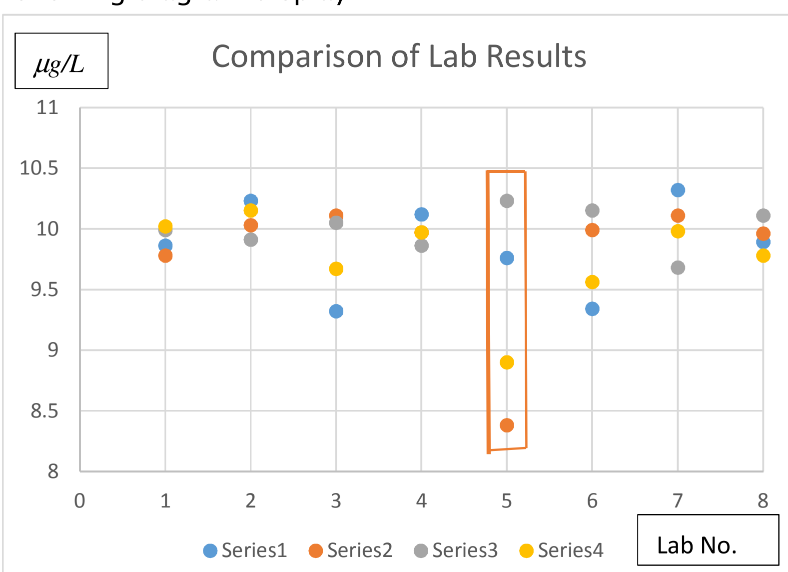
Cochran's C Test Statistic – What, Why and How

There are many test statistics for checking outlying sample mean results in a homogeneous population but relatively fewer for variance outliers. It is however, important for us to know if the estimate of a variance is significantly larger than a group of variances or standard deviations with which the single estimate is supposed to be comparable before deciding any outlier on the mean results.

For example, in a new test method development exercise, 8 laboratories were selected to study the ruggedness of a newly developed procedure in the analysis of a toxic analyte in drinking water. Due to the lab's respective technical competence, one would expect variability in their reported test results.   The collated test results were summarized below:

(µg/L)	LAB1	LAB2	LAB3	LAB4	LAB5	LAB6	LAB7	LAB8
Rep1	9.86	10.23	9.32	10.12	9.76	9.34	10.32	9.89
Rep2	9.78	10.03	10.11	9.97	8.38	9.99	10.11	9.96
Rep3	9.99	9.91	10.05	9.86	10.23	10.15	9.68	10.11
Rep4	10.02	10.15	9.67	9.97	8.9	9.56	9.98	9.78
Mean =	9.913	10.080	9.788	9.980	9.318	9.760	10.023	9.935
SD =	0.112	0.140	0.368	0.107	0.833	0.375	0.268	0.138

By plotting the replicated results against the participants, we obtain the following diagram display:



From the above diagram, we see that Lab 5 had rather big variation in its replicates.   We would like to know if this large variance is significantly different from the others.

The Cochran's C test statistic is popular for multi-sample test for equal variances. It is a one-sided upper limit variance outlier test, and is simple to use with the following assumptions:

a. The data set considered is a balanced design, i.e. each subject has equal number of replicates
b. Each set of replicates are normally distributed

The C test has been used as an alternative to Bartlett's, Levene's and Brown-Forsythe's tests in the evaluation of homoscedasticity (literally same variance) such as in a linear regression model. It is important not to mix this Cochran's C test up with the Cochran's Q test which is used in the analysis of two-way randomized block designs with different treatments in a design of experiments.

The Cochran's C test is built from the F-test, so we would expect it to be sensitive to deviations from normality. It has a simple estimation equation, involving the ratio of the maximum variance in the data set and the sum of all the variances:

$$C = \frac{s^2{}_{j,L}}{\sum_{i=1}^{N} s^2{}_i}$$

where $s_{j,L}$ = largest standard deviation in the data series j amongst the data set, s_i = standard deviation of data series with $1 \leq i \leq N$, whilst N being the number of sample groups (i.e. the number of laboratories in this example) involved in the data set.

The Cochran's upper critical value C_{UL} is obtained by considering factors such as N (number of groups/subjects), n (number of replicates in each group), α (level of significance desired) and a F_c value which can be obtained from the F distribution table or derived from a computer software for the F function, in the following equation:

$$C_{\mathrm{UL}}(\alpha, n, N) = \left[1 + \frac{N-1}{F_c(\alpha/N, (n-1), (N-1)(n-1))}\right]^{-1}.$$

In the above example, the largest variance (square of standard deviation) was noted in Lab5, and hence, $C = 0.629$. With $N = 8$, $n = 4$ and $\alpha = 0.05$, the C upper critical limit $C_{UL} = 0.438$. As $C > C_{UL}$ with 95% confidence, the null hypothesis of 'equal' variances was rejected and hence Lab5 was removed as an outlier.

It may be noted that the C test detects one exceptionally large variance at a time in the data groups. Upon eliminating the outlier, the test can be repeated until no further exceptionally large variance found. But by doing such elimination of multiple outliers, one may run into a risk of unnecessary excessive rejections, particularly when the data series is not normally distributed as it is sensitive to such departure from normality. Hence, the Cochran's C test for outlier must be carried out with caution.