PhUSE 2014

Paper IS04

Unuttered Questions of Statistical Programmers

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

ABSTRACT

We know that all the bright folks from various backgrounds, be it medical, engineering, pharmacy, biotechnology or statistics, enter our world of clinical programming. Together they are called 'statistical programmers'. Data issues to complicated coding, ambiguous specifications to stringent timelines, nothing can beat their enthusiasm. The only thing they hesitate to do is to ask a particular question – "why?" Whenever they are asked to provide standard error instead of standard deviation; whenever they are suggested to use different formulae each time they calculate p-value; whenever they are asked to plot mean and confidence interval in the errorbar graph instead of usual mean +/- SE plot; they follow the instructions immediately and deliver the output in a quick and efficient manner. Though "What to do" and "How to do" have become common questions now; "Why to do" is still something very rarely asked. This paper tries to provide an approach to find some quick and easy answers to those unuttered questions from statistical programmers keeping above examples in focus e.g. how to approach p-value or where exactly standard error is different from standard deviation.

INTRODUCTION

Usually the life of a statistical programmer starts with reading instructions from protocol or statistical analysis plan and ends with delivering TFLs. The communication between the study statistician and study programmer concentrates on mostly "What to get" and "How to get it". A programmer asking "Why to get that" is a rare event. Though programmers may always have this question on their mind, they somehow hesitate to ask it. So, the only option which lies open to them is searching for statistical concepts on the internet which most of the time leads to complicated mathematical expressions. And it becomes a bit difficult to relate those formulas to the actual data in clinical trials.

For example, if we want to approach p-value from a programmer's perspective, his/her scope will be limited to the formula to get it. At the most we might understand what to do if the derived p-value is less than 0.05 and what happens if it is more than 0.05. But why is it necessary to use a particular procedure, what is the meaning of the p-value in the context of the data under study and most importantly, what decision is based on the p-value is never discussed with statistical programmers. Maybe it is beyond scope for statisticians to do that. So instead, this paper is an attempt to give statistical programmers a direction towards how to get some of these answers.

This paper lists out some such basic questions that may come into a programmer's mind but are not answered directly through study documents like SAP. Along with answering them, it also tries to describe how to approach these questions. Additionally, this paper explains some basic terminologies used in clinical trials in a practical way instead of going into a lot of theoretical detail. All discussions are based on real life examples.

CASE STUDY 1: SD AND SE

It does not have to be a complicated analysis to start the list of the unuttered questions. The example below is a table for Systolic Blood Pressure (SBP) – a vital signs parameter which is a very common parameter used for safety analysis. The table presents the summary statistics of SBP for different dose groups and visits.

Table 1
Summary of SBP (mmHg) by Visit (Partly)
Safety Analysis Set
(Phase 1 Study to Evaluate the Effect of Study Drug on Adult Subjects with Advanced Solid Tumors)

	0.1 mg/kg (N = 3)	0.5 mg/kg (N = 3)	Total (N = 6)
Baseline			
n	3	3	6
Mean	132.3	133.3	132.8
SD	32.1	10.7	22.5
SE	18.6	6.2	11.9
Median	144.0	139.0	140.5
Min, Max	96, 157	121, 140	96, 157
Day 1			
n	3	3	6
Mean	128.7	134.7	131.7
SD	21.4	9.6	14.2
SE	12.3	5.5	8.6
Median	140.0	133.0	134.5
Min, Max	104, 142	126, 145	104, 145
Change form Baseline from Day 1			
n	3	3	6
Mean	-3.7	1.3	-1.2
SD	11.5	13.6	12.6
SE	6.6	7.9	7.2
Median	-4.0	6.0	1.8
Min, Max	-15, 8	-14, 12	-15, 12

Even though all the summary statistics are pretty common and widely used; there can be a slight confusion between two of them – SD and SE. these are abbreviations of Standard Deviation and Standard Error respectively. However knowing that does not explain their purpose. The definitions and formulae of these two summary statistics are –

- In statistics and probability theory, the **standard deviation** measures the amount of variation or dispersion from the average. [1]

  Formula: $\sqrt{\frac{1}{N}(\sum_{i=1}^{N}(x_i - m)^2}$, where N is the sample size, m is the sample mean and X_1, X_2…X_N are the data points.

- The **standard error** is the standard deviation of the sampling distribution of a statistic. The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate. [2]

  Formula: $\sqrt{\frac{1}{N(N-1)}(\sum_{i=1}^{N}(x_i - m)^2}$, where N, m and X_1, X_2…X_N are same as above.

But the definitions, again, are not very clear about the purpose of having both of these in the table or about the difference between the two.

To go in depth of the interpretation of these two numbers in the table, it is necessary to understand the concept of population and sample first. The set of subjects included in a clinical trial is considered as a set of representatives from the target population in the world who can be treated using the investigational product. For example, in above table, the sample size of dose group 0.1 mg/kg is three. That means the summary statistics derived in the first column of the table are based on 3 subjects getting 0.1 mg/kg of study drug. But the purpose of this clinical trial is to extrapolate the interpretation of the results from these 3 subjects to millions of similar patients in the world. So the sample mean at day 1 (128.7 mmHg) in the table does not interpret as average SBP of these 3 subjects only. It serves as an estimate of average SBP in the entire population of adult people having advanced solid tumor if they would receive 0.1 mg/kg of the study drug one day prior.

A similar statement cannot be made about Standard Deviation though. As per above definition and formula, the standard deviation would measure the amount of variation from the average. This means that the SD of SBP of the three subjects in 0.1 mg/kg dose group will only measure how scattered these three SBP values are from their average i.e. 132.3 mmHg. Unlike sample mean, the standard deviation does not provide any good estimate of the population value. So it is not used in above table.

Rather, if the formula is slightly changed, it becomes a better estimate of population standard deviation. The tweak is changing the divisor and the new formula becomes –

$\sqrt{\frac{1}{\mathbf{N-1}}(\sum_{i=1}^{N}(x_i - m)^2}$, where N is the sample size, m is the sample mean and X_1, X_2…X_N are the data points.

Though both the standard deviations, with divisor N and (N - 1) are biased estimators of population standard deviation; the later has less bias [3] than the former one [4] and it is named “Sample standard deviation” and hence presented in summary tables. So when SAS® calculates the SD using standard PROCs for summary statistics, it is the sample standard deviation which is presented.

So, the SD of SBP in the dose group at day 1 (21.4 mmHg) is the sample standard deviation. And it is an estimate of the standard deviation of SBP for the entire target population of adult people having advanced solid tumor if they would receive 0.1 mg/kg of the study drug.

The story of the Standard Error is a different one. Unlike other summaries in the table, SE does not estimate any population value directly. The full name of SE in the table is “Standard Error of Mean” or in short SEM and it is based on the probability distribution of the sample mean.

In reality there is only one sample considered in a clinical trial; but theoretically there can be many. So there can be many sample means like the one in above table. But, as all the samples are to be drawn from a particular population, which is the adult patients with advanced solid tumor here, all the sample means of SBP would be around the population mean of patient SBP. Individually all of them will be an estimate of the population mean and together they will form a probability distribution, with keeping the population mean at center, as mean value. If the variation of the sample means could be measured in terms of standard deviation, it would be Standard Error of Means. But, as it cannot be measured, for the population mean is unknown and only one sample is drawn instead of many, Standard Error of Means or SEM is rather estimated using the available data of a single sample. Which means, the difference between the average SBP of all the adult patients with advanced solid tumor if they would receive 0.1 mg/kg of study drug one day prior and 128.7 mmHg (sample mean SBP calculated in the table for same dose at visit day 1) is estimated as 12.3 mmHg.

So, the standard deviation provided in the tables is – “Sample Standard Deviation” and provides an estimate of population standard deviation. And the standard error in the table is – “Estimate of Standard Error of Mean” or SEM which gives an estimate of how far the sample mean can be from the population mean. Also, they will have less bias for larger sample size. For the estimate of standard error of mean, with n = 2 the underestimate is about 25%, but for n = 6 the underestimate is only 5%.

CASE STUDY 2: ERRORBAR PLOT

A popular way to observe the average trend of any data is to have error-bar plot. For example, considering the example in the previous case study, observing the average trend of SBP over visits would be easier to analyze graphically than as a table.

There are two types of error bar plots which are popular in clinical trials. One is for Mean +/- SE and another one is Mean with its confidence intervals.

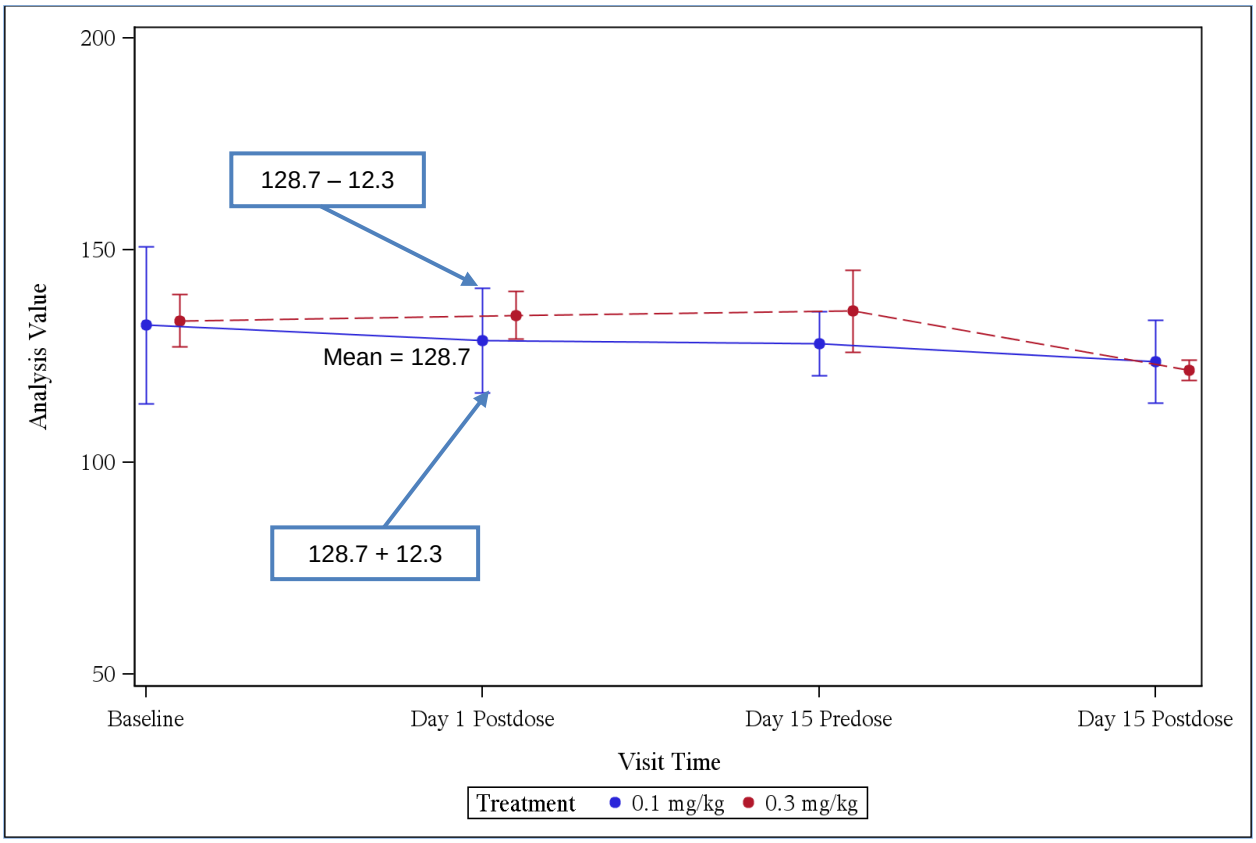
But the question is, which one is better? Or is there even a comparison?

This is one of those situations where the answers are obtained by asking more questions – for example, why are the error bars needed? As mentioned in the previous case study, all the results calculated in clinical trial analysis are not absolute values. They are just estimates and might vary from the actual population value. So it is necessary to estimate the possible span of that variation as well to get a proper interpretation. A couple of examples are discussed below.

Standard Error provides an estimate about how much a sample mean can vary from the population mean. The values (Mean – SE) and (Mean + SE) provide an estimate of a possible interval for the population mean around the sample mean. Thus it gives a more robust estimate than a single value of sample mean. The concept of providing an interval instead of a single value is more commonly known as interval estimation.

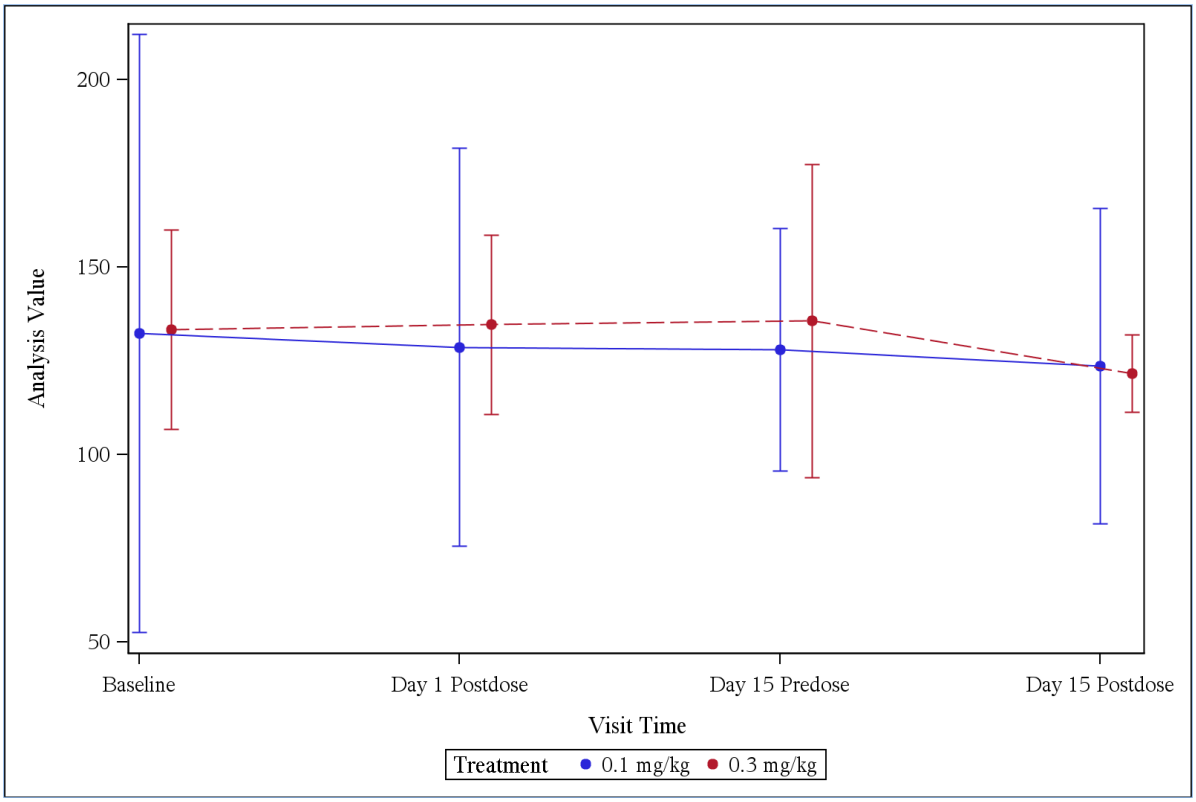
Figures 1 and 2 are graphical presentations of the same data tabulated in Table 1. In Figure 1, the estimated possible ranges of average population SBP are displayed for different dose levels and visits along with the average sample SBP value.

Figure 1
Mean +/- SE Plot for Systolic Blood Pressure (mmHg)
Safety Analysis Set



The confidence interval adds value to above mentioned interval. The interval from (Mean – SE) and (Mean + SE) provides an estimate of possible range of population mean but cannot measure the probability of the population mean falling in that range. Confidence interval, on the other hand, estimates that information. In Figure 2, the error bars are the estimated interval that would contain the population mean with 95% confidence.

Figure 2
Mean with 95% CI Plot for Systolic Blood Pressure (mmHg)
Safety Analysis Set



From the definitions, it might seem that confidence interval of sample mean is better than a crude range of Mean +/- SE. But calculating confidence intervals of means has its own limitations. The usual and easily available formula to get confidence intervals of mean is based on the assumption that the data in question follows normal distribution (discussed further in case study 3). The common procedures in SAS® also have the same underlying assumption. There are other ways as well to get confidence intervals without any prior assumption of the probability distribution of the data; but they are not as easy to get as the above.

Also, to assure a certain probability that the population mean would be included, the range becomes very wide as can be observed in above figure. In this particular case, sometimes the interval has covered almost the total possible range of SBP (Baseline 95% CI for 0.1 mg/kg). This happens mainly for small sample size. In case confidence intervals are needed for drawing conclusion in a clinical trial, sample size should be carefully chosen based on the trial design.

So, confidence interval plot may be better than having Mean +/- SE plot only if the data follows normal distribution and the data has decent sample size of each analysis group. Otherwise, Mean +/- SE plot would give much robust estimate.

CASE STUDY 3: LOG TRANSFORM DATA PRIOR TO ANALYSIS

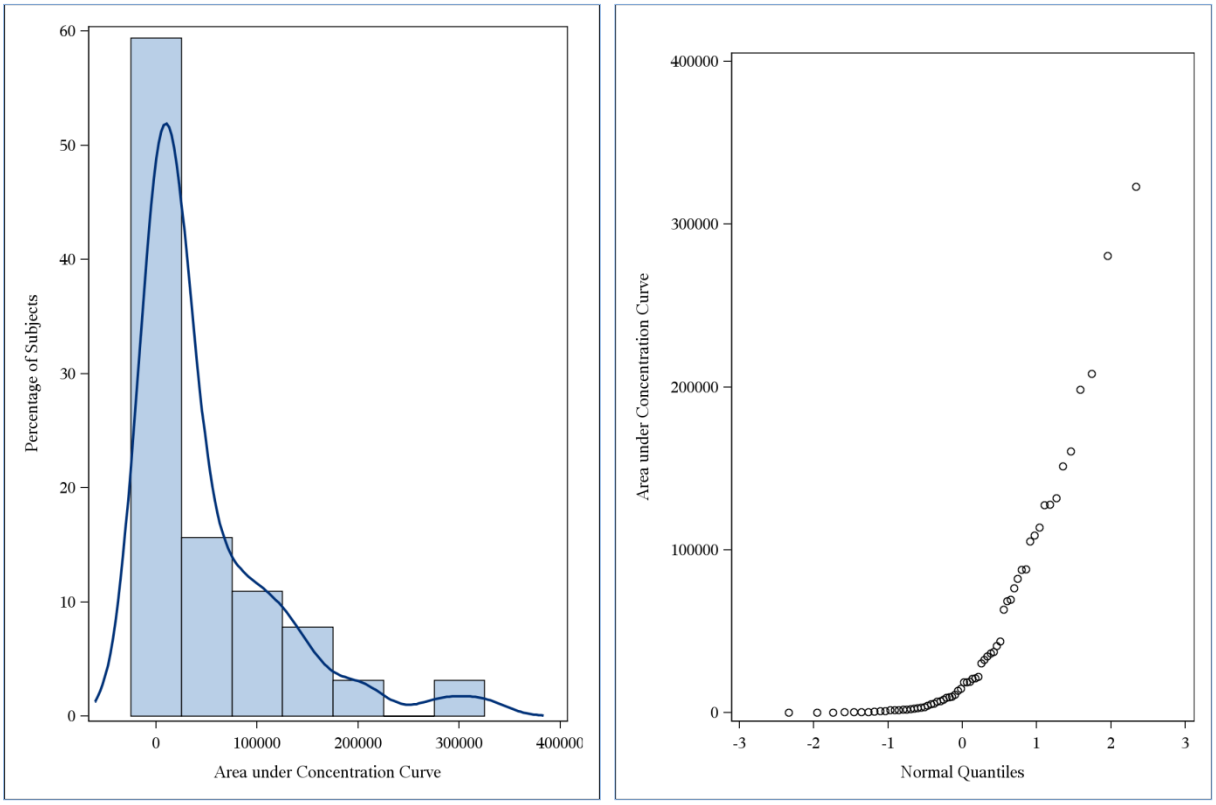
For some of the parameters used in clinical trials, instead of absolute values the logarithm of the values is used for analysis. For example, pharmacokinetic parameters and some laboratory test values.

Like mentioned in case study 2 above, many analyses, formulae and models used in clinical trials are based on normality assumption; for example T test, Analysis of variance/covariance, mixed model, correlation, regression etc. This means that the data needs to follow normal distribution to apply these analysis tools. There are a variety of tests to check the normality of data but the most common finding is that clinical data is not usually normally distributed.

There are a few ways to overcome this problem. One of them is transforming the data to follow a normal curve so that standard statistical analysis methods can be applied. There are various transformation methods as will like log transformation, square root, log (1+x), Box-Cox power transformation etc. A good example is of non-normal data is area under curve of drug concentration over time. It does not follow normal distribution but if we take log transformation, it fits to normality.

There are various tests available to check the normality of the transformed data. A famous one is Q-Q plot (Quantile - Quantile plot). Figure 3 shows a histogram (left) and corresponding Q-Q plot (right) for area under curve of drug concentration over time.

Figure 3
Histogram and Q-Q Plot for Area under Concentration Curve

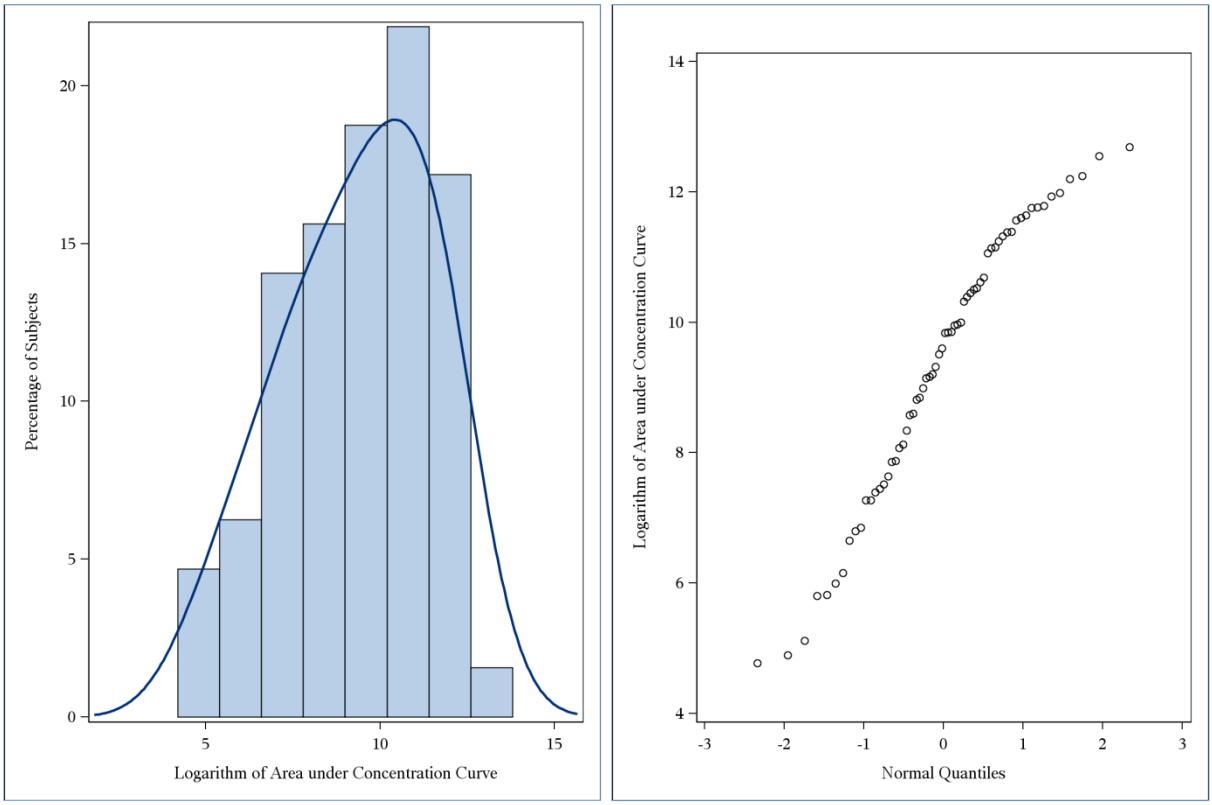


The Histogram has actual area under concentration curve values plotted on horizontal axis and percentage of subjects in the data having corresponding area under concentration are plotted on vertical axis. If the data follows normal, the histogram curve should be like a symmetric bell curve. And the corresponding Q-Q plot should look like a diagonal straight line.

It is evident from the histogram that the values are not normally distributed because they are not forming a symmetric bell curve. Also this deduction is supported by the Q-Q plot which is not close to a straight line.

Figure 4 displays same plots for logarithm of area under concentration curve.

Figure 4
Histogram and Q-Q Plot for log-transformed Area under Concentration Curve



The histogram curve for log-transformed values is approximately a symmetric bell shape. And the Q-Q plot is also close to a straight line. This indicates that normal distribution can now be assumed for these log-transformed values. Of course the analysis results from log-transformed values need to be back-transformed to get them related with the actual variable.

Though there are certain variables which cannot be transformed to normal distribution at all, for example time to event variables. For these types of data, analysis without normality assumption is usually chosen.

CASE STUDY 4: LINEAR MODEL RANDOM EFFECT

Often we need to run a linear model to carry out analysis of pharmacokinetic parameters and sometimes some safety and pharmacodynamic parameters as well. Below is a PK parameter table for example –

Table 2
Summary Statistics and Statistical Comparisons for the Plasma Pharmacokinetic Parameters after Single Dose Administration of the Study Drug or the Co-administration of Two Other Marketed Drugs

	Study Drug			Co-administration of Two Other Marketed Drugs			Study Drug / Co-administration	
Pharmacokinetic Parameter	N	GM	90% CI	N	GM	90% CI	GMR	90% CI
$AUC_{0-\infty}$ ‡ (nM.hr)	22	8027	(7767, 8297)	20	7931	(7675, 8196)	1.01	(1.00, 1.03)
C_{max} ‡ (nM)	21	895	(849, 945)	20	867	(822, 914)	1.03	(0.99, 1.08)
‡ Back-transformed least squares mean and confidence interval from linear mixed effects model with ***treatment and study period included as fixed effect*** and ***subject included as random effect***; performed on natural log-transformed values;								
GMR = Geometric least squares mean ratio, GM = Geometric Least-Squares Mean, CI = Confidence Interval								

Here we will focus on the footnote since that gives the programmer a hint as to how the SAS code should be set up. The SAS® code used to get above table is –

```
proc mixed data = pkdata;
  by pkparm; /*Run the model for each PK parameter*/
  class treatment period subject;
  model logval = treatment period/ddfm = KR;
  random subject;
run;
```

The analysis varies from ANOVA, ANCOVA or Mixed Model based on the requirement; the variables in the model are changed based on the scenario. The only consistent part is to use subject as random effect which leads to several questions:

- Why is subject being considered as random every time?
- Why isn't any other variable in the model considered random?

This is because the effect associated with a sampling procedure is considered as a random effect to the model. This will be more prominent if it is compared with any other factor, such as treatment. The variable "treatment" above has two values in the data – "Study Drug" and "Co-administration of two other marketed drugs". The situation will not be different for any other sample or even for the population. But, the effect of the subject on the pharmacokinetic parameters is not fixed; it will change as soon as a different sample is taken. So it is considered as a random effect coming from a random sample.

CASE STUDY 5: P-VALUE

One of the most seemingly mysterious but important number which is derived in clinical trial data analysis is p-value. P-value is needed almost everywhere as far as any statistical testing is concerned. There are so many questions that come with this single term so it would be convenient to divide that into some parts based on different types of questions that can arise.

WHAT IS P-VALUE?

From previous case study it has been very clear that the statistical analysis of clinical trials is focused around population and sample size. The aim is to get information about the target population using the available sample data. And that directly leads to the purpose of a clinical trial. The purpose can vary based on the sponsors' interests, like –

- Is the study drug effective?
- Does the study drug behave like any other marketed drug in human body?
- Is the study drug better than any other drug involved in the clinical trial?

  And so on

These are some examples of usual questions faced in clinical trials. The method to get answers to these particular questions is by "Testing of Hypothesis". This method carries out a test to check the validity of a statement, called 'Hypothesis' in statistical terms, about a population based on sample(s) taken. Below is an example –

Consider a city with population of around 20 million where an anti-smoking campaign is being conducted. The campaign claims that the proportion of smokers in the city has gone down to less than 12% of the total population which was much higher 1 year ago. To confirm the validity of this claim, hypothesis testing can be done.

Testing of hypothesis should necessarily have two hypotheses statements at the very least. First is the null hypothesis, which is the "no change" statement, against the alternative hypothesis, "the change" statement. And most importantly, the alternative hypothesis should always support the purpose of the trial.

So for this example,

Null Hypothesis – the percentage of smokers is greater than or equal to 12%, vs.
Alternative Hypothesis – the percentage of smokers is less than 12% (the claim of the campaign)

The most accurate method would be to count the total population of the city as well as the number of smokers. But that would involve enormous time and money. So the testing would be based on samples. There could be one or

more sample, with fixed or various sizes. After getting the sample data, the overall sample proportion of smokers would be calculated to estimate proportion of smokers in the overall population, exactly like the analysis done in clinical trials.

Now the million dollar question is – how to decide about the population even when the sample percentage is available? For example, if the sample percentage of smokers is 99%, it would not support the claim; or if the sample percentage is estimated as 1%, it would support the claim. And both of this can be simple guess work if the sample proportions are such extreme values. But the decision will not be that easy if the sample percentage comes somewhere close to the crucial proportion of 12% for example 14% or 8%. So to come to any conclusion about the population based on sample data there should be some pre-defined rule or boundary.

So, if the sample proportion comes as 14%, it can be because the actual proportion of smokers in the city is above 12% and it is reflected in the sample. Or the sample can just happen to have 13% smokers whereas the percentage of smokers in the city is less than 12%. Here comes p-value in picture. P-value is the probability of getting a particular sample value if the null hypothesis were true. And the null hypothesis will be rejected if p-value comes below a particular level. In most of the cases the level is defined as 0.05. P-value provides an estimate of the validity of the sample and it does not give any direct probability of any statement about population.

WHAT IS THE RANGE OF P-VALUE?

As mentioned above, p-value is a probability value. So it should be strictly between 0 and 1, both inclusive.

WHY REJECT NULL HYPOTHESIS IF P-VALUE COMES BELOW 0.05?

Lower value of p-value indicates that it is a very less likely to get these sample values if the null hypothesis were true. For above example – suppose the sample proportion value comes as 1%. Then the corresponding p-value will be – probability of sample proportion coming as 1% given that null hypothesis is true. Which is the probability of getting sample proportion 1% given that the actual proportion of smokers in the city is greater than or equal to 12%. Certainly the probability of this event taking place would be very low, almost close to zero.

Which means the smaller is the p-value, the less likely to get the sample proportion if the null hypothesis were true. So, for lower p-value it is reasonable to reject the null hypothesis instead of concluding that a rare event has occurred by chance.

WHY IS 0.05 A MAGIC NUMBER?

It's about defining a boundary. The truth about the population remains unknown forever. So to decide based on probability would require a strict boundary and beyond that the null hypothesis would be rejected. The boundary can be considered as 0.01 if we want to make the test stricter. This boundary in probability is the level of significance for the test.

Also above example of building hypothesis was not taken from clinical trial just to make the example simple enough to understand the concept of p-value. In clinical trials, the null hypotheses are something like –

Null Hypothesis:	The average improvement in BP for the study drug is equal to the same for placebo
Alternative Hypothesis:	The average improvement in BP for the study drug is greater or same as for placebo
Null Hypothesis:	The ratio between the study treatment and control treatment in terms of maximum drug concentration in body plasma is greater than/equal to 1.25 or less than/equal to 0.80
Alternative Hypothesis:	The ratio between the study treatment and control treatment in terms of maximum drug concentration in body plasma is between 0.80 and 1.25

WHAT IS THE FORMULA TO GET P-VALUE?

Unfortunately, there is no universal formula to get p-value. P-value holds different significance for different tests. So the formula to get p-value changes as well.

To consider the first hypothesis example above – it is testing the equality of mean between two groups; one group is change in blood pressure after getting study drug and another group is change in blood pressure after getting placebo. This analysis can be carried out using linear model as blood pressure closely follows normal distribution.

For the second test, the pharmacokinetic parameter T_{max} needs to be compared for study drug and another marketed drug. Now T_{max} does not follow normality and there isn't any possible transformation of T_{max} that follows normal

distribution as it is a time-to-event data. So the test must be carried out without any normality assumption – using non-parametric tests.

P-value can be obtained from both the cases and they will have similar interpretation towards corresponding null hypotheses. But, as the test methods are totally different in these two cases, so will be the formula to get p-value.

At this point, it is important to mention that there are some tests for which we would not want to draw any conclusion based on p-value even when we can calculate them. This mainly happens when the null hypothesis of a statistical test becomes somewhat complicated so that the p-value, the definition of which is fully dependent on the null hypothesis, does not remain that straight forward to interpret.

WHY DO WE DISPLAY SMALL P-VALUES LIKE "<0.0001"?

If a p-value is way less than the significance level (which is 0.05 or 0.01 in most of the cases), then it really does not matter what the exact value is. It is sufficient to know that the p-value is less than the significance level or not so that a decision can be taken whether or not to reject the null hypothesis.

CAN WE DECIDE THAT THE NULL HYPOTHESIS IS TRUE IF WE GET A HIGH P-VALUE?

No. We can only reject or not reject the null hypothesis. If the p-value comes greater than significance level, we can only say that we cannot reject the null hypothesis based on the given sample. Small p-values indicate very strong evidence against the null hypothesis, because it indicates that random variation was not responsible for the observed value of the sample. Large p-values indicate no evidence against the null hypothesis.

SUMMARY

It is true that gaining expertise in Statistics and that too while concentrating mainly on programming is extremely difficult. This paper is an attempt to answer some basic questions that arise with early phase clinical trial analyses. Though the examples covered here are certainly not sufficient to understand the vast range of clinical trial analyses, hopefully they can serve a starting point. And the main point here is to note that statistics is not about crazy formulae but a set of tools which help us learn more about the data we collect. And as seen here, statistics is not very difficult to understand if we ask the right questions.

REFERENCES

[1] http://en.wikipedia.org/wiki/Standard_deviation
[2] http://www.investopedia.com/terms/s/standard-error.asp
[3] http://en.wikipedia.org/wiki/Bias_of_an_estimator
[4] https://www.lhup.edu/~dsimanek/scenario/errorman/distrib.htm

ACKNOWLEDGEMENT

I would like to thank Meghana Bhagwat and my other colleagues in Cytel who shared their experiences, answered queries, read through my drafts and gave constructive feedback.

CONTACT INFORMATION

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

aparajita.dey@cytel.com

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