# CASE STUDY REPORT:
# STATGRAPHICS CENTURION XV PROFESSIONAL

By

Christopher C.E. Hopkins

Professor/Director, AquaMarine Advisers®, Sweden: http://www.aquamarine.se

## 1. PREAMBLE

In the first half of the 1980s, I started using Statgraphics for biostatistical analyses related to marine ecology and fisheries when on the staff of the Norwegian College of Fishery Science at the University of Tromsø, Norway. However, I progressed to using Systat, BMDP and SAS due to these program packages being, at that time, substantially more advanced for the serious user than Statgraphics. However, in May 2008, after about 25 years' absence, I returned to the Statgraphics package, using Statgraphics Centurion Version XV Professional as my core statistical analytical tool. In the following, I'll describe my personal experiences with using Statgraphics Centurion Version XV Professional applied to a major project that I have collaborated with co-workers from Stockholm University, Sweden, and BONUS EEIG, Helsinki, Finland. In a relatively short document as this, I cannot deal with all aspects of Statgraphics in similar detail. So, I will concentrate here on the main aspects that I regard as most relevant based on my personal experiences.

## 2. THE TEST CASE STUDY

The test case project in which Statgraphics Centurion XV Professional has been applied is that of analyzing the peer-reviewed publications on Baltic Sea marine science, and its various disciplines, in the five-year period 2002 – 2006 by the nine Baltic Sea countries (*i.e.* Denmark, Estonia, Germany, Finland, Latvia, Lithuania, Poland, Sweden and Russia), and other countries. The database used for the statistical analyses includes, for example, information on the numbers and identity of the authors, what the scientific focus of the particular paper is, the research institutions involved, the journals published in, countries publishing either alone or through collaboration, and various bibliometrics (*e.g.* impact factor, immediacy index, citation half-life). The aim of our project has been to record and analyze the scientific output from Baltic Sea research during 2002-2006, based on 1975 Institute for Scientific Information (ISI) ranked international scientific publications. The analyses have concentrated on elucidating the following aspects of research in relation to countries, R&D funding and identification of strengths and gaps:

- Volume in terms of funding (inputs) and number of papers published (outputs) – effectivity in output per unit cost, *etc.*;
- Focus regarding various areas and sub-areas (*e.g.* disciplines and topics) of science;
- 'Quality' based on bibliometric indices;
- Cooperation (co-publication) at the national and international levels.

## 3. THE FOCUS OF THE APPLICATIONS AND SELECTED TESTS

### 3.1 Data storage and handling facilities

As with many researchers, we started our database in Microsoft Excel. It was good to experience the ability to easily transfer data to and from Excel and Statgraphics data-files. This is obviously advantageous when several researchers, in different institutions, are collaborating using different statistical packages. Within Statgraphics there is an impressive number of Operators containing special functions for mathematical transformations, statistical summaries, distribution functions, data selectors and manipulators, creating new variables, *etc*. When working on data and analyses, these can be saved and cohesively interlinked via the a) *StatGallery* function providing a container in which graphical outputs may be saved and placed together for comparative viewing, including overlays, b) *StatReporter* which collects the results of one's data analyses in a presentational quality report, c) *StatAdvisor* which produces easy to understand interpretations of one's statistical results, and c) *StatFolios* which save the full range of analyses allowing them to be re-run again later. These aspects are creative and highly effective. The only advance I would like to see with these is an improvement in the way one identifies exactly which parts of one's database was used underlying the analytical print-outs. For example, I would like to see the ability to easily (automatically) write the name of the database file being used on the generated analytical results in the *StatReporter* and/or the *StatFolios*.

### 3.2 Exploratory data analyses

For me, the exploratory data analyses are some of the most useful and notable aspects of Statgraphics, as knowing one's data characteristics (*e.g.* normally or non-normally distributed and tests for normality, outlier identification techniques, the ability to compare with and approximate to various types of statistical distributions) are essential before deciding what are the appropriate analytical techniques to apply later. In Statgraphics, these exploratory techniques are numerous, and include: Box-and-Whisker Plots, Frequency Histograms, Barcharts and Pie Charts, Normal Probability Plots, Bubble Charts, Distribution Fitting (Uncensored Data), *etc*. The Distribution Fitting (Uncensored Data) was very useful - providing the opportunity to fit/examine 45 probability distributions to numerical data, as well as having a diverse range of analysis summaries, tests for normality, goodness-of fit tests,

Chi-squared and Kolmogorov-Smirnov tests, *etc*., production of frequency histograms, quantile- and quantile-quantile plots, and normal tolerance limits to name a few.

For categorical data one has at one's hands some excellent Tabulation, Crosstabulation and Contingency Table (CT) techniques. I particularly found the CT most useful for setting up tests of independence/dependence in frequency tables, via construction of observed and expected frequencies, generation of Chi-square statistics, and related *post hoc* tests such as calculation of 'adjusted residuals' identifying which 'cells' do not significantly correspond with the Null Hypothesis (*Ho*). For the study of co-publication (association) between Baltic Sea countries, this allowed one to set up a *Ho* of independence between countries, and to identify and quantify which countries were actually collaborating with others to an extent 'greater' or 'less' than expected according to the *Ho*.

### 3.3 A wide range of very good analytical techniques and approaches

I am particularly pleased with the wide range of applications/tests regarding non-normally distributed data where ANOVA and related techniques ought not to be used. Here having identified non-normality of the data, one should move into the world of non-parametric tests including underlying ranking-based techniques (*e.g.* Mood's median test, Mann-Whitney (Wilcoxon) test, Kolmogorov-Smirnov test, Kruskal-Wallis test). The Two Sample Comparisons and Multiple Sample Comparisons are well developed in Statgraphics and easy to comprehend and use, with the option of using the appropriate tests depending on the normality or otherwise of the data. I found the Multiple Sample Comparison (MSC)/Analysis procedures very good with respect to both normally and non-normally distributed data. For the normally distributed data, the Multiple Range Tests in the MSC were first-class regarding the *post hoc* options (*e.g.* LSD, Tukey HSD and Duncan tests). I have rarely seen better presented options and intelligible results. Where I would appreciate some more development by future versions of Statgraphics is in the non-parametric equivalents, or options, of the just-mentioned *post hoc* tests, which have progressed quite extensively in the methodology sphere in recent years. Bring this in soon, please, and then just about all one can wish for will be in place in this important area of statistics.

Bearing in mind that really rather a lot of data deviates significantly from normality, I was also gratified that such relatively challenging/advanced techniques as Box-Cox transformations/fitting are available to normalize data when appropriate. Moving in this direction allows one to then apply the standard tests for normally distributed data, subject to a good Box-Cox fit.

The Six Sigma menu of Statgraphics brings in another increasingly useful capability which organizes statistical procedures into a DMAIC (Define, Measure, Analyze, Improve, Contol)

paradigm (a favourite of engineers) for quality definition and control purposes (*e.g.* batch production or sub-samples). This is potentially very applicable also in the biostatistical sciences.

I am motivated in producing my outputs for high quality presentations and peer-reviewed publications. Here I was pleased by the graphical presentation possibilities and options of Statgraphics Centurion XV Professional, whereby one can select the appropriate layout, fonts, colours and line types/thickness of graphs/figures and tables. The outputs can also easily be transported into HTML files for web-browsers using the *StatPublish* facility. Exporting into Microsoft PowerPoint presentations is painlessly easy.

Finally, I'll draw attention to the essential techniques which I have applied in Statgraphics for tackling multivariate analyses and developing informative predictive models. These have included Multifactor ANOVA (for constructing a model describing the impact of two or more categorical factors on a dependent variable), Variance Components Analysis (designed to estimate the contribution of multiple factors to the variability of the dependent variable), Regression Model Selection (for helping selecting the independent variables to use in building a multiple regression model to predict a quantitative dependent variable), and – last but not least – General Linear Models (GLM, designed to construct a model describing the impact of one or more factors – quantitative or categorical, crossed or nested, fixed or random - on one or more dependent variables). These have functioned very well in Statgraphics Centurion XV Professional. In particular, the GLM facility provided an exceptionally versatile, practical and robust tool for analysis and modeling purposes in relating predictors/factors of the 'quality' and abundance of the publications produced by the countries we analyzed.

## 4. CONCLUSIONS

Having comprehensively tested Statgraphics Centurion XV Professional for about six months, I can categorically state that it is an impressive piece of statistical-analytical software. It is supported by first-class interactive help 'pop-ups' and extensive documentation. Perhaps one can, in expectation of even more, hope for additional emphasis to be given to providing some additional literature references/citations regarding the various statistical analyses. Further attention could also be devoted to developing non-parametric *post hoc* tests comparable to the parametric Multiple Range Tests. All-in-all, I can highly recommend Statgraphics Centurion XV Professional to serious statistics practitioners. I can't wait to test future updates and developments of this software.