

SOME BASIC STATISTICS FOR NON-STATISTICIANS

Gary Stevens, Biogen Idec Ltd, Maidenhead, UK

ABSTRACT

In our industry as clinical or statistical programmers we are at the beck and call of statisticians for all sorts of weird and wonderful numbers to describe the clinical trial data. This paper will present an explanation of a few of the common statistics we are asked to produce and in turn hopefully provide a greater understanding of the true meaning of these statistical terms in relation to our clinical trial analyses. The intended audience is statistical or clinical programmers who would like to expand their knowledge into understanding the statistics they are actually producing.

INTRODUCTION

When I began my programming career in SAS many, many years ago everything was going fine. Not long after I began I moved into the pharmaceutical industry as a clinical programmer and all of a sudden statistics appeared. This paper is an attempt to help understand everything that “frightened the life out of me” all those years ago. I will begin by defining statistics, discussing normal and skewed population curves and their effect on positioning of the mean and the median. I will then move onto the standard deviation, standard error, confidence intervals, p-values and finally correlations.

WHAT IS STATISTICS?

DEFINITION: Statistics is the study of populations based upon samples taken from these populations.

Statisticians use a method of comparison. They want to know the effect of treatment on a response. To find out they compare responses of a treatment group with a control group. If a control group is comparable to the treatment group, apart from the treatment, then a difference in the responses of the two groups is likely to be due to the effect of the treatment. Statisticians use statistics to report the outcomes of clinical trials.

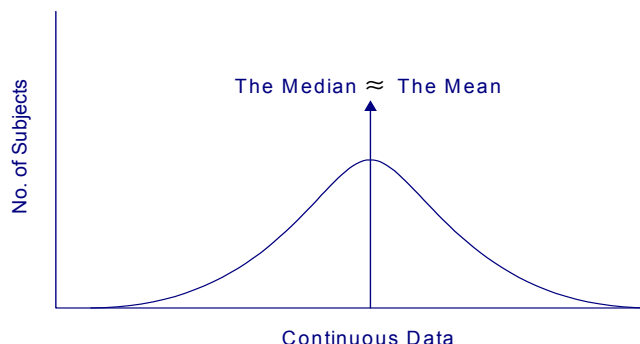
MEAN OR MEDIAN?

DEFINITION OF MEAN AND MEDIAN:

The Mean (or Average) of a list of numbers equals their sum, divided by how many there are.

The Median is the middle value when all the values are lined up lowest value to highest value (or vice-versa).

The Median is best shown on a histogram. The Median of data we have (shown on the histogram below) is the value with half the area under the curve to the left and half the area under the curve to the right.



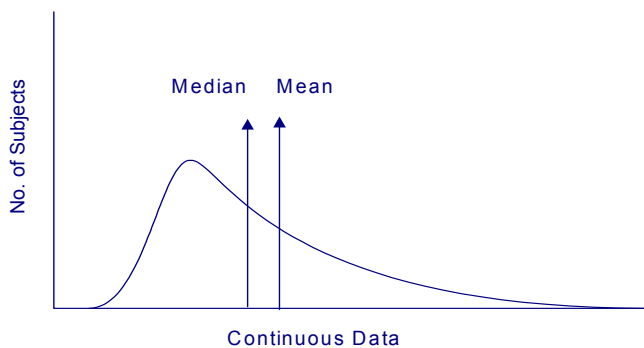
NORMAL OR GAUSSIAN DISTRIBUTION:

There are strong theoretical reasons for continuous data to be normally distributed. The type of bell shaped graph above shows a typical normal distribution. This is called a Gaussian distribution due to the fact a German Statistician called Carl Friedrich Gauss in the 18th Century described it. In a Normal distribution such as this the Mean and the Median are very close together.

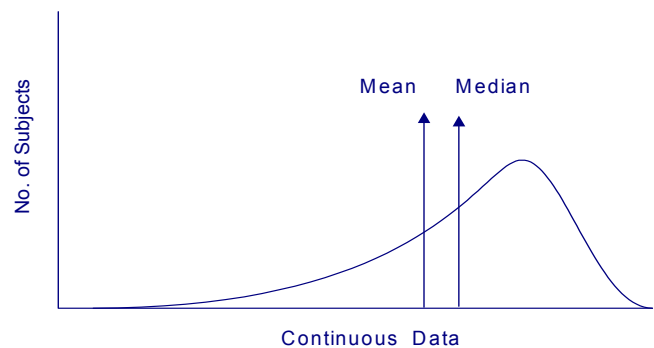
Examples of continuous data are heights of subjects, weights of subjects or blood alcohol levels. Essentially data that can vary numerically from subject to subject.

SKEWED DISTRIBUTION:

Sometimes the distribution is not normal. The distribution can be skewed. There are three main categories when looking at normal distribution. Symmetrical as we have discussed and the other two are shown below.



Long Right Hand Tailed



Long Left Hand Tailed

THE EFFECT OF THE DISTRIBUTION ON THE POSITIONING OF THESE TWO STATISTICS:

As you can see from the above graphs;

Long Right Hand Tailed – Mean is larger than the Median
Symmetric – Mean is approximately the same as the Median.
Long Left Hand Tailed – Mean is smaller than the Median

STANDARD DEVIATION

DEFINITION OF STANDARD DEVIATION:

The Standard Deviation (SD) says how far away numbers on a list are from the Mean.

CALCULATION/DERIVATION:

The Standard Deviation is the root mean square (r.m.s.) of the deviation from the Mean. To calculate by hand the SD follow the 4 steps below

1. Calculate the Mean
2. Calculate the difference between each individual value and the Mean
3. Take the mean of the squared differences using $n-1$. (This is known as the variance)
4. Take the square root of this and you get the Standard Deviation.

Hence the formulae for the SD is where d is the difference from the average

$$SD = \sqrt{\frac{d1^2 + d2^2 + \dots + dn^2}{n-1}}$$

In Example suppose we have the numbers 20 10 15 15

$$\text{The mean} = \frac{20+10+15+15}{4} = 15$$

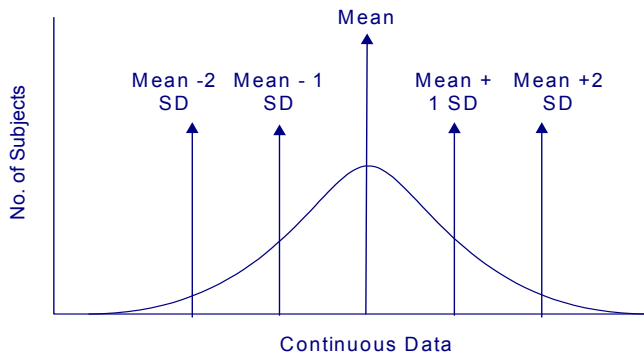
The difference from the mean for each of our numbers is 5 -5 0 and 0

$$\text{Therefore the SD} = \sqrt{\frac{5^2 + (-5)^2 + 0^2 + 0^2}{4-1}} = 4.08$$

WHAT A STANDARD DEVIATION ACTUALLY MEANS:

For a normal distribution approximately 68% of the numbers on a list are within one SD of the mean. The other 32% are further away. Approximately 95% are within 2 SDs of the mean the other 5% are further away. Therefore we can say that the smaller the SD the closer to the Mean our numbers are.

GRAPHICAL REPRESENTATION:



STANDARD ERROR

DEFINITION OF STANDARD ERROR:

The Standard Error can be applied to any statistic. In our Example I will use the Mean.

The Standard Error of the Mean (SE) measures the mean-to-mean variation unlike the SD, which measures the subject-to-subject variation.

CALCULATION/DERIVATION:

The formulae for calculating the SE of the mean value is as follows:

$$\frac{SD}{\sqrt{n}}$$

Where n is the number of subjects in the sample.

WHAT A STANDARD ERROR ACTUALLY MEANS:

Suppose we take a sample of our Subjects and measure their Blood Alcohol levels. We could calculate the mean of these values. If we take another sample of these subjects we will also get a mean but it is likely that the mean will not be the same. Depending on the variation in data you will get a number of different means from all the samples that you take.

PhUSE 2006

In practice we do not normally have the luxury of this repeat sampling but we can estimate the repeated Mean values based on just a single sample. This estimate of the SD given by the above formulae is the SE of the Mean.

The SE of the mean is a measurement of how reliable our sampling group is with the group as a whole. Therefore a small SE means that we have a reliable sampling group where the mean values from all sample groups will be close together.

A large SE means we are in an unreliable sample setting where the Mean values could be very different.

CONFIDENCE INTERVALS

DEFINITION:

The Confidence Interval expresses a range of values in which we are fairly certain that a value lies. The Confidence Interval has two parts. The Upper Confidence Interval and the Lower Confidence Interval.

CALCULATION:

In this example we will calculate the Confidence values for a single Mean. More often than not in Clinical Trials we are asked to present 95% Confidence intervals:

Lower 95% Confidence Interval

$$Mean - \left(1.96 \times \frac{SD}{\sqrt{n}} \right)$$

Upper 95% Confidence Interval

$$Mean + \left(1.96 \times \frac{SD}{\sqrt{n}} \right)$$

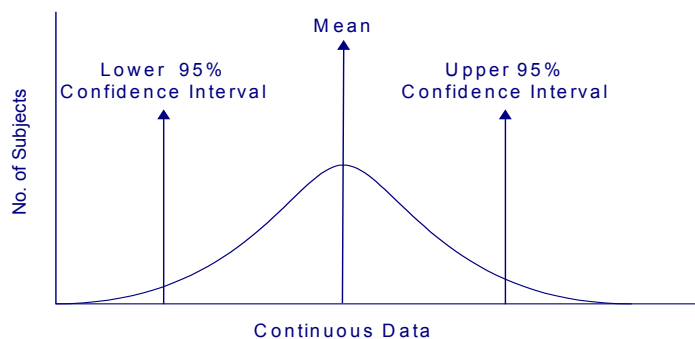
The value of 1.96 is the multiplying factor used in calculating 95% Confidence Intervals for normally distributed data. Other Confidence Intervals can be calculated using different factors for a normal distribution.

- 1.645 for 90% Confidence Intervals
- 1.960 for 95% Confidence Intervals
- 2.576 for 99% Confidence Intervals

WHAT THE CONFIDENCE INTERVALS ACTUALLY MEAN:

We can be 95% certain that the Mean of our sample lies within these two values.

GRAPHICAL EXPLANATION:



P-VALUES

DEFINITION OF A P-VALUE:

The p-value of a test is the chance of getting a test statistic of that magnitude or larger – assuming the null hypothesis to be true. The p-value is not the chance of the null hypothesis being right but a measure of evidence against the null hypothesis.

NULL HYPOTHESIS – ALTERNATIVE HYPOTHESIS:

To give a more of an understanding of a p-value first we have to introduce the idea of the Null and Alternative Hypothesis. The Null Hypothesis expresses the idea that an observed difference is due to chance. The Alternative Hypothesis expresses the idea that the observed difference is real. The simplest scenario is the comparison of two Mean values Mean_1 & Mean_2.

First we have to formulate the Hypotheses.

Always structure the Alternative Hypothesis to be the desirable outcome.

Therefore

The Null Hypothesis states that the two Means are equal

The Alternative Hypothesis states that the two Means are not equal

Suppose our SAS® Code gives us a p-value of 0.037

What does this mean?

Definition A. There is a 3.7% probability that Mean_1 = Mean_2

Definition B. There is a 3.7% probability that the observed difference between Mean_1 & Mean_2 is due to chance.

Definition B is the correct Definition.

The p-value of a test is the chance of getting a test-statistic of that magnitude or larger assuming the null hypothesis to be true. Small p-values are evidence against the null hypothesis; they indicate something else besides chance was operating to make the difference.

The smaller the p-value, the more evidence we have against the null hypothesis.

We are trying to disprove the null hypothesis.

As an analogy: If you are driving down a country lane and you see a field of sheep and they are all white you can state a hypothesis that all sheep are white. Further on down the lane you see more fields of sheep and these are also all white. These other fields of sheep support your hypothesis that all sheep are white but they cannot confirm this. It is, however, very easy to disprove your hypothesis if eventually you come across just one single black sheep.

WHAT THE P-VALUE ACTUALLY MEANS:

The p-value is a probability and is therefore expressed as a number between 0 and 1. A value of between 0 and 0.05 is declared statistical significant and we conclude, in the above case, that the means are different. If the p-value is greater than 0.05 then we declare non-significance and conclude no differences.

CORRELATIONS

DEFINITION OF CORRELATION:

If there is a strong association between two variables, then knowing one helps in predicting the other. But when there is a weak association, information about one variable does not help in guessing the other.

The correlation coefficient is a measure of linear association, or clustering about a line. The relationship between two variables can be summarized by;

1. The mean of the x-values, the SD of the x-values.
2. The mean of the y-values, the SD of the y-values.
3. The correlation coefficient.

CORRELATION COEFFICIENT CALCULATIONS:

PhUSE 2006

Convert each variable to standard units. The Mean of the products gives us the correlation coefficient. (A value is converted to standard units by seeing how many SDs it is above or below the Mean)

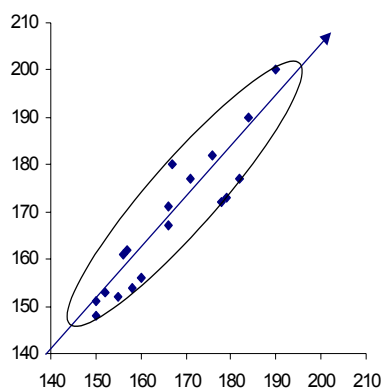
GRAPHICAL EXPLANATION OF CORRELATION COEFFICIENT:

Correlations are always between -1 and 1 , and can take any value in between. A positive correlation means that the spread of scattered points (sometimes known as the cloud) slopes up; as one variable increases so does the other. A negative correlation means that the cloud slopes down: as one variable increases the other decreases.

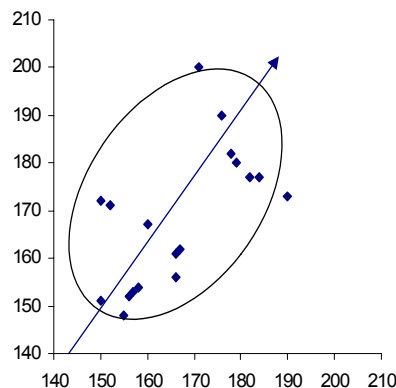
The following graphs are of two arbitrary data points plotted against each other. The first two scatter plots have positive correlation, as the line of best fit shows an increase in the x-axis data point increases the y-axis data point. The cloud shown for the first plot shows a good clustering of points and we could say that the correlation coefficient of this data is quite high and not too far away from 1 . The second is a little less convincing in that the clustering is fairly widespread and therefore the correlation is lower.

Positive Correlations

High Positive Correlation

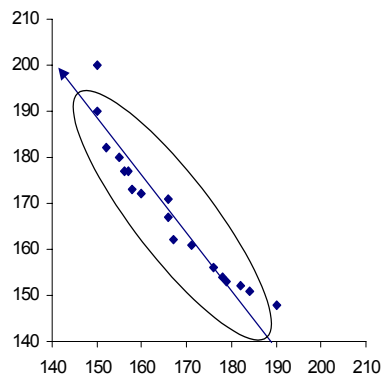


Lower Positive Correlation

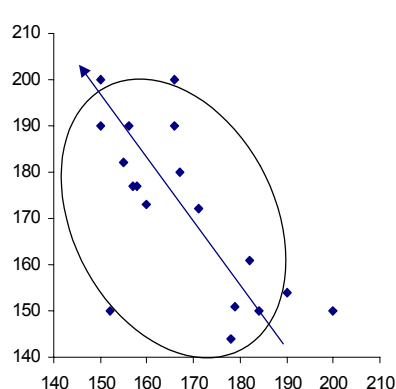


The next two graphs are similar although they show negative correlation.

High Negative Correlation



Lower Negative Correlation



REFERENCES

Statistics – Third Edition, Freedman, Pisani & Purves

CONTACT INFORMATION

Gary Stevens
Biogen Idec Ltd.
Thames House
Foundation Park
Maidenhead
Berkshire SL6 3UD
UK

Tel. +44 (0) 1628 501000
Direct +44 (0) 1628 512583
Fax +44 (0) 1628 501010

gary.stevens@biogenidec.com