform the procedure as well as computations for the gain in data quality will be presented.

INTRODUCTION
Maintaining data integrity in clinical trials is a very important process. It can also be very resource intensive. While some clinical trials are currently using remote data entry, a significant number of trials still collect the data on paper case report forms. This data must be accurately transcribed into the computer system for appropriate analysis. The most common procedure for this transcription procedure is double data entry.

Double data entry consists of two separate data entry technicians entering the data into the computer independently. Any discrepancies in the data entered are resolved either during the second entry process or as an additional step. Gibson, et.al. [2] have shown that in some cases the gain in data quality may not justify the cost of the process.

The continuous sampling plan (CSP) approach requires the data to be entered into the computer one time. A visual record verification check is then performed on the data. The number of records checked is a function of the data quality and the CSP plan selected. King and Lashley have shown the improvements in average outgoing quality (AOQ) by applying CSPs to data consisting of randomly occurring errors. This paper will investigate the results of applying CSPs to data with other types of error distribution.

ERROR MODELS
There are two common, non-random types of errors that appear in data entry. One type we will define as “burst” errors. These errors typically occur when there is a large amount of numerical data being entered. The technician may get “off-cut” in entering the data, causing several errors to occur in a short timeframe. Once the technician realizes the problem, the alignment is corrected and data entry continues with the original error rate. This can also be seen if the technician is frequently interrupted during the data entry process.
Another error model can be characterized by technician fatigue or interruption. As the day proceeds, the technician’s ability to concentrate erodes, yielding a slowly increasing error rate.

**SAMPLING PLANS**

To improve the quality of data for statistical analysis, a quality assurance (QA) process is performed against the data. Before the QA step, the data has an initial (input) error rate. The QA step reduces the error rate of the data yielding an output error rate less than or equal to the input error rate. One way to reduce the error rate is to apply a continuous sampling plan to the data as the QA process.

Two CSP plans were chosen for examination. A CSP-1 plan presented by Dodge [3] has a sampling frequency, f, of .1 and a clearing interval, i, of 10. The clearing interval is the number of consecutive records that must be free of data entry errors before the sampling of a fraction, f, of the records will begin. For the other plan used, f=.2, i=5. Many other CSP plans are given in the literature.

Execution of a CSP plan is easy. (1) Inspect i successive records. (2) If no errors are found, randomly sample fraction f of data records and check for errors. (3) Whenever an error is found, correct the record and repeat step 1.

The parameters needed to quantify the effectiveness of a CSP are average outgoing quality (AOQ) and the average fraction inspected (AFI). All other relevant measures of CSPs can be computed from these two values. In terms of the plan parameters,

\[
AOQ = p \left[ \frac{(1-f)q^i}{f + (1-f)q^i} \right]
\]

and

\[
AFI = \frac{f}{f + q^i(1-f)}
\]

where p is the proportion of incorrect data records in the file before sampling and q=1-p.

We can also compute the average records inspected (ARI=N * AFI), average time necessary to perform the visual records check (ATQA=AFI*N*t, t=time required to verify one data record), and the percent gain in average quality (PGAQ=1 - \frac{AOQ}{p} *100).

**SIMULATED DATA**

SAS® macro programs were written to simulate data from a clinical study. Macro variables were used to hold the input error rate, p, and number of observations to generate for each trial, N. This allowed the testing of various error rates and sample sizes. Each data record consisted of one variable. A ‘0’ was defined as a correct record, and a ‘1’ an incorrect record. The binomial random number generator was used to generate a sequence of correct and incorrect records. For each simulation, 10,000 sets of size N records were generated.
Each error rate model used a base error rate of \( p = 0.04 \). For computation purposes, we assumed that the visual record check would take 1 minute. The random model produced data with a constant error rate.

The burst model simulated a problem approximately twice a day, or every 250 records. The burst of bad quality lasted 25 records. Each record during the burst was generated with an error probability of 0.3.

An exponential model simulated an increasing error rate beginning approximately every 2 hours (125 records) and lasting for 30 records. At the end of the cycle, the probability of a bad record was approximately 0.33. The error rate model for this situation was

\[
p(t) = \begin{cases} 
  p + \exp(b(t - x)) & \text{if } 0 < t - x < 30 \\
  p & \text{otherwise}
\end{cases}
\]

where \( x = 125 \)

this model was repeated every 165 records.

For each non-random model, the error rate returned to the original rate at the end of the problem areas. Each simulated data set contained 5000 records.

RESULTS

For each error model, 10000 simulated data sets were generated. Each data set contained \( N = 5000 \) records.

In all cases shown, the simulated \( p \) and AOQ values closely matched the random errors theoretical values of \( p \) and AOQ. The results for the \( i = 10, f = .1 \) plan are shown in table I below.

<table>
<thead>
<tr>
<th></th>
<th>Random error model</th>
<th>Burst error model</th>
<th>Exponential error model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Error Rate ( p )</td>
<td>.0400</td>
<td>.0653</td>
<td>.0653</td>
</tr>
<tr>
<td>AOQ</td>
<td>.0343</td>
<td>.0470</td>
<td>.0469</td>
</tr>
<tr>
<td>AFI</td>
<td>.1433</td>
<td>.1874</td>
<td>.1874</td>
</tr>
<tr>
<td>ARI</td>
<td>717</td>
<td>937</td>
<td>937</td>
</tr>
<tr>
<td>ATQA (hours)</td>
<td>11.94</td>
<td>15.62</td>
<td>15.62</td>
</tr>
<tr>
<td>PGAQ (%)</td>
<td>14.25</td>
<td>28.02</td>
<td>28.18</td>
</tr>
</tbody>
</table>

The burst error model and the exponential error model yielded almost identical results in overall error rates, AFI, and AOQ. The PGAQ for both non-random models was almost double the random model. The average number of records inspected increased 31%.

Similar results were obtained for the \( i = 5, f = .2 \) sampling plan. These are shown in Table II.
### Simulation Results for I=5, f=.2 CSP Plan with N=5000, p=0.04

<table>
<thead>
<tr>
<th></th>
<th>Random error model</th>
<th>Burst error model</th>
<th>Exponential error model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Error Rate (p)</td>
<td>.0400</td>
<td>.0653</td>
<td>.0652</td>
</tr>
<tr>
<td>AOQ</td>
<td>.0306</td>
<td>.0413</td>
<td>.0413</td>
</tr>
<tr>
<td>AFI</td>
<td>.2348</td>
<td>.2675</td>
<td>.2675</td>
</tr>
<tr>
<td>ARI</td>
<td>1174</td>
<td>1338</td>
<td>1338</td>
</tr>
<tr>
<td>ATQA (hours)</td>
<td>19.57</td>
<td>22.29</td>
<td>22.29</td>
</tr>
<tr>
<td>PGAQ (%)</td>
<td>23.5</td>
<td>36.75</td>
<td>36.66</td>
</tr>
</tbody>
</table>

The PGAQ for the non-random models was over 13% higher than the random model while the number of records inspected was 14% more in the burst and exponential models. Inspecting more records yields better outgoing data quality, but it also increases the time necessary to execute the continuous sampling plan. In choosing a CSP, both AOQ and ATQA must be considered.

#### References


#### Author Contact

Roderick Lashley
STATKING Consulting, Inc.
759 Wessel Drive - Unit 6
Fairfield, OH 45014
Phone: 513-858-2989
Fax: 513-858-3022
email: rod@statkingconsulting.com

---

**CONCLUSION**

Errors in the data entry process for clinical trials research do not always fit the random error model. The data is often skewed by having bursts of “bad” data, or by having an increasing error rate over time. These distributions can be caused by fatigue, distractions, or by a technician getting the data misaligned with the data entry form.

Continuous sampling plans can be successfully used as an alternative to double data entry for these error models. In fact, simulations have shown that the percent gain in average quality is much greater for the non-random error models than for the random model.