

Statistics: The Fourth Dimension of a “Statistical Programmer”

Gauri Khatu, Cytel, Pune, India
Vibhavari Inamdar, Cytel, Pune, India

ABSTRACT

In this world of clinical programmers, we have people coming from various backgrounds; some of them are engineers, biologist, few may be pharmacy or biotechnology graduates while only some are from statistics background. The one thing that is common to all is that they all are titled as “statistical” programmers. Few may have had a course or two in Statistics during their academic years, but some are completely new to this world of numbers.

So does it make any difference if you have a prior Statistics knowledge? Can you acquire the skills to look at numbers differently? This paper discusses why it is important for a programmer to be aware of at least the basics of statistics.

INTRODUCTION

In the physical world, two dimensions reflect a plane surface picture; the third dimension gives it a depth and fourth dimension represents time. Similarly the journey of a “statistical” programmer initiates with two dimensions- programming aptitude and communication skills. Domain knowledge adds depth to his/her experience. To race in time more efficiently, one needs to have a good understanding of statistics.

In clinical domain, statistics supports the understanding of the large volume of efficacy and safety data that represent the effects of the drug on humans. It plays a central role in the entire process of drug development from planning, conduct, interim analysis, final analysis and reporting. Now, we programmers come into picture during analysis and reporting phase. Statistical awareness can help us understand the data in better way and notice when data aren't as expected. The routine tasks of a programmer usually consist of working on datasets, tables and figures. Being efficient on these tasks can make one a “good” programmer. However what sets one apart from others is the vision to look beyond just the records. Assuming that the programmers may not be statisticians, hence obtaining in-depth statistical details for interpretations may not be necessary. Even then, knowing the fundamentals will serve the purpose of getting some insight of the broader picture.

The word “Statistic” may portray some complicated analysis, modeling and inferences. However, as commonly referred, statistics is not just mere number crunching, but a numerate common sense.

This paper discusses through few case studies which are selected from real life work examples, how some basic statistical understanding can change the approach in looking towards the numbers. However, the details of the statistical theory or the SAS® procedures used for the derivations are out of scope of this paper.

The examples covered are as follows:

1. Basic demographic summary
2. Summary of a laboratory parameter
3. Shift table
4. Summary from efficacy analysis
5. Survival graph

CASE STUDY 1

One of the basic table to be submitted in a clinical study report is the baseline demographic summary. It describes the characteristics of the population participating in that study. In clinical trials, the aim is to study the comparison between the study treatment with placebo or any other active comparator. The investigators are interested in observing the response to the given drug in the active treatment group as against the control group. To avoid bias in the results, the baseline characteristics across all such groups need to be comparable. For example - the age groups, gender and race proportions and such other demographic parameters should be approximately in similar range. Many disease areas and as well some treatments may have prognostic or treatment effects linked to certain baseline characteristics. The demographic summary is thus useful in verifying that a fair comparison is made across the different groups. Also at times, it is a protocol pre-requisite that subjects are to be enrolled in defined proportions (e.g., enrichment designs, stratified designs) like females should be double in number to that of male population. These points can be easily tested using the elementary demographic table.

Let us try to understand a similar table.

Table 1: Demographic Summary

Paramter	Descriptive Statistics	Placebo (n=115)	Treat A (n=112)	All (N=227)	
Age (years)	N	115	112	227	} Calculated Descriptive Statistics
	Mean (SD)	27.8 (3.59)	27.3 (4.18)	27.5 (3.89)	
	Median	28.2	27.1	28.8	
	Q1 - Q3	25 - 30	25 - 30	25 - 30	
	Min - Max	19 - 41	19 - 41	19 - 41	
Age category (N%)	19-27 years	54 (47.0)	59 (52.7)	113 (49.8)	} Calculated Frequency & Percentages
	28-32 years	49 (42.6)	41 (36.6)	90 (39.6)	
	33-50 years	12 (10.4)	12 (10.7)	24 (10.6)	
Gender	Female	16 (13.9)	12 (10.7)	28 (12.3)	
	Male	99 (86.1)	100 (89.3)	199 (87.7)	
Race (N,%)	Asian	0	0	0	
	Black	0	0	0	
	White	115 (100.0)	112 (100.0)	227 (100.0)	
	Other	0	0	0	
BMI at baseline	N	115	112	227	
	Mean (SD)	22.26 (2.594)	22.22 (2.377)	22.24 (2.484)	
	Median	21.90	21.88	21.88	
	Q1 - Q3	20 - 24	21 - 23	21 - 24	
	Min - Max	17 - 30	18 - 31	17 - 31	

Observe that descriptive statistics have been calculated for few parameters where as frequencies for the others. So, what are the factors which decide these calculations? This can be determined by the type of variable: Categorical or Continuous.

The variables like Age category, Gender, Race and Ethnicity can be grouped into distinct classes, like male-female, White-Black-Asian and so on. Finding averages and medians of this data will not yield any rational or useful information. Just the counts and percentages of subjects in these groups can give the idea about distribution of the data in the active and placebo groups.

As against this, for the variables like BMI, Age, Height and Weight - the descriptive statistics like Mean, Minimum and maximum are the 'representatives' of the data. Let us consider a part of the above table to understand the interpretation of each statistic.

PhUSE 2013

Table 2: Subpart of Table 1

Parameter	Descriptive Statistics	Placebo
		(n=115)
Age (years)	N	115
	Mean (SD)	27.8 (3.59)
	Median	28.2
	Q1 - Q3	25 - 30
	Min - Max	19 - 41

The table signifies that the average age of the 115 subjects belonging to Placebo group is approximately 28 (27.8) years with a deviation of 3.5 years.

- The same is evident from the median age of 28 years, with half of the subjects below the median and remaining above it.
- The Q1 and Q3 stand for the lower and upper quartiles of the data with 25% of the subjects lying below and above these values respectively, and the remaining 50% in between. So it is understood that 50% of the subjects belong to the 25-30 yrs age group.
- The youngest of the subjects is 19 and the eldest is 41 with the rest in between.

Thus the demographic table, although known to be a simple one and the SAS® procedures to obtain these numbers also very straightforward, nevertheless it gives vital information about the subjects involved in the trial. Hence a statistical programmer should not just restrict at the creation of deliverables, but develop an inclination to identify with the outputs.

CASE STUDY 2

The laboratory parameters are one of the crucial factors determining the safety profile of the drug, and at times could be an important efficacy endpoint. The observations on the subjects are recorded at baseline visits prior to dosing. Similarly, the postdose records are collected at pre-specified interval as mentioned in the protocol. The analysis of differences in pre and post drug administration results is used for the safety assessment. Now let us walk through a example of a standard laboratory summary.

Table 3

Summary of Observed Values of Laboratory Parameters by Treatment Group						
Safety Population						
Parameter: Calcium (mmol/L)						
Treatment	Visit	Descriptive Statistics				
		N	Mean	SD	Median	Min;Max
Placebo	Predose	2	2.220	0.0424	2.220	2.19; 2.25
	Day 1, 24 Hours	2	2.280	0.0283	2.280	2.26; 2.3
	Day 3, Predose	2	2.225	0.0778	2.225	2.17; 2.28
	Day 7, Predose	2	2.260	0.0566	2.260	2.22; 2.3
	Day 10, 24 Hours	2	2.360	0.0283	2.360	2.34; 2.38
Treatment A	Predose	6	2.228	0.0542	2.245	2.14; 2.29
	Day 1, 24 Hours	6	2.342	0.0488	2.335	2.3; 2.43
	Day 3, Predose	6	2.290	0.0525	2.290	2.22; 2.36
	Day 7, Predose	6	2.322	0.0397	2.330	2.25; 2.36
	Day 10, 24 Hours	6	14.165	28.8232	2.405	2.34; 73

This is a simple summary with the basic descriptive statistics on one of the laboratory parameter. At first glance, nothing seems wrong with the table. On observing it carefully, it can be noticed that the numbers from the last row (highlighted) do not seem to be in sync with the rest of the data. This is a signal that it needs further exploration. Studying the numbers in the table can give some indications.

The Min; Max column for all the visits except the last has values in the range of 2-2.50, the last visit has the max result as 73.

To cross check, going back to the data and searching for the range of column 'Result' for 'Calcium' test also suggests that 73 could be a possible 'outlier' as seen in Figure 1 below. A more confirmatory approach would be a box plot of the raw values (Figure 2).

Figure 1: Snapshot of the distinct result values

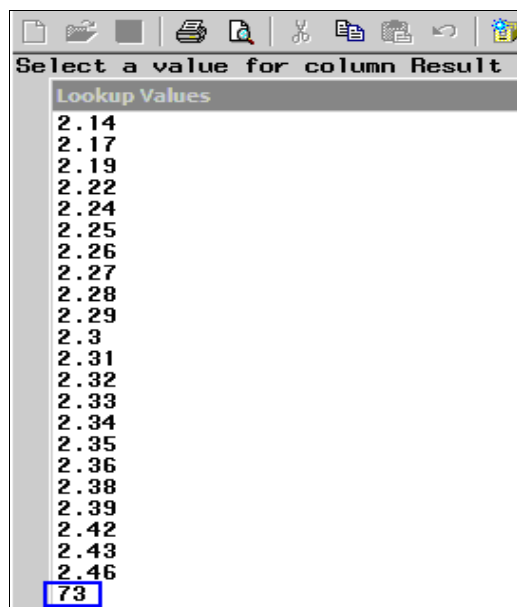
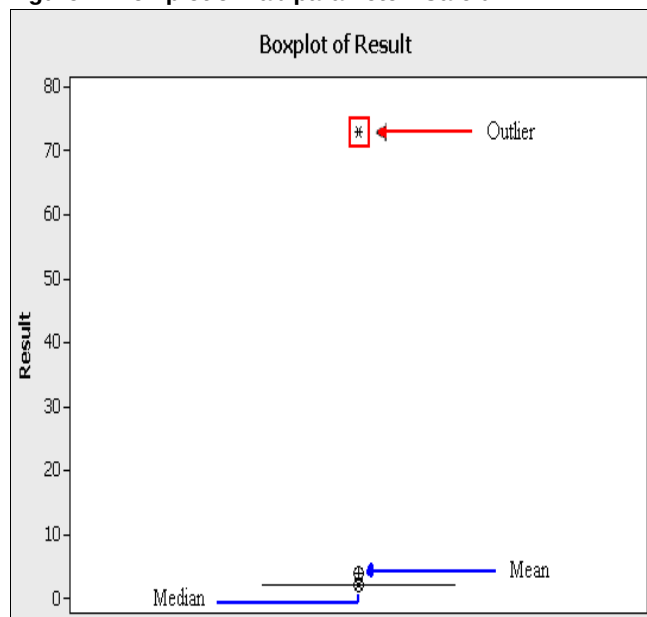


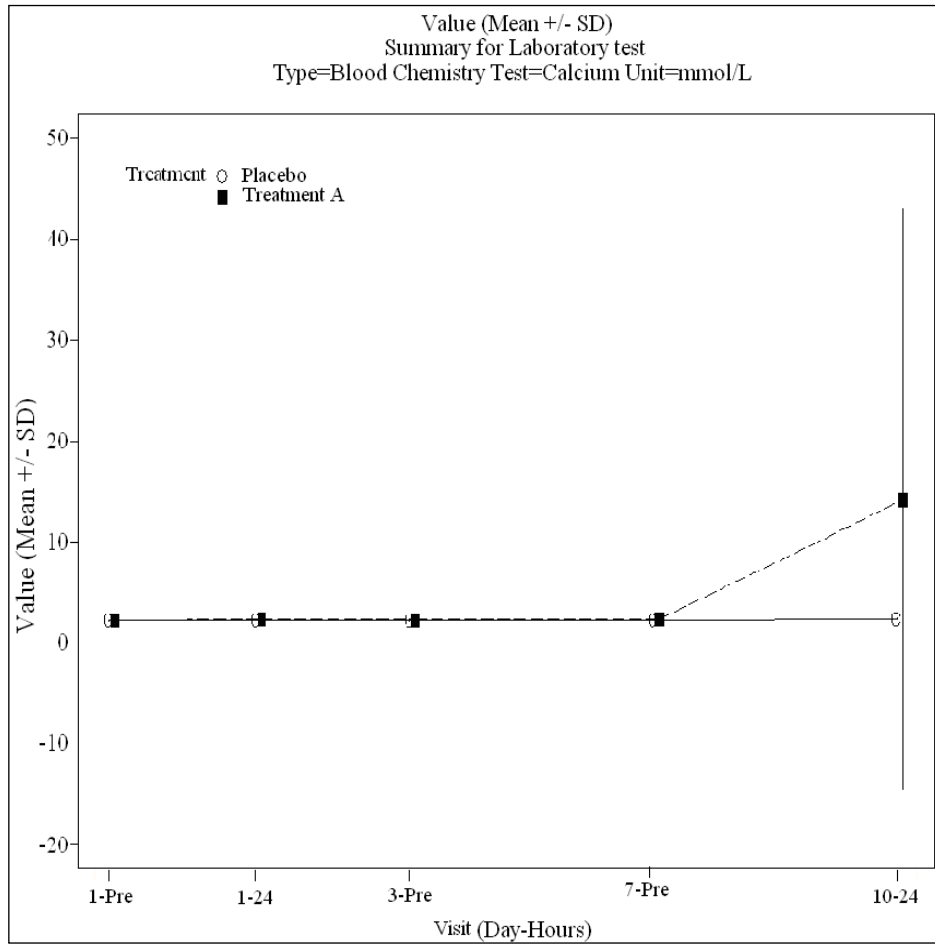
Figure 2: Box plot of Lab parameter: Calcium



The first image shows the distinct result values for the Calcium parameter and the second displays the box plot of the same. As seen above, both the figures indicate the extreme value for the result. This could be a data issue or a possible typographic error. It should be confirmed from Statistician or clarified through queries raised by data management to sites, and then decide about the appropriate analysis, possibly excluding the outlier.

The laboratory tables are usually very lengthy and 'observing' each and every value may not be feasible. So such outliers may not be easily caught, even during validation. This difficulty can be overcome by looking at the corresponding graphs vigilantly. Below is the graphical representation of the summary statistics presented in the table above.

Figure 3: Summary of Laboratory Parameter: Calcium



As seen from the graph, the result value of 73 blows up the mean as well as the standard deviation at the last time point. This abnormality in the data is easily captured through this figure.

The below figures show the box plot and profiles of the means of the same data after excluding the extreme record.

Figure 4: Boxplot for Lab Parameter excluding the outlier

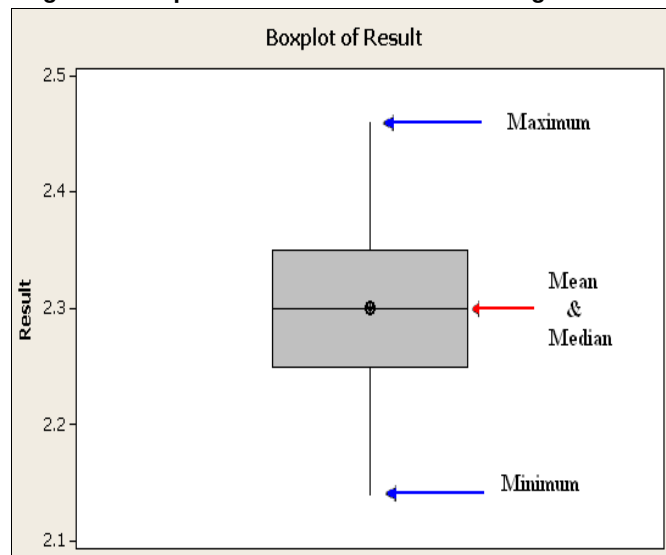
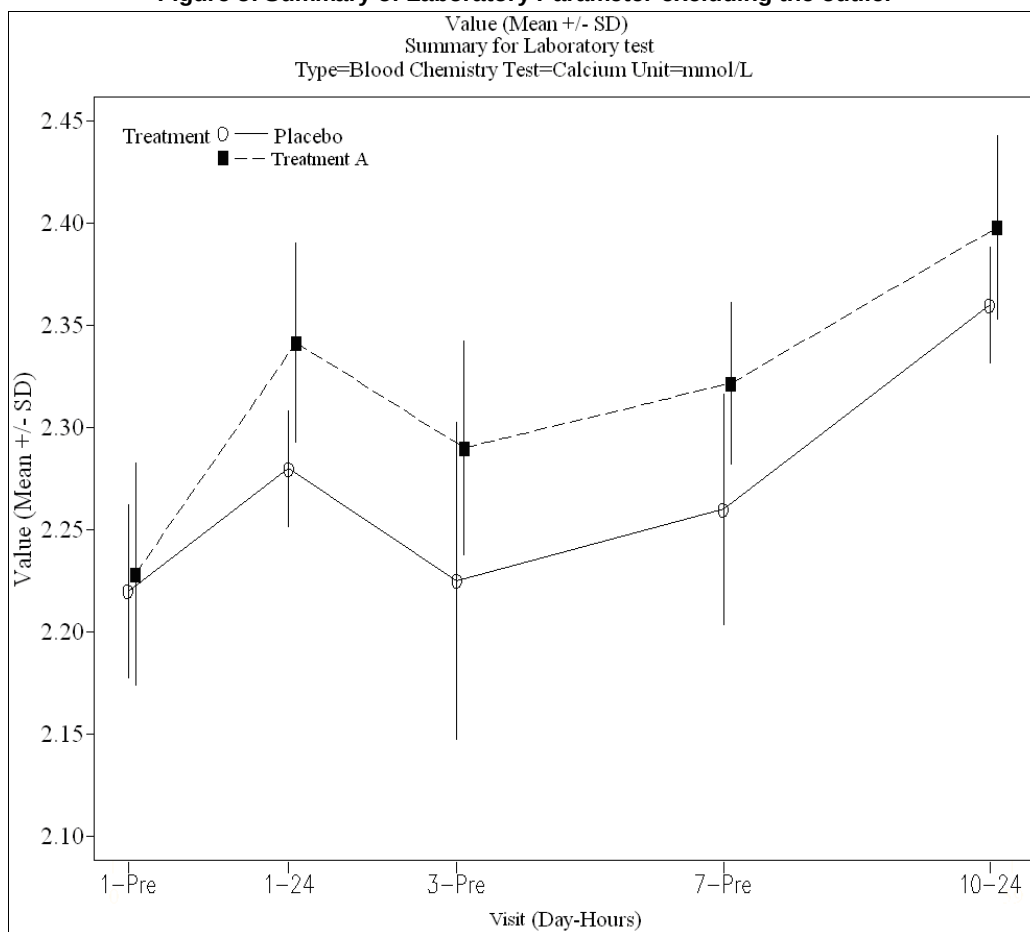


Figure 5: Summary of Laboratory Parameter excluding the outlier

Thus the outputs that are produced during the analysis are interrelated and one can try to figure out the relationship between the data, table, listings and figures to get better picture.

CASE STUDY 3

Let us now go to another analysis from a Laboratory Shift Table.

The main purpose of shift tables is to determine how the result varies from one timepoint to another. Shift tables are used to understand treatment-emergent patterns of shifts in grades of categories from baseline vs post-baseline (either at specific visits, or worst post-baseline). The lab values are classified into various grades. These grades are then summarized based on the change of categories between baseline and some subsequent time point

In the given example the aim is to observe the change in the grades from baseline to post baseline in two different treatments. Consider a mock shell for the shift table to be produced for CTC grades at baseline to postdose value for Fasting Glucose.

PhUSE 2013

Table 4: Mock shell for Laboratory Shift Table

Table for Laboratory Parameter							
Fasting glucose shift table - Safety analysis set							
Treatment	Baseline		Post Baseline				
	n	(%)	Grade 1 n (%)	Grade 2 n (%)	Grade 3 n (%)	Grade 4 n (%)	
TRTA (N=xxx)	Grade 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 3	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 4	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Total	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
TRTB (N=xxx)	Grade 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 3	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 4	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Total	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Overall (N=xxx)	Grade 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 3	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Grade 4	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Total	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)

Glucose unit is mg/dL.
Baseline is defined as the last non-missing value prior to the first study drug.

The things to be taken into account before starting with the coding:

- The lab data has multiple postdose visits.
- As per the given mock shell, it is required to have a unique record per subject per treatment for the baseline and the post-baseline. So there should be a defined logic to derive the unique post-baseline grade per subject – example the worst grade or best grade. So where to find this information?
 - Ideally, this information should be provided in the footnote of the mock shell.
 - If not, to locate this one can go back to the SAP and study the requirements.
 - If this is not mentioned explicitly in the any of the documents as well, the programmer needs to confirm it with the statistician.
- The above mock shell does not accommodate for any missing grade at either of the timepoints. However if the dataset has such a scenario, "missing" category should be added in the table layout. So it is also required to have a look at the details of the data for deriving the table.

Below figure shows the actual table derived referring the above layout.

PhUSE 2013

Table 5

Table for Laboratory Parameter							
Fasting Glucose (mg/dL) Shift Table - Safety Population							
Treatment	Baseline		Post-Baseline				
		n (%)	Missing n (%)	Grade 1 n (%)	Grade 2 n (%)	Grade 3 n (%)	Grade 4 n (%)
2							
TRT A (N=19)	Missing	2 (5.9)	2 (50.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	1 Grade 1	13 (38.2)	0 (0.0)	3 (18.8)	5 (83.3)	4 (66.7)	1 (100.0)
	Grade 2	2 (5.9)	0 (0.0)	1 (6.3)	0 (0.0)	1 (16.7)	0 (0.0)
	Grade 3	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Grade 4	2 (5.9)	0 (0.0)	1 (6.3)	0 (0.0)	1 (16.7)	0 (0.0)
	Total	19 (55.9)	2 (50.0)	5 (31.3)	5 (83.3)	6 (100.0)	1 (100.0)
TRT B (N=15)	Missing	2 (5.9)	1 (25.0)	1 (6.3)	0 (0.0)	0 (0.0)	0 (0.0)
	3 Grade 1	12 (35.3)	1 (25.0)	9 (56.3)	1 (16.7)	0 (0.0)	0 (0.0)
	Grade 2	1 (2.9)	0 (0.0)	1 (6.3)	0 (0.0)	0 (0.0)	0 (0.0)
	Grade 3	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Grade 4	5 0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Total	14 (44.1)	2 (50.0)	11 (68.8)	1 (16.7)	0 (0.0)	0 (0.0)
OVERALL (N=34)	Missing	4 (11.8)	3 (75.0)	1 (6.3)	0 (0.0)	0 (0.0)	0 (0.0)
	Grade 1	25 (73.5)	1 (25.0)	12 (75.0)	6 (100.0)	4 (66.7)	1 (100.0)
	Grade 2	3 (8.8)	0 (0.0)	2 (12.5)	0 (0.0)	1 (16.7)	0 (0.0)
	Grade 3	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Grade 4	4 2 (5.9)	0 (0.0)	1 (6.3)	0 (0.0)	1 (16.7)	0 (0.0)
	Total	34 (100.0)	4 (100.0)	16 (100.0)	6 (100.0)	6 (100.0)	1 (100.0)

Baseline is defined as the last non-missing value prior to first dose.
Percentage is based on the Baseline total n

Refer the numbering of the highlighted portions in the above table and read the corresponding explanation to see what these numbers signify.

1. Out of 19 subjects under Trt A, 13 subjects have grade 1 at baseline and three of them remain at grade1, five subjects shift to grade 2, four to grade 3 and one subject to the 4th grade post baseline.

So it should be noted that the total number of subjects under post baseline grades should add up to number of subjects at baseline grade category in this row.

2. Total 19 should match the Total number of subjects under the respective treatment (N). The vertical total as well as horizontal total should get add up to this number.

3. As seen in Trt B, the total of 14 subjects at baseline does not seem to be consistent with the corresponding column total as well as the treatment total (N = 15). Focusing on the grade 1 category, it can be spotted that the counts from the post-baseline records do not sum up to 12. As a result the corresponding values from the Total row give the incorrect addition of 14, which in reality should have been 15. Consequently this has affected the counts in the 'Overall' group as well.

4. The second footnote of the table directs towards the denominator to be used for percentages. In this table, the percentages are calculated over the total of each grade at post-baseline. However the correct approach would be to use the totals at baseline value (highlighted as **bold (point #5)**) for each of the corresponding three treatment groups. Thus the respective percentages would be 100% for these three totals.

Many time the mock shell do not include the "Total" row which adds up the total of all grade counts. In such cases it is programmer's responsibility job to ensure that the total comes up to 100 before finalizing the table.

Thus the shift tables, although look very simple with just some frequencies and percentages to be derived, need to be understood thoroughly and cross-checked very cautiously.

CASE STUDY 4

Consider a table on an efficacy endpoint from a neurology study. The study 123(Ext) is an extension of the previously conducted study (Study 123). The subjects enrolled in the extension arm are the same as that of Study 123. The following table reports the basic descriptive summary and the statistics for inferring the comparison between the medians of the Total T2 lesion volume in the two studies.

Table 6

Total T2 Lesion Volume at Visit 54 by Treatment Efficacy Population Subjects randomized to Trt A in Study 123 and to Trt B in Study 123 Extension			
Endpoint	Statistics	Study 123 Trt A (N = 225)	Study 123 (Ext) Trt B (N = 225)
Total T2 Lesion Volume	N(Missing)	216(9)	207(18)
	Mean +/- SD	12087.6 +/- 14230.90	11009.8 +/- 12690.05
	Median	7819.0	6492.0
	Min;Max	423;122157	210;98404
Inferential Statistics	Relative Reduction (1) % (Study 123 (Ext) vs. Study 123)		16.97
	97.5% CI (2)		(-186.00,796.00)
	p-value(3)		0.048
(1) Calculated as the ratio of the difference in median Total T2 Lesion Volume (study 123- study 123(Ext)) relative to the Total T2 Lesion Volume in the study 123. (2) CI is for the median change from Study 123 to Study 123(Ext). (3) p-value based on two-sided Wilcoxon signed rank test.			

Let us try to understand the table in details. The initial parts of this paper have covered the basic statistics at length. So this section will focus on the more complex part of the table

Table 7: Subpart of table 6

Endpoint	Statistics	Study 123 Trt A (N = 225)	Study 123 (Ext) Trt B (N = 225)
Total T2 Lesion Volume	Median	7819.0	6492.0
Inferential Statistics	Relative Reduction (1) % (Study 123 (Ext) vs. Study 123)		16.97
	97.5% CI (2)		(-186.00,796.00)
	p-value(3)		0.048
(1) Calculated as the ratio of the difference in median Total T2 Lesion Volume (study 123- study 123(Ext)) relative to the Total T2 Lesion Volume in the study 123.			
(2) CI is for the median change from Study 123 to Study 123(Ext).			
(3) p-value based on two-sided Wilcoxon signed rank test.			

. From the table and its footnotes, it can be observed that

- Relative reduction is derived from the medians of the two treatment groups (Refer footnote (1)). Thus the Lesion volume has reduced in the extension study by more than 16% in comparison to the core study. As this is a comparative analysis with the same subjects over a period of time, one can out of curiosity, check if this reduction is a desired change in favor of the treatment!

- The CI does not include the difference between the median lesion volumes ($7819-6492 = 1327$). This may give a hint that the difference is significant. This is justified by the p-value which is less than 0.05; hence difference between the medians is significant at 5%.

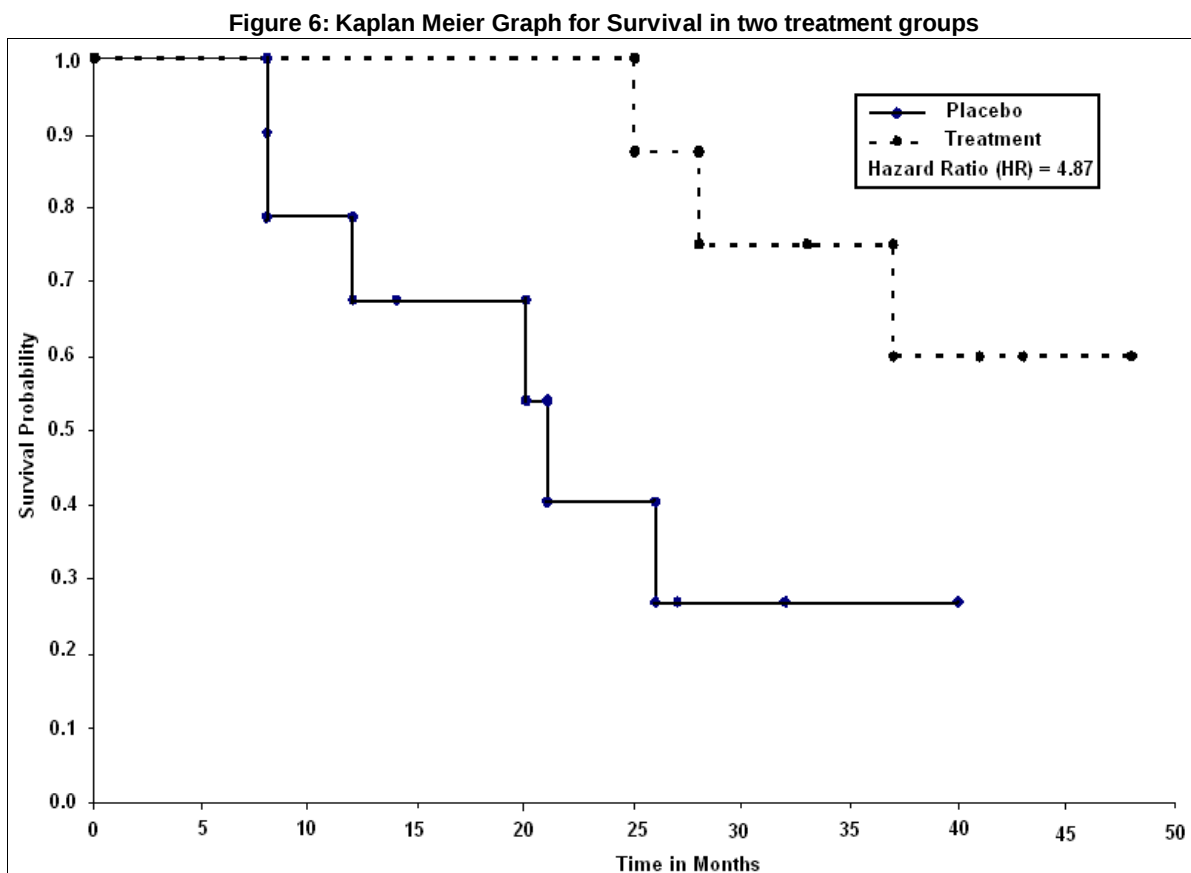
Confidence intervals and p-values are critical in establishing the significance of the comparison of the endpoints between the treatment groups. Thus these are very important pieces in supporting understanding of treatment effect. These are mandatory for the regulatory submissions. The method to calculate these CIs and p-values is pre-specified in the protocol. The procedures to get to these numbers may be readily available in the SAP or input documents. The output may also be derived by passing the correct variables from the dataset, but the challenge lies in taking that one extra step to recognize the meanings of the numbers.

CASE STUDY 5

Many clinical trials involve follow-up of the patients to observe the events such as death, relapse, adverse drug reaction or deterioration of disease condition. The follow-up time for the study may range from few weeks to many years. Specialized techniques like survival analysis are developed to handle such kind of data.

Kaplan-Meier analysis is one the tool in survival analysis which helps in estimating the probability of the survival of the subjects up to a certain point of time.

Consider the following Kaplan Meier Survival Probability graph for two treatment groups. Here the event of interest is death.



Looking at the graph the points to be noted are:

- The time on X-axis where the graph takes a downward step marks the occurrence of the event; death in this case. For example: The first death in the treatment group occurs around 25th month, while the same for the placebo group happens at around the 8th month.

PhUSE 2013

- The filled circles symbolize censored subjects i.e. these subjects cannot be evaluated any further for the event of interest. For example – Death cannot be recorded for subjects withdrawing from the study or lost to follow-up.
- Thus the death incidence happens at a faster rate in the placebo group as compared to the treatment. Hence the possibility of survival in the latter is more as compared to that of the former.

Now, let us see if the other results presented in the figure are consistent with the above three points. The Hazard ratio (HR) in a simple language is the ratio of the risk of experiencing the event of interest in the treatment population relative versus the control. It is already seen that the probability of event (i.e. death) in the placebo group is greater. Thus the HR should have been less than 1 due to the larger denominator, which is not consistent with the above result of $HR = 4.87$

So what could be the reason behind this mismatch? The possible answer to this is how the censoring variable has been defined and its value that has been entered in the SAS® procedure to calculate the HR. Suppose the censoring flag is defined as 1 when the record is censored and 0 when the death is observed. So while specifying in PHREG procedure which calculates the HR, the value 1 should be entered and not 0. This can be a quick check performed by the programmer before getting this as a QC finding from the validator.

DISCUSSION

As seen from above case studies, the scope of a programmer's role is not restricted to just coding. It is crucial for a programmer to identify key data issues, inconsistencies in the input documents at the right time. Ideally this should take place during the dry runs. The discrepancies in the data eventually affect the quality of the data and the date of the data base lock (DBL). This in turn impacts the overall study timelines and consequently the budget of the trial.

From this paper we have also seen that it is imperative that programmers take initiative to understand “why” particular analysis is performed. The SAP or protocol may not give this information readily. However a statistician can help in answering the question. Programmers need to be proactive in raising early discussion points with statisticians or clinical teams while coding. However the stringent timelines, complexity and the volume of the deliverables will not always give programmers the luxury of time to go into minute details. But whenever there is downtime, one can go back and get their questions answered. By using every minute of free time, a programmer can investigate, research and learn many aspects of the study. These extra efforts will definitely make one a strong and valuable programmer and help contribute better.

CONCLUSION

Statistics as a subject is a vast domain and it is difficult to have a complete expertise. We have tried to touch base on some basic topics in this paper to initiate a thought process to look at the numbers diligently and differently. One can be a successful programmer purely on the basis of the SAS expertise without any reference to Statistics. However, a SAS programmer with some statistical knowledge will be in a position to take decisions while programming and hence have a competitive edge over a pure SAS programmer. Thus he/she would be more independent and be able to do better justice to the role by understanding and developing more meaningful outputs with higher quality.

ACKNOWLEDGMENTS

We would like to thank all our colleagues at Cytel who shared their experiences, patiently answered all our queries, read through our drafts and gave us constructive feedback.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Gauri Khata: gauri.khatu@cytel.com

Vibhavari Inamdar: vibhavari.inamdar@cytel.com

Cytel Statistical Software & Services Pvt. Ltd., Pune, India

Brand and product names are trademarks of their respective companies.