Chemistry Statistics

Lee Kennedy

20 April 2016

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.

# 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

Using Excel’s descriptive statistics function you can get an indication of a dataset’s ‘normality’. Similarly, in R, 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.

summary(data.in)

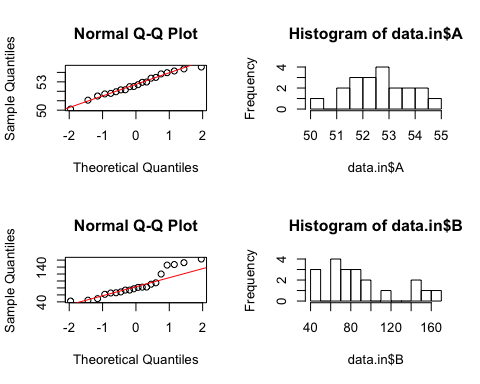
## A B   
## Min. :50.12 Min. : 42.35   
## 1st Qu.:51.91 1st Qu.: 65.98   
## Median :52.60 Median : 79.56   
## Mean :52.69 Mean : 89.20   
## 3rd Qu.:53.56 3rd Qu.:101.02   
## Max. :54.57 Max. :163.10

describe(data.in, ranges = FALSE)

## vars n mean sd skew kurtosis se  
## A 1 20 52.69 1.18 -0.20 -0.79 0.26  
## B 2 20 89.20 36.86 0.74 -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 (higher values) has a positive skew. \* An asymmetrical distribution with a long tail to the left (lower values) has a negative skew.

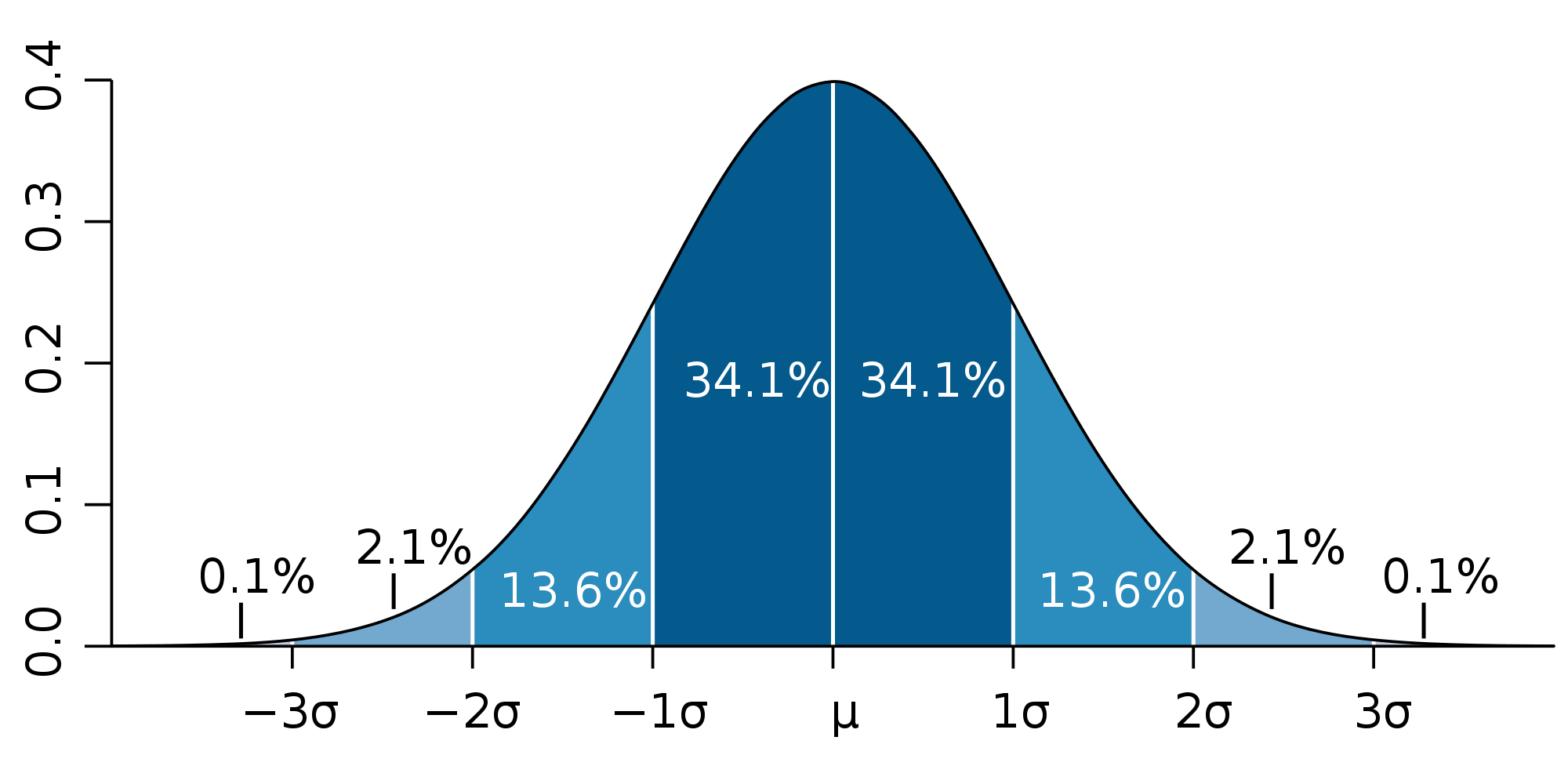
### Kurtosis

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.

## Standard Deviation

The standard deviation is just a way of describing the spread of the curve. The spread is described by the variance. The standard deviation is the square root of the variance, which means that it has the same units as the mean.

A small standard deviation means a narrow shaped ‘bell’, a large standard deviation means a low flat ‘bell’.



Standard Deviation shows how much variation there is from the "average" (mean). It may be thought of as the average difference of the scores from the mean of distribution, how far they are away from the mean.

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.

and

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. n − 1 is used instead of n for the same reason it may be used in estimating variances from samples (i.e. Bessel's correction).

### 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 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.

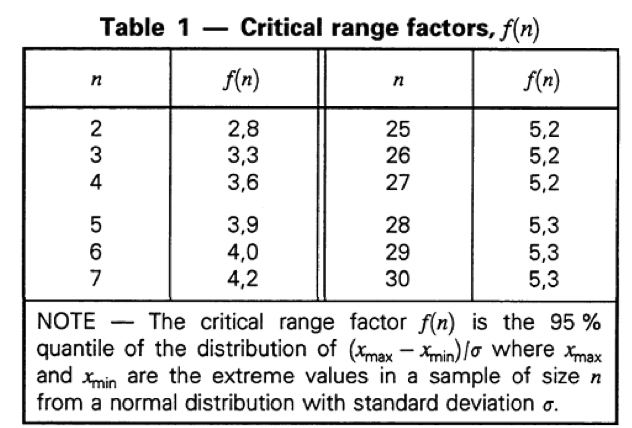
### Interim Precision

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.

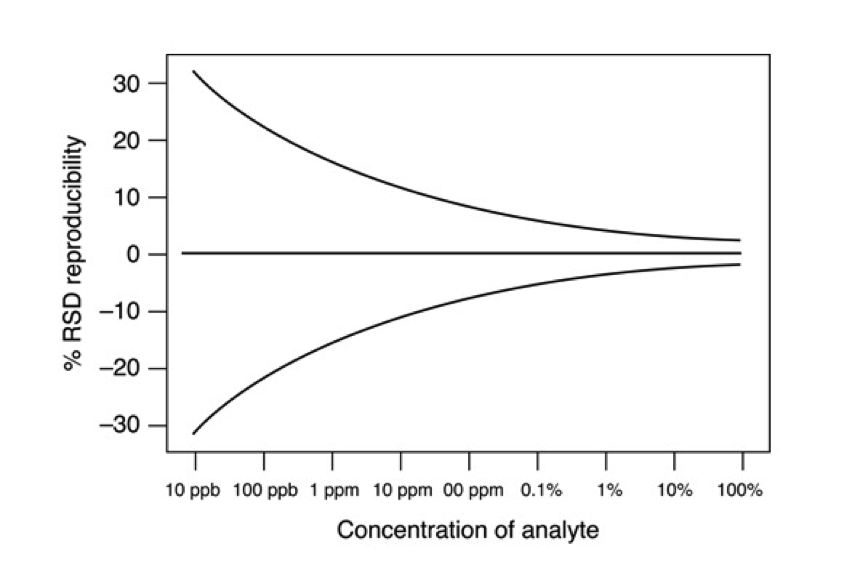


### Horwitz’s Trumpet

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)

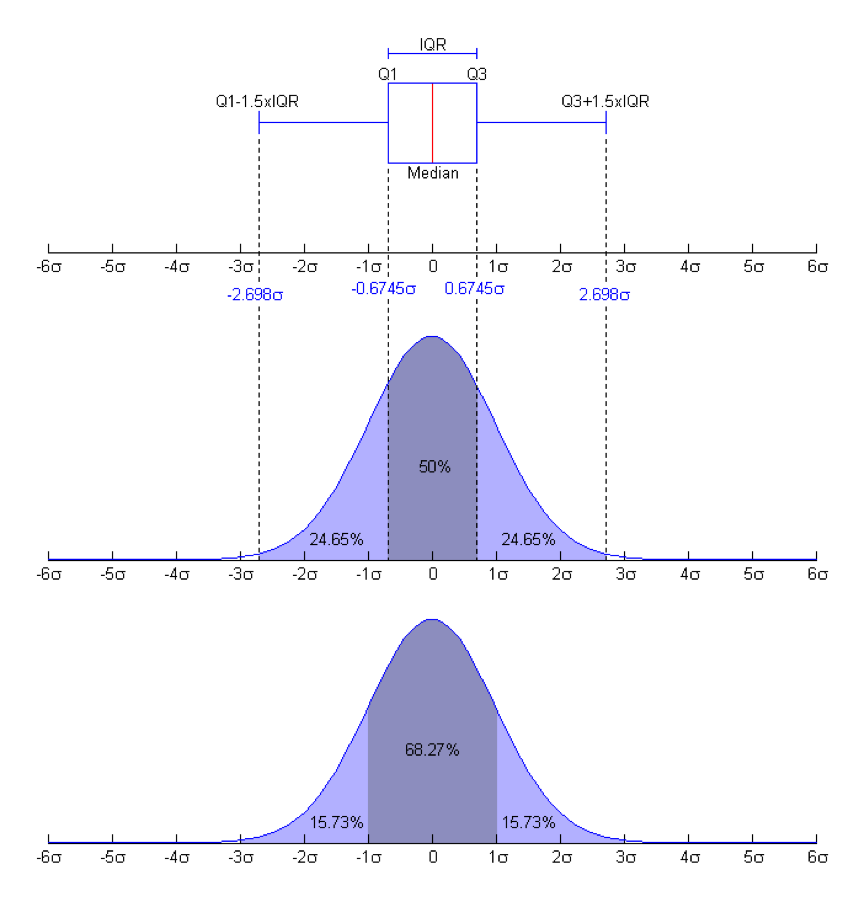
A Horrat Value is the ratio of a proficiency program’s RSD to Horwitz’s predicted value.



### 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

### Robust Statistics

## IQR etc

sd ~ 0.7413 x IQR

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. The asymptotic relative efficiency (ARE) of this estimator with respect to the one from sample standard deviation is 0.37. Hence, for normal data, it is better to use the one from sample standard deviation; when data is with thicker tails, this estimator can be more efficient.

### Robust z score

Robust z= (Result-Median)/(0.7413×IQR)

### Standard Difference

Std Diff= (Result-Median)/Limit

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

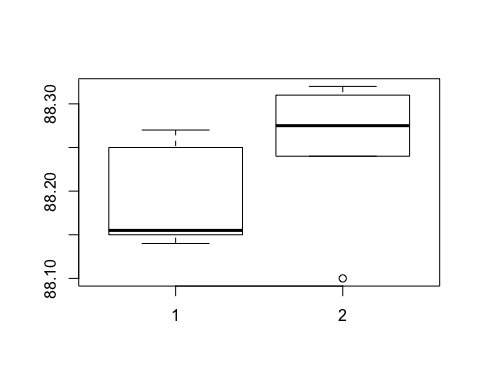
The value for the limit is an agreed value for the test. It is, in effect, a pooled standard deviation for the test and most useful when only a few laboratories are participating in a program.

### Student's t Test

The data:

## A B  
## 1 88.25 88.32  
## 2 88.27 88.31  
## 3 88.14 88.26  
## 4 88.16 88.29  
## 5 88.15 88.10  
## 6 88.15 88.24

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.

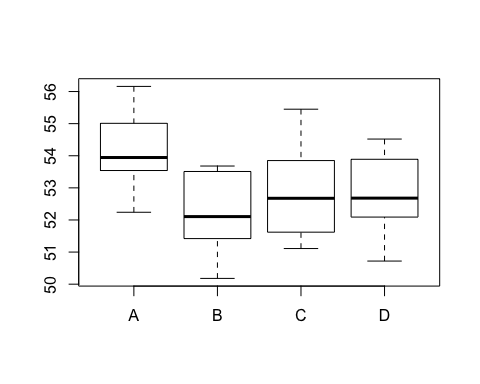
### ANOVA

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

## A B C D  
## 1 52.92 52.37 55.45 53.89  
## 2 52.24 51.19 51.23 50.72  
## 3 53.90 51.54 52.34 54.52  
## 4 53.89 53.58 51.11 52.48  
## 5 53.99 53.68 54.44 51.81  
## 6 55.01 50.18 53.85 52.68  
## 7 53.54 53.51 52.47 53.35  
## 8 56.16 52.26 52.88 52.68  
## 9 54.36 51.42 52.89 52.09  
## 10 55.07 51.95 51.62 54.52

## A B C D   
## Min. :52.24 Min. :50.18 Min. :51.11 Min. :50.72   
## 1st Qu.:53.63 1st Qu.:51.45 1st Qu.:51.80 1st Qu.:52.19   
## Median :53.95 Median :52.10 Median :52.67 Median :52.68   
## Mean :54.11 Mean :52.17 Mean :52.83 Mean :52.87   
## 3rd Qu.:54.85 3rd Qu.:53.23 3rd Qu.:53.61 3rd Qu.:53.76   
## Max. :56.16 Max. :53.68 Max. :55.45 Max. :54.52

boxplot(data.in3)



data\_st <- stack(data.in3)  
  
anova\_x <- aov(values~ind, data=data\_st)  
summary(anova\_x)

## Df Sum Sq Mean Sq F value Pr(>F)   
## ind 3 19.65 6.551 4.325 0.0106 \*  
## Residuals 36 54.53 1.515   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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

## 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. From the following table, only the A-B comparison is significant (p-value = 0.0062).

TukeyHSD(anova\_x)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = values ~ ind, data = data\_st)  
##   
## $ind  
## diff lwr upr p adj  
## B-A -1.940 -3.4223779 -0.4576221 0.0061798  
## C-A -1.280 -2.7623779 0.2023779 0.1110452  
## D-A -1.234 -2.7163779 0.2483779 0.1313395  
## C-B 0.660 -0.8223779 2.1423779 0.6313692  
## D-B 0.706 -0.7763779 2.1883779 0.5796337  
## D-C 0.046 -1.4363779 1.5283779 0.9997866

## Curve Fitting and Least squares

r R2 In statistics, the coefficient of determination, R2 is used in models for the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

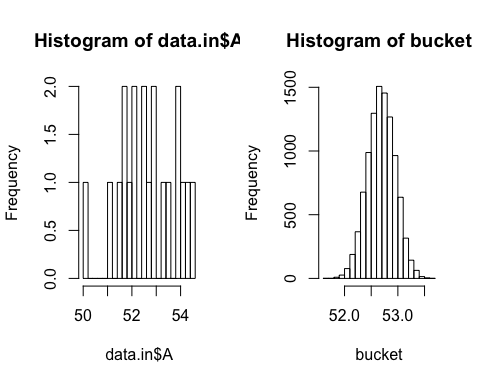
In this case, R2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values, or in the case of simple linear regression, between the outcome and the values being used for prediction. In such cases, the values vary from 0 to 1.

"Approximately seventy percent of the variation in the response variable can be explained by the explanatory variable. The remaining thirty percent can be explained by unknown, lurking variables or inherent variability."

Significance of R2

### Bootstrapping

data.in <- read.csv("/Users/Study Old/Documents/GitHub/Chemistry\_Stats/data/Book1.csv", header=TRUE)  
  
bucket <- rep(NA,10000)  
n <- nrow(data.in)  
  
for(i in 1:10000){  
 samp = sample(data.in$A, n, replace=TRUE)  
 bucket[i] = mean(samp)  
}  
par(mfrow = c(1,2))  
hist(data.in$A, breaks = 20)  
hist(bucket, breaks = 20)



## Monte Carlo Estimates

# Laboratory Statistics

## Data Exploration Guidelines

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

## Stats Test Decision Tree

## Tidy Data - Presenting Data for statistical analysis

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

## 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

## 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.

# n sample points  
n <- 22   
  
# p = desired probability  
p <- 0.95  
  
# Expansion Co-efficient  
cat("Expansion Co-efficient = ", round(abs(qt((1-p)/2, n-1)),2))

## Expansion Co-efficient = 2.08

## TOST

## Standard addition

## Retest Acceptability

## Distribution Calculator

[On-line Distribution Calculator] (<https://gallery.shinyapps.io/dist_calc/>)