

A Programmer's Guide to Statistical Procedures

Jim Edgington, Gilead Sciences, Inc., Foster City, California

ABSTRACT

Programmers and statisticians have distinct jobs in the creation of reports used in clinical development. Statisticians write required statistical methods and examine the data to determine its limitations or constraints. In turn, the statistician informs the programmer as to what type of SAS procedure to use. Programmers, with their specialized knowledge of SAS code, build a program around the procedure received from statisticians. Effective statisticians understand the programming side of the problem, foresee potential issues, and collaborate with the programmer. The goal of this paper is to give the programmer insight into how a statistician chooses the statistical method for a given set of data. As a result, a programmer can provide input to improve the quality of the final product and cut the report process development time.

INTRODUCTION

In programming for clinical trials, both programmers and statisticians have distinct roles to play in the development of statistical reports used in regulatory filings. The statisticians design and implement statistical analysis using protocols that provide more specific information in a Statistical Analysis Plan (SAP). Based on a SAP the programmers create analysis files and reporting programs to produce the tables, figures, and listings required to complete a clinical study report. While the jobs of a programmer and statistician are separate, the most effective statisticians understand the implications of their requests on programs. When posed with two equally valid statistical methods, the statistician should pick the simpler method to program. Programmers are not in a position to question if there are two valid statistical methods. However, with a fundamental understanding of statistics, the programmer can provide valuable feedback to the statistician. Alternately, a lack of common understanding can lead to unnecessary work. The goal of this paper is to enable programmers to ask intelligent questions and reduce the amount of work required to complete the statistical analysis.

DEFINITIONS

In order to make intelligent assumptions as to which procedures and methods should be used there are several key ideas that must be addressed. This paper will examine which procedures to use when there is one variable in question, with two samples or groups of data being discussed. For example, the two samples would be an active and placebo arm of a clinical trial. Multiple variables, multiple samples, or the combination of the two are beyond the scope of the paper.

The **Null Hypothesis**, which states the hypothesis put forth is false, is crucial to understanding how statistical trials are run. For example, the hypothesis is the drug in question will lower a person's blood pressure as opposed to a placebo. A statistician finds the probability that the hypothetical statement is false.

A **p-value**, probability value, is the percent chance that a given statement assumed to be false is in fact, true. A p-value represents a percentage. In order to be 95% confident that a hypothesis is true, a p-value must be less than 0.05.

Data with a normal distribution, when graphed, plots as a bell curve. The data has the same value for the mean and median. For each point on one side of the "middle" there is a corresponding point an equal distance from the mean on the other. As data points move from the middle, fewer data points exist. When there is a normal distribution, then assumptions can be made about the data, greatly simplifying the statistics involved in the analysis. There are few cases in clinical trials, due to limited sample sizes, that are normal distributions.

Variance, a measure of the normal distribution, describes how close individual values of the data are to the data's mean. While normal distribution quantifies how the data falls around the mean, variance is a determinant of how close those values are in relation to the mean.

When data is not normally distributed, the programming involved in the calculations is problematic. In answer to this problem, statisticians have developed the **central-limit theorem**, a method to treat non-uniform data as normally distributed for purposes of analysis. Despite the data's lack of normality, the central-limit theorem allows inferential statistics to be preformed as if it was a normal distribution. If the central-limit theorem fails, then different statistical methods must be used.

Assuming sufficiently normal data, comparing mean values, the next step is to determine if the two samples are **independent**. Independence implies that variables that affect one group have no bearing, positive or negative, on the other group. An example of independent variables is age and gender; age does not affect a person's gender.

Comparatively, dependency would be a person's height and weight. While height does not determine weight, it does affect the value.

Assume the following conditions:

- Mean of one variable in question
- Two samples or groups
- Samples are not independent
- Underlying distribution is normal or the central-limit theorem holds

These conditions imply that the appropriate statistical test to determine the significance of the difference between the means of the two treatment groups is a t-test. In its simplest form, SAS provides PROC TTEST for two samples. When the conditions fail, there are several other versions of the t-test built into SAS. If two samples are independent, an **f-test** uses the variances of the means to determine how to proceed based on the significant difference. If the two means' variances are shown to be significantly different, then use a t-test designed for equal variances. Otherwise, use a t-test designed for unequal variances.

In clinical trials it is not practical to sample enough of the population to have normal data. A representative sample is a smaller group of people that reflects the population at-large. A reason to perform descriptive statistics is to demonstrate that sample reflects the population at-large. Additionally, when the sample is subdivided into two groups for analysis, descriptive statistics illustrates that the two randomized treatment arms are like each other, and there is no bias of characteristics in one treatment arm versus the other.

Degrees of freedom quantify how close a set of data is to a normal distribution; the fewer degrees of freedom in the data, the fewer assumptions that need to be made about the data. A greater number of degrees of freedom imply more variability in the data, leading to less exacting results. A **confidence interval** is the range of values in which the expected mean of the entire population can be found. Fewer degrees of freedom allow for a smaller confidence interval.

For clinical trials, a typical confidence interval is 95%. A representative sample mirrors the characteristics of the population at large. For example, the average man's height in the population was 68.1 inches, with a 95% confidence interval of 66.7 – 69.5. There is a 95% chance that the true mean of the entire population falls inside of these two values.

APPLICATIONS

Due to design limits of clinical trials, it is not always possible to support that the central-limit theorem holds true, and the data must be tested further. A **binomial distribution** means the data has only two possible outcomes. For example, the outcomes could be yes or no, pass or fail, true or false. Also, a binomial distribution implies that there is a constant percentage of the sample that will achieve one or the other result.

People on an active blood pressure drug will have a diastolic blood pressure of less than 140. At the end of the study, the investigator will be able to say if this statement is true or false. For the remainder of this paper we will only consider binomial distributions.

The data independence is again tested. If the samples are not independent then use **McNemar's Test**. When programming using SAS, the McNemar's test is a special case of a large class of tests called **Cochran-Mantel-Haenszel Statistics**. Sample code to produce the test:

```
PROC FREQ;
    Tables tstgrp*bpresult / cmh1;
Run;
```

If the samples are independent then a closer look at the data itself is required. If all counts are not greater than or equal to 5 then use **Fisher's Exact Test**. In SAS, Fisher's Exact Test is also calculated by the frequency procedure:

```
PROC FREQ;
    Tables tstgrp*bpresult / exact;
Run;
```

Note that the default options of PROC FREQ produce a significant amount of output. An out= statement will allow formatting into a customized report.

If all the counts of the samples are expected to be greater than 5, then different determinations need to be made. A **confounding variable** is when another variable plays a part in determining the final value. With the example of height and weight, height does have an effect on weight, but a confounding variable would be a person's frame type.

A small frame, tall person could weigh less than a large frame person of average height.

If the expected results can be broken down into a 2*2 grid, such as treatment 1, treatment 2, by pass, fail, then one of three types of statistics are required. If there are no confounding variables present then either a two sample test for binomial proportions or a 2*2 contingency table is used:

```
PROC FREQ;  
    Table x / binomial;  
Run;
```

In most clinical trials there is at least one confounding variable. In that case, Cochran-Mantel-Haenszel Statistics are used, the SAS code being the same as that used for the McNemar's Test.

If, instead of a 2*2 table, a 2*many table is produced there are two choices based on the presence of confounding variables. With the presence of confounding variables, the Mantel Extension test is used:

```
PROC FREQ;  
    Table x / cmh;  
Run;
```

A chi-square test is used if there are no confounding variables or if there is no interest in the trend of the binomial proportions across the possible outcomes. Again, PROC FREQ is used to do the calculation:

```
PROC FREQ;  
    Tables tstgrp*bresult / chisq;  
Run;
```

Finally, if the possibilities include a many-by-many resulting grid, the chi-square test is chosen.

CONCLUSION

With a more comprehensive understanding of what they are programming, programmers are better positioned to support the work of the statistician. By better understanding why statisticians design statistical analysis in the way they do, programmers can ask more informed questions and gain heightened understanding of the statistical process. This will lead to greater programming efficiency. This will afford programmers a better feel for what to review in a SAP - which data needs to be better cleaned during the data management stage of a trial, and why complex statistics are required in certain cases. Programmers often question why a given statistical method is used during a clinical trial. For maximum efficiency, it is imperative that a programmer understand the methods chosen by the statistician prior to completing the programs of the statistical analysis plan.

ACKNOWLEDGMENTS

Nicole McBeth provided excellent editing and style suggestion, making this paper far more readable without her efforts.

CONTACT INFORMATION

Jim Edgington
Manager, Gilead Sciences, Inc.
333 Lakeside Drive
Foster City, CA 94404

650 522-6335
james.edgington@gilead.com

BIBLIOGRAPHY

Weiss, Neil and Hasestt, Matthew; [Introductory Statistics](#)

Rosner, Berbard; [Fundamentals of Biostatistics](#)

The SAS Institute; [SAS/STAT User's Guide Volume 1, ACECLUS-FREQ](#)

The SAS Institute; [SAS/STAT User's Guide Volume 2, GLM-VARCOMP](#)

