

Using Continuous Sampling Plans to Quantify the Effectiveness of Quality Assurance Checks on Data from Clinical Trials

Dennis W. King, PhD
Rod Lashley, BS
STATKING Consulting Inc., Fairfield, OH

Introduction

One of the most important steps in the process of conducting clinical trials in all product areas is the transcribing of data from the case report forms (CRFs) into the computer. Because correct entry of the data into the computer is so critical to the analysis of the data, many quality assurance procedures have been developed to insure that incorrect data is not entered into the system.

For SAS users, the finished products from this part of the process are the cleaned, locked SAS data sets for the study. In this paper, an "extra" quality assurance check of the data is described which uses the SAS data sets containing study data that were created from double data entry. The effectiveness of this additional QA step can be quantified by what is known as a continuous sampling plan (CSP).

In this paper, the application of CSPs to the quality assurance process for data sets from clinical trials will be discussed. The parameters of the CSP plan will be presented along with examples of how to use these parameters to design efficient sampling plans for the QA of data sets from clinical trials.

The Data Quality Assurance Process at SCI

At our company, the data QA process involves many steps but there are two main checks of the data quality. The first involves an initial entry of the data into the computer by a data entry operator followed by a verification of the data by a second data entry operator. This is commonly referred to as double data entry.

After double data entry, we print out the SAS data set containing the information and perform a (usually less than 100%) visual check of the data. For a data file containing N records, a sample of the records is taken and these records are compared to the corresponding CRFs. This step in the quality assurance process insures that the data have been read into the SAS data set correctly and corrects a portion of any data entry errors that were not detected with double data entry.

The specific procedure that we follow is to compare the first 10 records in the data set to their corresponding case report forms. If these ten records are correct then the data entry technician checks every tenth record until another incorrect record is found. When an incorrect record is found, the data entry technician corrects the record and returns to 100%

checking of the data records until ten correct records are found and so on. This procedure is analogous to what is known in quality control as a continuous sampling plan (CSP).

Continuous Sampling Plans

Continuous sampling plans were invented to monitor the quality of a steady stream of individual items from a process. For a manufacturing process, CSPs are used to monitor the quality of products (such as TV sets, personal computers) on an assembly line. Quality inspectors sample products from the production line according to schemes dictated by a specific CSP. Various CSPs and their effectiveness have been discussed in quality control texts such as those by Schilling (1982) and Montgomery (1991).

For SAS data sets generated from clinical trial data, each observation in the data set can be thought of as a product on an assembly line. The quality inspector (data entry technician) moves down the line of data records sampling some of the records and checking them for correctness. This situation is similar to the sampling of products moving on an assembly line.

The CSP-1 Plan

Continuous sampling plans are referenced by the parameters f , the sampling frequency and i , the clearing interval. For our QA application, the clearing interval gives the number of records with no data entry errors that must be examined consecutively before the sampling of a fraction, f , of the records will begin. The most commonly used CSP is the CSP-1 plan given by Dodge (1947). A sketch of the logic for this plan, as it applies to our QA procedure, is shown below.

The values needed to quantify the effectiveness of the CSP-1 plan in improving the quality of a data set are the average outgoing quality (AOQ) and the average fraction inspected (AFI). The AOQ gives an indication of the quality of the data records after the extra QA step. The AFI gives the proportion of the records verified (on average) in the visual QA check. In terms of the plan parameters, these are

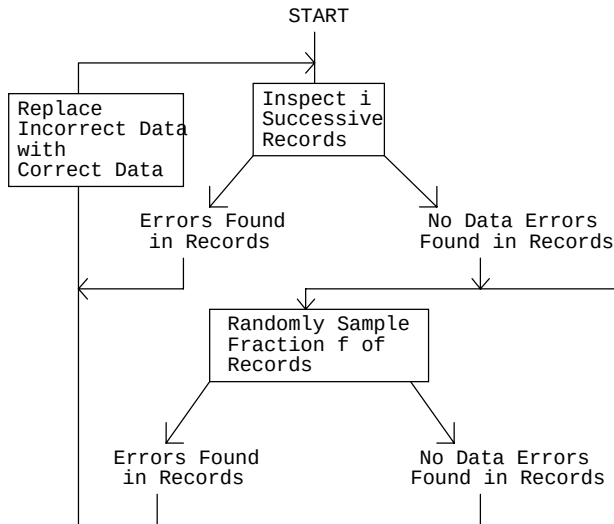
$$AOQ = p \frac{\sum_{i=1}^N (1-f)q^i}{N + (1-f)q^i}$$

and

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

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

Logic of CSP-1 Plan



The other relevant quantification measures are computed from AOQ and AFI. For a file of size N records, the average number of records inspected (ARI) is

$$ARI = N \cdot AFI$$

The average time necessary to perform the QA (ATQA) is then

$$ATQA = AFI \cdot N \cdot t = ARI \cdot t$$

where t is the time needed to QA one record in the data file. The relative gain in record quality from the extra QA step is the percent gain in average quality (PGAQ) given by the ratio

$$PGAQ = \frac{N}{p} \frac{AOQ}{p} \cdot 100$$

Application of CSPs to the QA of SAS Data Sets

The formulas in the preceding section can be used to evaluate the effectiveness of a CSP-1 plan used for data QA. The following discussion demonstrates the quantification of the effect on data quality of a particular CSP-1 plan.

For example, consider the QA plan which alternates between 100% verification and 10% verification as described earlier in the paper. This is a CSP-1 plan with $i=10$ and $f=.1$. Assuming that the proportion of defective records in the data file after double data entry is .001, the AOQ of this CSP-1 is

$$AOQ = (.001) \frac{N(1-.1)(.999)^{10}}{1 + (1-.1)(.999)^{10}} = .0009$$

The relative gain in record quality after this CSP-1 plan has been applied to the data set is $PGAQ=(1-(.0009/.001)) \times 100 = 10\%$. The average fraction of records verified in this QA step will be $AFI=.1009$. For a file of $N=10,000$ records, the average number of records inspected is $ARI=1,009$.

A key issue is the time needed to perform this extra QA. The average time needed to complete this QA step can be computed for any CSP-1 plan using the formula for ATQA. Continuing our example, assume it takes one minute to QA each record. The $ATQA=(1,009)(1/60)=16.82$ hours. For the CSP-1 plan with $i=10$ and $f=.1$, the ATQA, along with other CSP-1 plan parameters, are shown for various values of the proportion of incoming defective records in Table 1.

Table 1. ATQA for Various Values of the Proportion of Defective Records for the CSP-1 Plan with $i=10$ and $f=.1$ Assuming One Minute per Record for QA and $N=10,000$

Proportion Defective	AOQ	AFI	PGAQ	ARI (#Rec)	ATQA (hrs)
.01	0.009	0.11	10.0%	1094	18.24
.02	0.018	0.12	10.0%	1197	19.95
.03	0.026	0.13	13.3%	1309	21.82
.04	0.034	0.14	15.0%	1432	23.87
.05	0.042	0.16	16.0%	1565	26.09
.06	0.050	0.17	16.7%	1710	28.50
.07	0.057	0.19	18.6%	1867	31.12
.08	0.064	0.20	20.0%	2037	33.95
.09	0.070	0.22	22.2%	2220	37.00

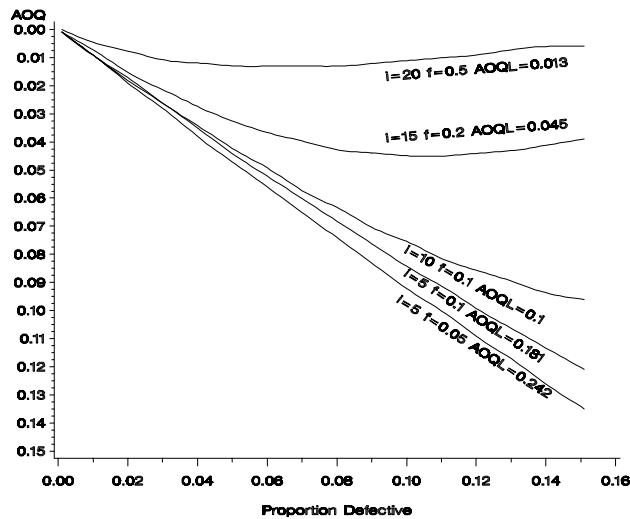
Selection of CSP-1 Plan Parameters

The ultimate goal of this paper is to demonstrate how to select a CSP-1 sampling plan for data QA. The selection of a CSP-1 plan involves choosing the plan parameters i and f based on the desired quality level after QA.

In order to select a CSP-1 plan for use in QA, a plot of AOQ versus the proportion of incoming defective records is necessary. This plot is shown for five CSP-1 plans in Figure 1. These plans provide a cross section of the available CSP-1 plans. The plans near the top of the graph require large amounts of sampling but produce excellent outgoing data record quality. The plans closer to the bottom of the graph require less sampling at the expense of outgoing data record quality.

For each of the CSP-1 plans in Figure 1, the average outgoing quality limit (AOQL) is drawn on the graph. The AOQL for a CSP-1 plan is the worst possible average outgoing data record quality expected regardless of the incoming proportion of defective records.

Figure 1. Plot of AOQ vs. Proportion Defective for Various CSP-1 Quality Assurance Plans



In order to use the graph in Figure 1, the user draws a vertical line from the incoming proportion of defective records (shown on the x axis) up to a particular curve. The user then extends a horizontal line across from the intersection of the vertical line with the curve to indicate the AOQ for that plan. This proceeds until a plan is located with the specified AOQ.

For example, assume the user has an incoming proportion of defective records of 6% and that he wishes the outgoing quality of the records to be 4% defective. Reading up from the bottom of Figure 1, the $i=5$ and $f=.05$ plan has AOQ of 5.75%, the $i=5$, $f=.1$ plan has AOQ=5.25% and the $i=10$, $f=.1$ plan has AOQ=5%. All of these plans produce unacceptable AOQ values. The $i=15$ and $f=.2$ plan has AOQ=3.75% so this would be the plan of choice.

SAS Macros

There are two SAS macros given in the appendix. The first, called CSPPLAN, prints the parameters discussed in the paper for any CSP-1 plan for a user specified number of records and QA time per record. The second, called CSPGRPH, graphs a set of CSP-1 plans on the same graph in order to allow the user to select a plan with appropriate AOQ. The code for both macros is shown in the appendix.

Conclusions

CSPs can be used to quantify the effectiveness of an additional visual quality assurance check on SAS data sets containing clinical trials data. Using the SAS macros shown in this paper, data management personnel can select a CSP-1 plan with a specified output data record quality target. Additionally, the data management personnel can compute the relative gain in data record quality for the use of any CSP-1 sampling plan and the time needed to perform the additional QA procedure.

REFERENCES

- Dodge, H.F. (1947), *Sampling Plans for Continuous Production, Industrial Quality Control*, vol. 14(3), 5-9.
- Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*, New York, NY: John Wiley & Sons.
- Schilling, E.G. (1982). *Acceptance Sampling in Quality Control*, New York, NY: Marcel Dekker, Inc.

SAS Code for CSP-1 Sampling Plan QA Procedure

```
%macro csplan(interval,fraction,plow,phi,pinc,t,n,dsn);
*****
* macro to create data set containing values
* for CSP-1 Sampling Plan
* STATKING Consulting, Inc.
* interval - parameter I from CSP-1
* fraction - parameter f from CSP-1
* plow - starting proportion defective.
* phi - ending proportion defective.
* pinc - amount to increment proportion defective.
* t - time to QA one record ( in hours).
* n - number of records.
* dsn - name of output data set.
*****
data &dsn;
length plan $12;
format pgaq percent7.1 afi 4.2;
f = &fraction; * fraction to sample;
i = &interval; *consecutive records to sample;
time = &t; * time to QA one record (minutes);
n = &n; * number of records;
plan = 'i=' || put(i,2.) || ' f=' || put(f,4.2);
* store plan info for use in plotting;

do p= &plow to &phi by &pinc;
q = 1 - p;
num = (q**i) * (1-f);
*compute values once and use later;
denom = f + num;
aoq = round(p * num / denom,.001);
pgaq = round(1 - (aoq / p),.000001);
* format will convert this to percent;
afi = round(f / denom,.00001);
ari = round(n * afi,1.);
atqa = round(afi * n * time,.01);
output;
end;
keep p q num denom aoq pgaq f afi atqa n time i ari plan;
%mend csplan;

%csplan(interval=10,fraction=.1,plow=.01,phi=.09,t=1/60,
n=10000,dsn=list,pinc=.01);

%macro csgrph(dsn,vorder,horder,numplans);
*****
** Macro to plot various CSP-1 plans.
** STATKING Consulting, Inc.
** dsn - input data set name
** vorder - order info for vertical axis.
** horder - order info for horizontal axis.
** numplans - number of plans being plotted.
*****
/* generate the symbols we will need */
%do i = 1 %to &numplans;
symbol&i i=join line=&i;
%end;

proc gplot data=&dsn ;
axis1 order=&vorder;
axis2 order=&horder label=("Proportion Defective")
minor=(n=1);

plot aoq*p=plan / vaxis=axis1 haxis=axis2 ;
title h=1 "Plot of AOQ vs. Proportion Defective for";
title2 h=1 "Various CSP-1 Quality Assurance Plans";
run;

%mend csgrph;
```

```
%cspgrph(dsn=all,vorder=0.16 to 0 by -.01, horder=0 to 0.16  
by .02, numplans=5)
```