Chemistry Statistics

Lee Kennedy

null

Table of Contents

# Introduction

1. There is nothing here to scare you.
2. Statistics are only a way of describing the properties of data.

## Assumptions

1. The data is normally distributed.
2. Never assume. Check.

Normality can be assessed with:

* Histograms (a density curve is also useful)
* Boxplots
* Normal probability plots (Q-Q)
* Skewness
* Kurtosis

# Part A : Basic Statistical Methods and Ideas

## Data

Most collections of data follow a Gaussian distribution. Commonly known as a ‘Bell’ curve or a ‘Normal’ curve.

Some data, notably microbiological data, does not follow the Gaussian distribution. In order to make it do so, you must transform the data by techniques such using logarithms.

### Descriptive Statistics

Descriptive statistics permit you to get an indication of a dataset’s ‘normality’. You can do a summary in base R or use "describe" in the Psych package. A quick check for normality may be just comparing the mean and median. They should be relatively close.

The following two data sets are displayed using the two processes. A shows good normality but B doesn't. This can be seen in the ranges and the difference between the mean and the median. Note also what happens to the means when the top and bottom 10% of data points are removed ("trimmed").

summary(data.in)

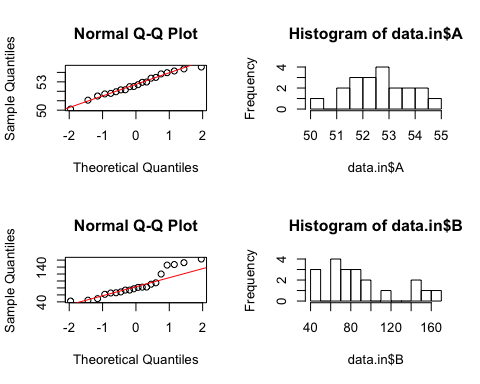
## A B   
## Min. :50.12 Min. : 42.35   
## 1st Qu.:51.91 1st Qu.: 65.98   
## Median :52.60 Median : 79.56   
## Mean :52.69 Mean : 89.20   
## 3rd Qu.:53.56 3rd Qu.:101.02   
## Max. :54.57 Max. :163.10

describe(data.in)

## vars n mean sd median trimmed mad min max range skew  
## A 1 20 52.69 1.18 52.61 52.73 1.25 50.12 54.57 4.45 -0.20  
## B 2 20 89.20 36.86 79.56 86.32 21.64 42.35 163.10 120.75 0.74  
## kurtosis se  
## A -0.79 0.26  
## B -0.80 8.24

### Q-Q Plot

A Q-Q plot gives an indication of normality. The previous data is assessed in the following plots. If normally distributed, the data will lie near the red line. The first set is normal, the second is not. The histograms provide a more visual perspective.



### Skewness

Skewness quantifies how symmetrical the distribution is.

* A symmetrical distribution has a skewness of zero.
* An asymmetrical distribution with a long tail to the right (more higher values) has a positive skew.
* An asymmetrical distribution with a long tail to the left (more lower values) has a negative skew.

### Kurtosis

Kurtosis characterizes the relative peakedness or flatness of a distribution compared with the normal distribution.

* Different to Excel, which sets a normal curve to a kurtosis of 3.
* A result less than 0 indicates a curve flatter than a normal distribution.
* A result greater than 0 indicates a curve sharper than a normal distribution.

### Transformations

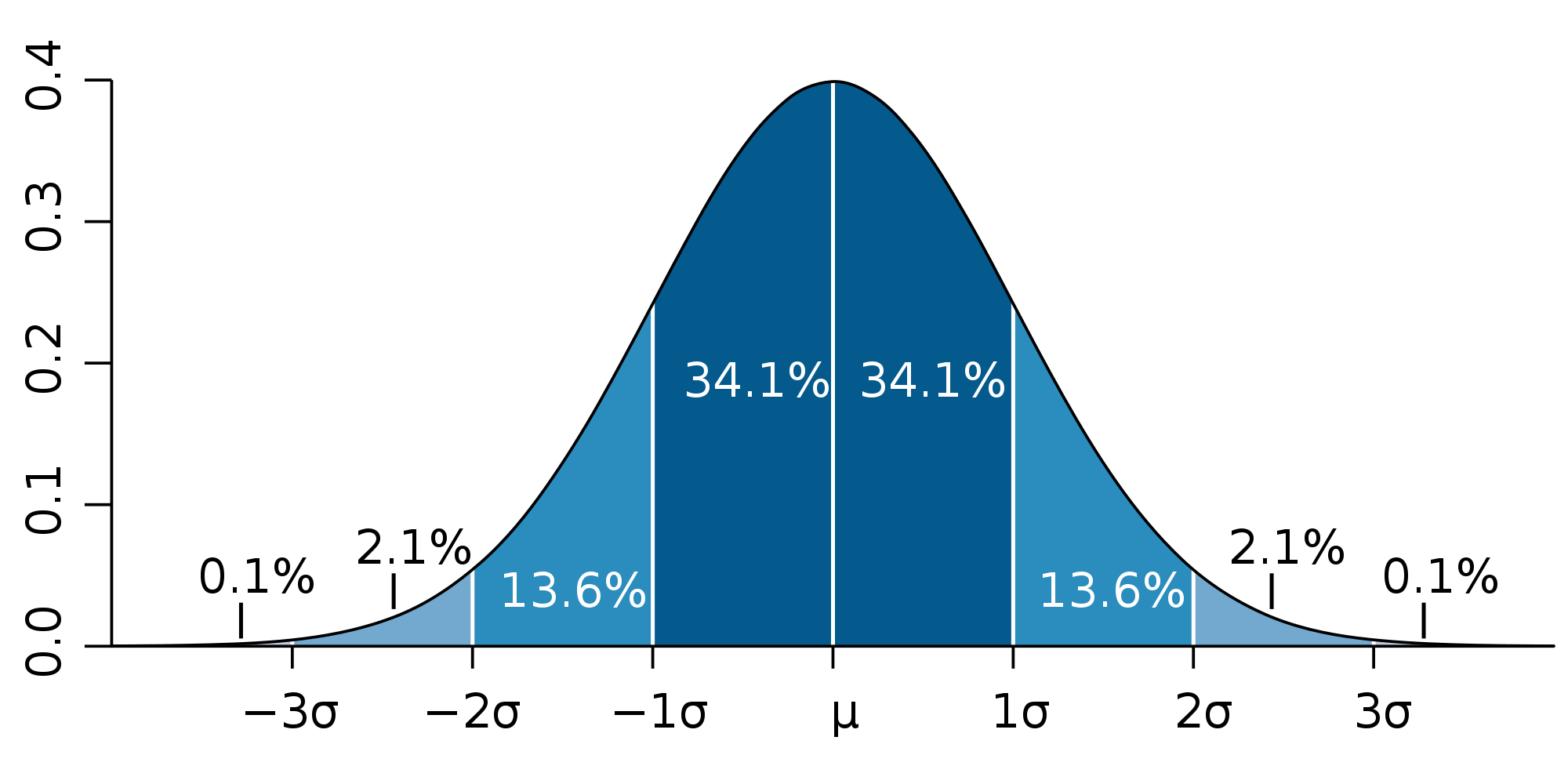
If a distribution is not assessed as normal it may be possible to transform the data to a normal set by, for example, converting by logorithms.

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | data\_review.R |
| Input | Data file, single column |

## Standard Deviation

The standard deviation is just a way of describing the spread of data in the normal curve.



A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

### Properties of standard deviations

For constant c and random variables X and Y:

where var() and cov() stand for variance and covariance, respectively.

NOTE: Usually, covariance is not an issue for chemists as the variables are usually independent. Usually. An example where they are not is in the calculation of the MU of Energy. The value for carbohydrate is not independent of the values for fat and protein and allowance must be made for covariance. This is done most easily by using the Monte Carlo method. (See later)

For addition and subtraction, where

For the compound result:

the combined standard deviation is:

For a complex product (and division) equation:

### z score

A z-score is a normalised standard deviation where the mean is zero and the standard deviation is 1. This permits the comparison of distributions by reducing them all back to a common metric.

### Pooled Standard Deviation

Pooled standard deviation is the square-root of the pooled variance.

The reason for pooled standard deviation is that you may sometimes have a number of small equivalent trials and a combined standard deviation will give a more robust value than any of the single, small population, estimates.

Pooled variance is calculated by

or with simpler notation,

where sp2 is the pooled variance and k is the number of samples being combined.

### Standard error of the mean

The standard error of the mean (SEM) is the standard deviation of the sample-mean's estimate of a population mean.

SEM is usually estimated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

where

sd = the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population), and

n = the size (number of observations) of the sample.

Consequently, the SEM is always smaller than the sample standard deviation and this is why an average is always a better estimate than a single test.

### Determining Standard Deviation in the Laboratory.

Where a sample is tested multiple times, such as a control sample, the standard deviation is easily determined by the normal equation or by the function STDEV(range) in Excel.

If you do not have access to multiple testing of a single sample, the standard deviation can be estimated from the difference of a series of duplicates.

Note: The data needs to be of comparable matrices and comparable levels of analyte.

where x1 and x2 are the duplicates and n is the number of duplicate pairs.

Repeatability standard deviation is determined from duplicates and reproducibility standard deviation is determined from retests.

### Effect Size

Calculate the effect size by subtracting the control group mean from the final group mean and dividing the result by the standard deviation of the control group. (Sometimes an arithmetic mean of the two sd’s is used.) The greater the effect, the greater the impact of the ‘treatment’. A result near zero indicates no effect.

## Hypothesis Testing

Hypothesis testing centers on the null hypothesis, Ho, that there has been no effect from the test or tests.

### Type I and Type II errors

Type I error: the null hypothesis is rejected, even though it is true.

Type II error: a null hypothesis is accepted, even though it is false.

### Confidence Intervals

The most common metric for the significance of a statistical test (and hence whether a hypothesis is true or not) is the p-value. Of equal importance, some say more importance, is to determine the confidence intervals around a result.

### Duplicate Repeatability

Repeatability, in the laboratory sense, is how close you expect two duplicate results to be. Same sample, tested in duplicate, together.

where sr = standard deviation for repeatability duplicates.

### Interim Precision

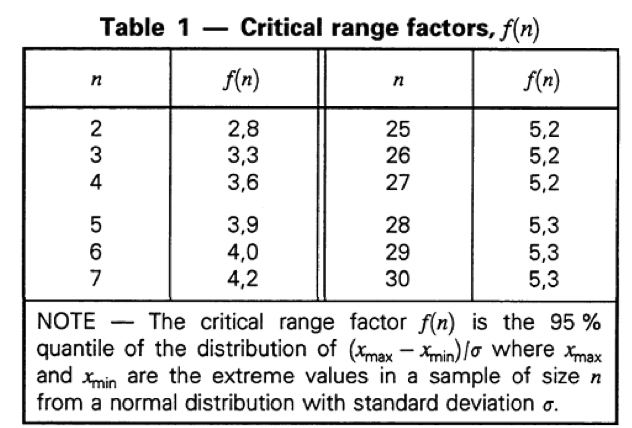
Interim Precision means reproducibility, in the laboratory sense, and is how close you expect two replicate results to be. Same sample, different batches (and possibly days, analysts and instruments.

where sR = standard deviation for replicates.

True reproducibility is a sign of the variance that can be expected between laboratories and is reflected in proficiency programs.

Reproducibility > Interim Precision > Repeatability.

The following table (From ISO 5725-6) gives the expected ranges associated with multiple testing of samples:

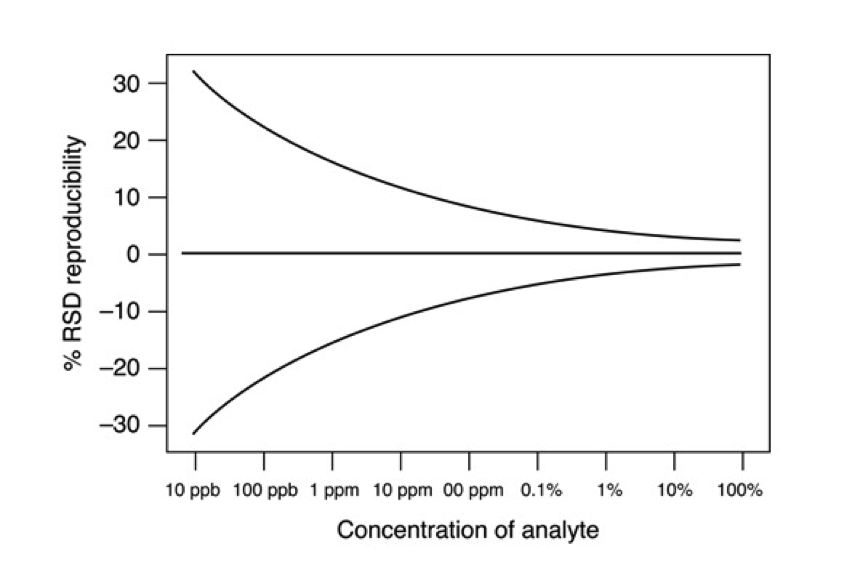


### Horwitz’s Trumpet

Dr William Horwitz looked at the relative standard deviation for several thousand proficiency programs and found that there was a relationship between the level of analyte being determined and the relative standard deviation of the results of the participating laboratories. He could then predict the RSD of another progam.

where c is the concentration of analyte, expressed in g/g. (eg 1 mg/kg = 0.000001g/g)

A Horrat Value is the ratio of a proficiency program’s RSD to Horwitz’s predicted value.



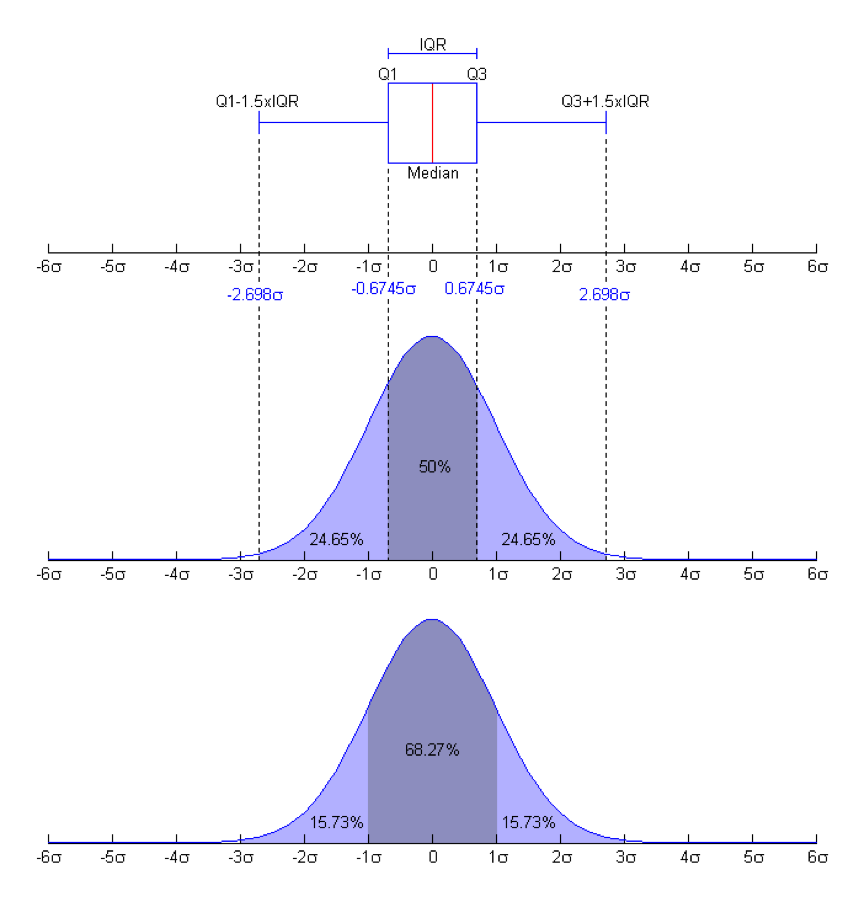
#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | horwitz.R |
| Input | xxx |

### Robust Statistics

Robust statistics are not influenced markedly by outliers or extreme results. For example the median will remain unchanged if one result is extreme where the mean can be dramatically affected.

Proficiency programs, such as Global Proficiency, use robust statistics because the participation group is often so small.



### Outliers

Outliers can be identified using Grubb's estimates but I prefer to use robust filtering using IQRs to omit outlying points.

## Robust Statistics

### IQR etc

The relationship between a regular standard deviation and a robust standard deviation is:

where IQR is the interquartile range of the sample, is a consistent estimate of σ if the population is normally distributed. The interquartile range IQR is the difference of the 3rd quartile of the data and the 1st quartile of the data.

### Robust z score

### Standard Difference

Conceptually, a standard difference is similar to a z-score.

The value for the limit is an agreed value for the test. It is, in effect, a pooled standard deviation for the test and most useful when only a few laboratories are participating in a program.

### Student's t Test

#### Assumptions

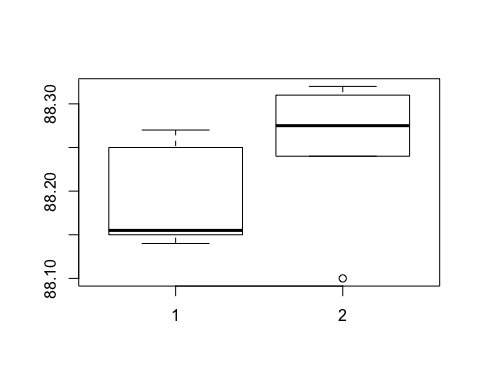
* Random sampling
* Normality

Types:

* One sample t-test : is the mean of the sample different to a hypothesised mean?
* Repeated measures t-Test : Test and retest. Are the means of the two result sets different?
* Independent groups t-Test : different participants have performed in each condition.

The data:

t1 <- data.in$A  
t2 <- data.in$B  
  
boxplot(t1, t2)



summary(data.in)

## A B   
## Min. :88.14 Min. :88.10   
## 1st Qu.:88.15 1st Qu.:88.25   
## Median :88.16 Median :88.28   
## Mean :88.19 Mean :88.25   
## 3rd Qu.:88.23 3rd Qu.:88.31   
## Max. :88.27 Max. :88.32

t.test(t2, t1, paired = TRUE)

##   
## Paired t-test  
##   
## data: t2 and t1  
## t = 2.4769, df = 5, p-value = 0.05605  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.002521843 0.135855177  
## sample estimates:  
## mean of the differences   
## 0.06666667

The p-value is (just) greater than 0.05 so we are not able to say that the two sets are different. But it is borderline.

Note the outlier (circle) in set 2.

Conceptually, a t-test is identical to a signal to noise ratio.

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | t.tests.R |
| Input | xxx |

### F test

The F-test is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be 1.

Assumptions / Notes \* The larger variance should always be placed in the numerator \* The test statistic is F = s1^2 / s2^2 where s1^2 > s2^2 \* Divide alpha by 2 for a two tail test and then find the right critical value \* If standard deviations are given instead of variances, they must be squared \* When the degrees of freedom aren't given in the table, go with the value with the larger critical value (this happens to be the smaller degrees of freedom). This is so that you are less likely to reject in error (type I error) \* The populations from which the samples were obtained must be normal. \* The samples must be independent

### ANOVA

If you need to compare more than one mean, you use ANOVA.

Assumptions

* Population normality
* Homogeneity of variance (Consider Levene's Test)

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | one.way.anova.R |
| Input | Data file, multiple labelled columns |

The 'ind' MS figure represents the between groups variance; the Residuals is the within groups.

ANOVA permits the comparison of two or more groups to compare their means. It will tell you that they are not all the same but will not tell you which one is different.

The wording of a typical statement: "There were no statistically significant differences between group means as determined by one-way ANOVA (F(2,18) = 0.399, p = 0.677)"

An estimate of repeatability standard deviation, sr, is obtained by the square root of the Within Groups Mean Square value. ( sqrt(0.0859) = 0.293, a variance of 0.0864, cf the three individual variance results).

An estimate of the between group standard deviation is given by

where n = the number of data points in a sets (=7 in the example).

An estimate of the intermediate precision standard deviation is given by

If the between variance is smaller than the within variance, then the means are really close to each other. When F<1 or F=1, between-group variance is smaller than within group variance, suggesting that the groups differ as much as one would expect due to sampling error. If there is no effect due to the different groups, the two MS values should be the same.

If differences across sample means are "large" relative to differences within samples, then we should reject the null hypothesis that the samples are all drawn from the same population. If the relationship is the other way, differences within samples is greater than differences between samples, we cannot say the means are different.

## Tukey HSD test

The Tukey HSD test assesses the differences between data sets. The ANOVA table has a p-value of 0.0106 indicating that the means are not all the same but it is Tukey who tells you which ones are different.

An estimate of repeatability standard deviation, sr, is obtained by the square root of the Within Groups Mean Square value. ( sqrt(0.0859) = 0.293, a variance of 0.0864, cf the three individual variance results).

## Rounding

The correct level of rounding for an analysis is indicated by the standard deviation of that analysis (2). If need be you can round to a less accurate level that the level indicated by this rule but not to a higher level of accuracy unless it is at the client’s specific request (eg project work).

**Rounding Rule:** A result should be rounded to the power of ten that is immediately less than half the standard deviation of that analysis.

**Example 1:** If the sd of the analysis is 0.5, half the sd is 0.25 and the power of ten immediately less than this is 0.1 so the results should be rounded to one decimal place.

**Example 2:** If the sd of the analysis is 3, half the sd is 1.5 and the power of ten immediately less than this is 1 so the results should be rounded to the nearest whole figure.

## Curve Fitting and Least squares

R2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values, or in the case of simple linear regression, between the outcome and the values being used for prediction. In such cases, the values vary from 0 to 1.

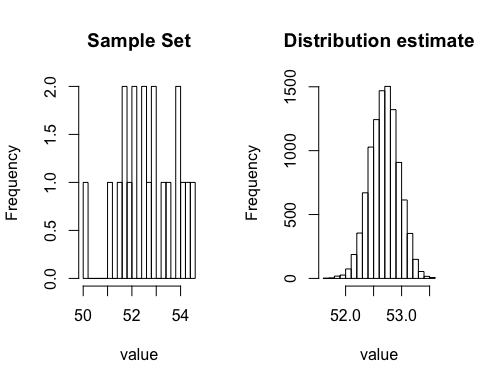
#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | X-Y.scatterplot.R |
| Input | Data file, Two labelled columns |

### Bootstrapping

Bootstrapping is where a small dataset is repeatedly resampled (with replacement) to produce a better estimate of the background distribution.

**Assumption:** the small set contains values that are all equally valid.



#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | xxx.R |
| Input | Data file, single labelled column |

## Monte Carlo Estimates

A Monte Carlo esitmate lets you estimate hard to calculate values. An example is the MU of Energy in a Nutrition Panel, where the value depends on Fat, Protein and Carbohydrate but Carbohydrate also depends on fat and protein. There is covariance that must be calculated.

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | xxx.R |
| Input | xxx |

# Part B : Laboratory Statistics

## Data Exploration Guidelines

1. Know what you are trying to prove. Correlation? Differences?
2. Determine what data you need to collect.
3. Always plot your data. Sometimes that is all you need to do for the answer to your question to be obvious.

* Regular plots
* X-Y plots
* Box Plots

Once you have had a look at your data, decide whether it is complete and whether it needs to be ‘cleaned’.

Incomplete data may have areas where your feel that more data is needed to present a complete picture.

Cleaning data is not a bad thing and can involve \* removing blanks, \* removing ‘less thans’ (<), \* removing text and \* removing obvious outliers.

Is your data normally distributed? Do you need to transform it to a format that is normally distributed? Check skewness and kurtosis.

Are there trends? Trending data can sometimes be normalised by plotting the point to point changes.

## Tidy Data - Presenting Data for statistical analysis

Tidy Data is easy to work with.

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be linked.

## Proficiency programs

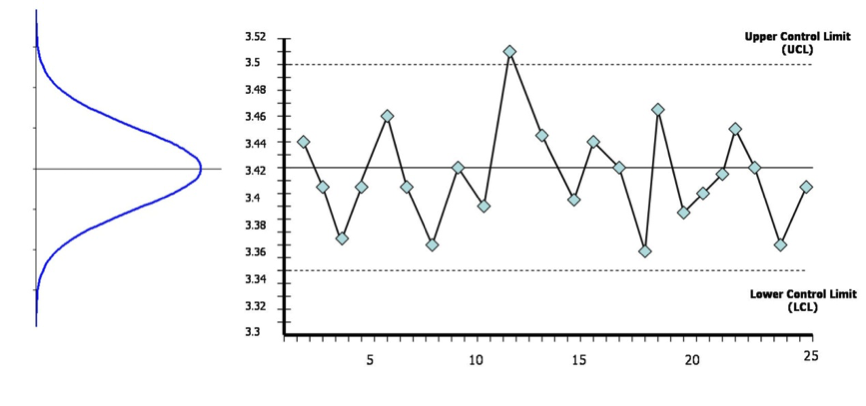
### En-Score

where

## Control charts

Control Charts, as the name suggests, help determine whether a process is under control.

Conceptually, they are a 3-D Normal curve. The majority of points should fall in the middle, evenly across both sides of the middle, with no trending.



### Trending rules

|  |  |
| --- | --- |
| Situation | Comment |
| One point outside 3sd (UCL, LCL) | Out of control |
| Seven or more consecutive points on one side of the mean | Bias |
| Six or more consecultive points trending up or down | Trending |

### Multiple Decision Points in a Control Chart

Sometimes it is decided to use more than one limit in a control chart. Be aware that this can result in needless retests. Say you are looking at four criteria in your control chart, any one of which has a 5% chance of exceeding the 2sd control limit. In parallel this way, there is a much larger chance of the batch failing. The probability of atleast one criteria failing is:

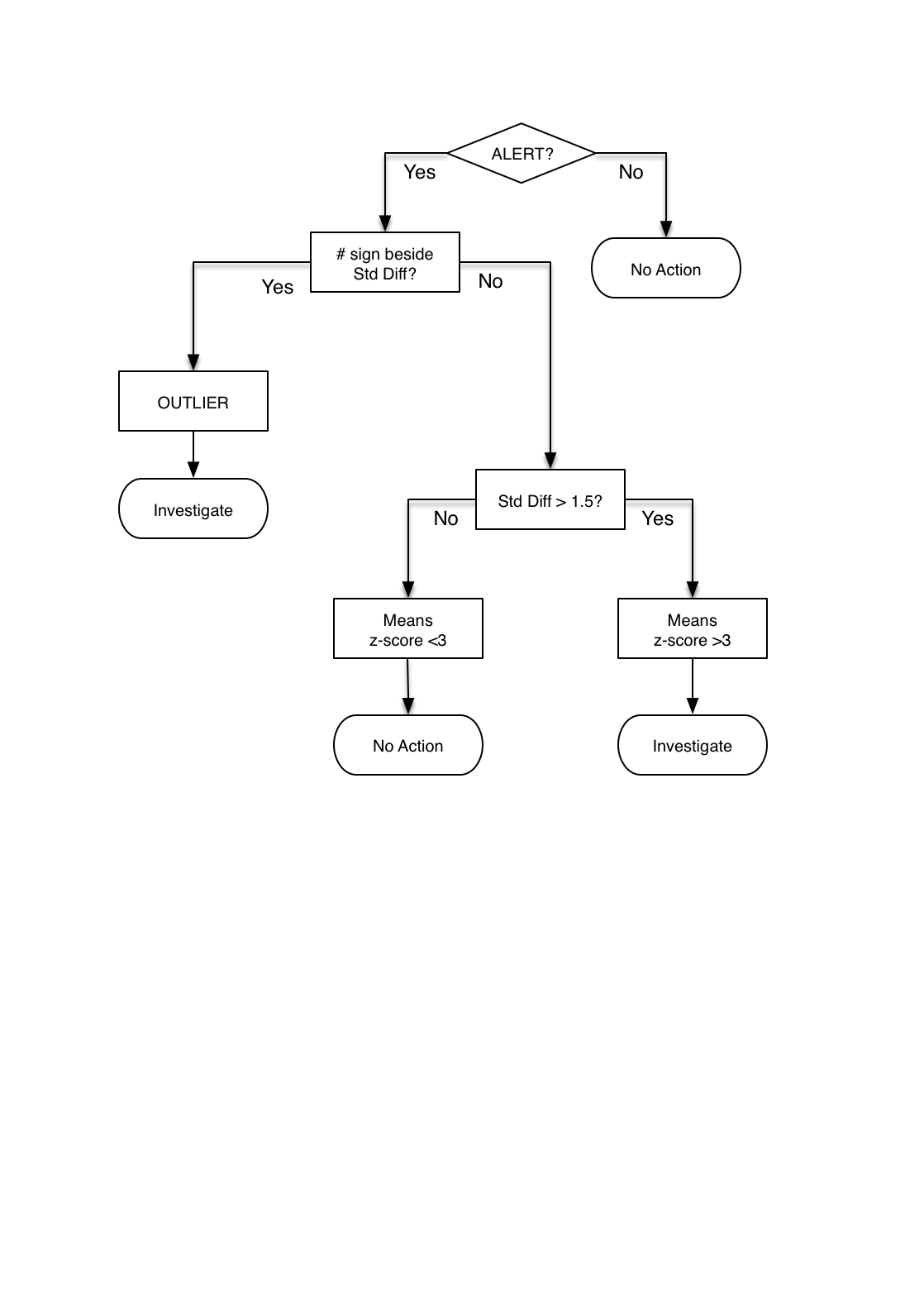
## Two IRM/SRMs in a batch

|  |  |  |
| --- | --- | --- |
| Criteria | Response | Rationale |
| Both IRMs within the ±2sd range. | System in control | no action |
| One IRM outside the ±2sd range, one within the ±2sd range | No action but monitor. | There is a 5% chance that this will happen randomly. |
| Both IRMs outside the ±2sd range. | Action required. | There is only a 0.25% chance of this happening randomly. |
| Either IRM is outside the 3sd range. | Action required. | There is only a 1% chance of this happening randomly. |

Note: the IRMs should not be run sequentially. With things like fat and ash it is not a big issue as the samples are all, in effect, tested separately but, in an instrumental batch, it is preferable that the IRMs are at each end of the batch.

## Interpreting GP reports

* The Limit = 95% confidence level (2sd)
* Outliers = 2 x limits = 4sd (so, seriously outlying). Identified with hash mark.
* Standardised Difference = (result - median)/limit
* Alert = Standard Difference >1 (>2sd)
* Where two samples are tested, Bias = average of the two standard differences.
* Bias alert when bias >0.5



Note: GP 'alerts' are based on 95% statistics, so there is a 1/20 change of a random alert. \* Check for transcription errors \* Check for calculation errors \* Ask for more sample and retest.

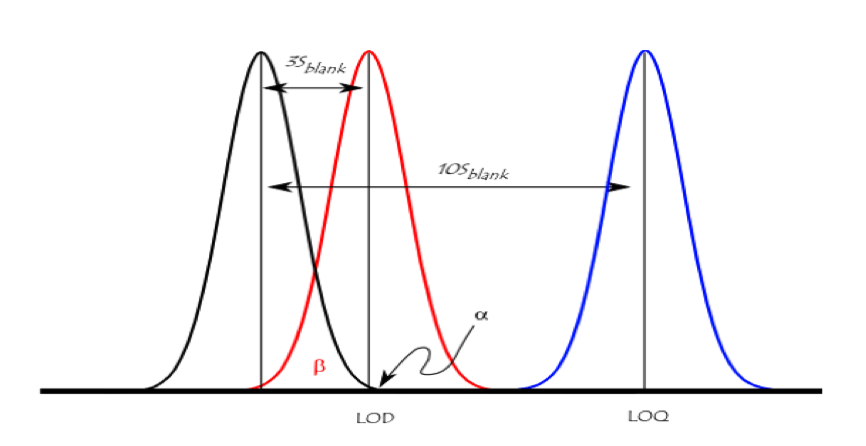
## Validation of methods

A typical validation project will consider the following items. Whether they are done will vary, case by case.

* Linearity
* LOD/LOR
* Matrix Effects
* Sensitivity
* Selectivity
* Precision
* Bias
* Ruggedness (~)
* MU

### Linearity

### Limit of Detection (LOD) & Limit of Reporting (LOR)



To assess LOD and LOR it is common to look at the standard deviation near the LOR. The black curve, above, represents the standard deviation as it brackets the blank. The red curve is centered 3sd and represents the point where the mean value idf clear of the blank variation but there is still an overlap between the blank and the detection curve. You may have detected the analyte but quatitationis uncertain. The blue curve, 10sd, is clear of the other curves and has a realistic likelyhood of quantitation.

### Standards curves

A standard curve is an X-Y curve and is a form of correlation.

There are a number of assumptions:

* Data is in pairs
* Data is normal
* The relationship is linear
* Homoscedacity - constant variance over the range.

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | X-Y.scatterplot.R |
| Input | Data file, two columns |

## MU

Measurement Uncertainty is a value range around a laboratory result that the lab is 95% confident contains the true value. It represents a plausible range of results around the nominal result.

### Precision

How close the laboratory results are to each other.

### Bias

How close the laboratory results are to a known value.

### UoBias

A function of the laboratory's precision and the uncertainty of the value for the standard or reference value.

Even when no bias appears to exist in a sample, there will be uncertainty as to whether any bias exists.

The diagram below compares the laboratory value to the ‘Truth’. The laboratory value has the appropriate uncertainty spread around it, however in reality there is no absolute ‘Truth’. Both the determined value and the certified “truth” value will have an uncertainty around them.

## Expansion Coefficient

Often 2 is used as a default value for k, the expansion coefficent. While this is adequate for large sample populations (>30), the true value should be used for small sample sizes.

## Expansion Co-efficient (n=10, p=0.95) = 2.26

## TOST (Two One-Sided t-Tests)

TOST tests assess whether there is a practical difference in means. You must pick a threshold difference for which smaller differences are considered practically equivalent. The most straightforward test to construct uses two one-sided t-tests from both sides of the difference interval. If both tests reject (or conclude that the difference in the means differs significantly from the threshold), then the groups are practically equivalent.

#### Data Assessment using the 'dts.quality' package

|  |  |
| --- | --- |
| Item | Input |
| R Script | xxx.R |
| Input | xxx |