Appendix 4

stark environmental

Freshwater ecology specialists

19 February 2010

Horizons Regional Council
Private Bag 11025
Manawatu Mail centre
Palmerston North 4442

Attention: Kate McArthur

Dear Kate

Re: Determination of ecologically meaningful changes in QMCI

Thank you for the opportunity to provide Horizons Regional Council (HRC) with some advice that may assist in formulating an agreed standard for inclusion in the One Plan whereby ecologically meaningful changes or differences in QMCI can be identified. I understand that such a standard should have a defensible statistical basis and would be used primarily in compliance monitoring programmes involving sampling of appropriately matched habitats at upstream (control) and downstream (impact) sites. However, a robust standard could also

- be used to define a threshold for a permitted activity to become a consented/discretionary activity
- serve as a minimum standard/target to aim for in measuring the effect of a consented point source
- provide a measure that assists with monitoring policy effectiveness.

The draft One Plan in Schedule D (Table D.16) contains the following condition:-

> *The QMCI shall exceed [5 or 6 depending upon the water management zone concerned], unless natural physical conditions are beyond the scope of application of the QMCI.*

In evidence presented at the One Plan hearing, various submissions were made that relate to the use of biotic indices for biomonitoring in the Manawatu-Wanganui region.

Stark Environmental Limited
PO Box 1831, Nelson 7040, New Zealand
10 Penzance St, Marybank, Nelson 7010, New Zealand
ph. 03 545 1766 027 340 3608 **home:** 03 545 1025
StarkEL@paradise.net.nz

Summary of One Plan submissions relating to invertebrate biomonitoring

HRC staff submitted that there be no more than a 20% reduction in QMCI between sites upstream and downstream of a discharge. This was intended to apply to the upstream control site and the first site downstream

Dr John Quinn recommended standards of 100 and 120 to apply to the MCI (analogous to the 5 & 6 standards for the QMCI that apply to defined sub-catchments). HRC noted that the MCI is used for State of the Environment monitoring and that the standard should apply also to the MCI-sb (for soft-bottomed streams and rivers).

Dr Russell Death considered that the QMCI was the most appropriate biotic index for assessing the effects of a potential discharge and/or other anthropogenic effects. Rather than choosing (arbitrarily, in his view) a 20% change as the threshold for determining whether a difference was environmentally meaningful or nor, he preferred "the more scientific approach of using the appropriate statistical tests and power analysis." However, he noted that "if a single percent change must be selected the 20% value is consistent with my observations of change in QMCI downstream from some of the worst sewage discharges in the region".

Mr Keith Hamill supported the development of standards/targets based on indices of macroinvertebrate communities such as the MCI and QMCI but suggested that the approach needed to be flexible are permit comparisons with references sites. Mr Hamill further recommended:-

- That the standard/target form MCI values should be set in relation to reference conditions (or use current MCI scores as a standard/targets when appropriate references conditions have not been defined
- That the standard/target form QMCI values should be set in relation to what causes a reduction in scores
- The inclusion of the MCI-sb (for use in soft-bottomed stream/river habitats)
- The inclusion of <u>discharges to water</u> to cause no more than a 20% reduction in QMCI between upstream and downstream of the discharge (in order to help focus attention on the cause of any decline in ecosystem health).

In his supplementary evidence (paragraphs 3.2 – 3.3) Mr Hamill discussed the complex matter of interpreting changes in QMCI values (with reference to Stark & Maxted 2007) and the difficulty of determining whether a discharge or other factors are responsible for the change. He suggested that the standard (i.e., no more than 20% reduction) should apply to changes in the QMCI that are "<u>caused</u> by the discharge in question".

In paragraphs 3.4 – 3.9 in his supplementary evidence Mr Hamill stated that QMCI targets should be set at what is an "acceptable difference". He noted that sampling effort affects the magnitude of difference that can be detected (with reference to Stark's (1998) table of Detectable Differences), and that the SQMCI could be included as a more cost-effective alternative to the QMCI.

Mr Hamill proposed the following wording for Schedule D:-

> *"Discharges to water to cause no more than a 20% reduction in Quantitative Macroinvertebrate Community Index (QMCI) score between appropriately matched habitats upstream and downstream of the discharge.*
>
> *Note: Where samples are collected using a hand net this standard shall also apply to the Semi-Quantitative Macroinvertebrate Community Index (SQMCI)"*

In the following sections of this letter I provide my views on statistical vs ecological significance and my comments on the submissions summarised above.

Biomonitoring

I have always believed that routine biomonitoring (whether it be for consent compliance or state of the environment) should be as cost effective as possible. There should be no need to measure or monitor everything, or to require a research-level of effort that becomes onerous and expensive for consent holders or ratepayers. Routine biomonitoring can be thought of as a watching brief. If adverse effects are detected, then that is the time for more intensive investigations in order to determine the cause/s. For more explanation see Stark & Maxted (2007).

HRC has proposed that the QMCI (which requires quantitative sampling) should be the biotic index used for consent compliance monitoring. Stark & Maxted (2007) maintained that the QMCI (and SQMCI) were more suited to compliance monitoring than to SoE monitoring, but noted that the MCI was also an appropriate index to use for compliance monitoring. In fact, New Zealand's longest-running freshwater biomonitoring programme (since September 1981) in the Kapuni Stream involves seasonal sampling (single kick-sample/site) from 10-14 sites with site-specific MCI targets and trends testing (Stark 2010).

I support the recommended wording for Schedule D as stated by Mr Hamill (see quote in italics above). I think the option of including the SQMCI would be a worthwhile addition, which would

open the way for much more cost-effective monitoring compared with that based on quantitative sampling and the QMCI. In my view, criticisms of the coded-abundance sample processing protocol (Duggan et al. 2003) are not serious enough to compromise the integrity of biomonitoring programmes or to out-weigh the cost-effectiveness of the SQMCI (cf. QMCI).

I agree with Dr Quinn's suggestion that standards of 100 and 120 should apply to the MCI, and see no reason why the MCI should be restricted to use on SoE monitoring programmes. Although bioassessments based on the QMCI, SQMCI, or MCI will inevitably differ in detail, all three indices are strongly correlated with each other (Stark 1993, 1998), and provide options for progressively reducing the costs of monitoring programmes. Perhaps there could be some discretion in the system that permitted more cost-effective monitoring programmes to be associated with some of the more minor discharges?

Comparisons with reference condition

Mr Hamill has suggested the use of a reference-based approach to determine natural reference condition for biotic indices. I concur with HRC's response (Supplementary Evidence of KJ McArthur, p47). Rather than aim for reference condition (which can be difficult to define in some cases – e.g., lowland rivers which are almost all impacted), I think there is plenty of scope for water managers to simply aim to ensure that best management practices are employed by industry (leading to improvement in discharge quality) and to monitor with the expectation that improvements in river health will occur with time. The reference condition approach can raise unrealistic expectations concerning the target river health that could be achieved.

The reference condition approach is not without its problems (and costs), is more complex, and, in my experience, its advantages tend to be over-sold (and disadvantages overlooked) by those who strongly advocate the abandonment of tried and true (existing methods) in favour of predictive models based on the reference condition (Stark 1999, Stark & Maxted 2007). Continued use of existing methods is consistent with my views on biomonitoring - simple and cost-effective is best.

An arbitrary standard or statistical testing?

I consider the proposed standard (i.e., no more than a 20% decrease in QMCI as a result of a discharge) to be a sensible approach to discharge consent compliance monitoring. Inherent in the choice of the 20% value is consideration of the errors associated with sampling and the degree of change that might be considered ecologically significant. The degree of change that

can be detected is very much dependent upon the between-replicate variability in QMCI and the sampling effort (i.e., number of replicates). A tighter standard would require greater sample replication in order to detect the difference with any confidence, and the smaller the difference the less ecological significance it is likely to have.

The 20% decrease in QMCI, although somewhat arbitrary, was selected because (in most situations) it is practical to detect such a difference with an acceptable level of effort (e.g., 5 replicate Surber samples per site) – more on that later. In addition, a 20% decrease in QMCI will normally result in a change to a lower quality class (e.g., Excellent → Good, Good → Fair, Fair → Poor) (using the quality classes defined by Stark & Maxted (2007)). A change of that magnitude is likely to have some ecological significance. Inevitably, though, a judgment call would be required in order to decide whether or not the decrease is entirely due to the effects of the discharge and/or really does have ecological consequences.

Statistical vs ecological significance

Karr & Chu (1999) noted that 'statistical decision rules are no substitute for biological judgment'. The objective of biological monitoring is to detect human-caused deviation from baseline biological integrity and to evaluate the biological – not statistical – significance of those deviations and their consequences. A statistically significant result (say $p < 0.05$) may not be indicative of an important ecological effect. Conversely, a statistically non-significant result ($p > 0.05$) might be biologically important. It is almost impossible to know for certain whether a significant biological effect is, or is not, present based solely on a *t*-test (or similar) result. Unfortunately, testing for statistical significant is (comparatively) easy, whereas determining whether something is ecologically significant is very difficult.

Although I agree with Dr Death that statistical tests should be employed when analysing biological data, I would not necessarily accept that the detection of a statistically significant decrease in QMCI is proof that a discharge has had a negative effect on the environment. As Mr Hamill has explained, there are other possible explanations (some of them 'natural') for between-site differences in QMCI values – the effect of the discharge is just one of them. In my view, it would be reasonable to expect a biomonitoring report to present the results of appropriate statistical testing, but unwise to have a rule in a plan that considers a statistically significant decrease in a biotic index value to be a breach of a consent. No statistical test that I am aware of can determine (by itself) whether or not a statistically significant difference is likely to have ecological significance, or has been caused solely by the discharge.

It is possible to detect statistically significant differences in QMCI values when the difference between means is very small if between-replicate variability is low. For example, if mean QMCI values at two sites based on 5 replicates were 6.48 (6.40, 6.50, 6.50, 6.50, 6.50) and 6.40 (6.40, 6.40, 6.40, 6.40, 6.40) a *t*-test reveals that this difference is statistically significant ($t = 4.0$, $p = 0.004$). In reality, QMCI values normally show greater between-replicate variability than this. However, this example does highlight the danger in relying solely of statistical tests for determining impact. No reasonable person would (or should) agree that a difference of only 0.08 in mean QMCI values between two sites has any ecological relevance (despite being statistically significant).

At the other extreme (and much more likely to be a problem) is the situation where a statistically significant difference exists but it is not detected. This can be an issue especially when sampling effort is insufficient to cover natural within-site variability, which then masks any between-site differences.

Arguments about whether or not statistically significant differences exist between natural populations might be irrelevant when it comes to considering whether or not ecologically significant differences exist. Often, statistical tests are based on the null hypothesis that there is no difference between populations. This may be a reasonable assumption when comparing the height of the same brand of beer bottles at two different bars, although when a statistically significant difference is not detected we cannot conclude that the heights are identical – only that a statistically significant difference was not proven. Lack of evidence for a difference is not the same as evidence of no difference. A larger sample could have lead us to find statistically significant differences, so a statistically non-significant result, in relation to a non-informative null hypothesis, means only that we simply don't know (Martinez-Abrain, 2007, 2008).

Given the inherent variability in natural populations, we would expect samples from two locations to be different, so testing the null hypothesis that they are the same is non-sensical (or inappropriate). We would inevitably find statistically significant differences provided we collect sufficient replicates. [Conversely, by not sampling enough we can sometimes guarantee that we will never detect a statistically significant difference. For this reason it is essential to have an appreciation of the relationship between sampling effort and statistical power].

A more appropriate question to pose would be whether the difference is greater than a given magnitude of interest. The difference could be expressed in various ways. For example, more than ½ the standard deviation, or more than a certain percentage (say 20%). Ideally the magnitude of the permitted difference should be less that the difference that could be considered to be an adverse ecological effect that was more than minor. It also needs to be detectable under a sampling programme that is not unreasonably onerous. The best

professional judgment of an experienced freshwater ecologist might be the best way to determine such a threshold (although some might consider that arbitrary).

Sample replication

Stark (1998) provided a table (Table 5) of detectable differences (DD) that indicated the difference between mean biotic index values that was required for the difference to be regarded as statistically significant. The primary value of this table was to provide some guidance when single samples were collected and it was not, therefore, possible to undertake statistically tests to determine whether or not differences were likely to be statistically significant. For example, two QMCI values (from single Surber samples) needed to differ by at least 1.37 for the difference to be statistically significant. That is a difference of about 27% around an average QMCI of 5.00. Cleary, collection of single Surber samples could be insufficient if one wanted to detect a 20% change in QMCI. At least two Surber samples would be required (DD = 0.97 or 19.4% for an average QMCI of 5.00). To detect a 20% difference at low QMCI values would require greater replication. For example at a QMCI of 2.00, detection of a 20% change could require as many as 12 replicates (Stark 1998: Table 5).

Power analysis can be used to determine the number of replicates required in order to detect a difference between QMCI values of a specified magnitude. For example, to detect a 20% difference in QMCI (say 5.00 vs 6.00), with a Type 1 error rate of 0.10, six replicate samples would be required to achieve a power of 80% (which usually is regarded as sufficient). [This is based on an estimate of the QMCI population standard deviation of 0.645 – from Stark (1998)). In other word, 80% of the time where a statistically significant difference exists, it would be detected. If 5 replicate samples are collected per site, power reduces to 72%. This analysis is based on an estimation of between-replicate QMCI standard deviation based on samples collected from throughout New Zealand, and may not be the best estimate for samples collected from the Manawatu-Wanganui region. With 5 replicates, 80% (or more) power would be achieved if the population standard deviation was 0.579 (or less). The results of power analyses generally are consistent with the DD method provided by Stark (1998).

However, the critical question is not the magnitude of the statistically significant difference in QMCI, but the change in QMCI that is ecologically significant. Statistically significant differences in QMCI can be detected with any degree of precision that is desired – it's a matter of deciding on the number of replicates to collect (and accepting the associated costs).

The table of detectable differences (Table 5 in Stark (1998)) can be used to determine how many replicate samples would be required in order to determine a 20% change in biotic index values. The number of replicates depends not only on the index concerned (i.e., MCI, SQMCI,

or QMCI) and the sampling method (i.e., hand-net or Surber), but also on the actual value of the biotic index. For example, more replicate Surber samples would be required to detect a 20% change when the QMCI is 2.00 compared to when it was 8.00. Table 5 in Stark (1998) would suggest that 12 replicate Surber samples would be required for the former, but a single Surber sample per site is likely to detect a difference of 20% for QMCI values around 8.00.

Table 1 Thresholds for biotic indices in order to detect a 20% change based on single samples or 2 – 12 replicates).

Replicates	Hand-net MCI	Hand-net SQMCI	Surber MCI	Surber QMCI
1	54	4.15	108	6.85
2	38	2.93	77	4.84
3	31	2.40	63	3.95
4	27	2.08	54	3.43
5	24	1.86	48	3.06
6	22	1.69	44	2.80
7	20	1.57	41	2.59
8	19	1.47	38	2.42
9	18	1.38	36	2.28
10	17	1.31	34	2.17
11	16	1.25	33	2.07
12	16	1.20	31	1.98

Table 1 provides thresholds of biotic index values and the number of replicates required to detect a 20% difference. For example, if the expected QMCI is 6.85 or greater, then a single Surber sample per site is likely to be required to detect a 20% difference. However, 8 replicates would be required to detect a 20% difference about a mean QMCI of around 2.42. Five replicate samples would be sufficient to detect a 20% change in QMCI at sites with mean QMCI values greater than 3.06. This means that five replicate Surber samples is unlikely to be sufficient to detect a 20% changes in QMCI in a river where the upstream QMCI is less than 3.06.

Table 1 also highlights the improved cost-effectiveness of the MCI and SQMCI when calculated from hand-net samples.

Interestingly, MCI values from Horizons' 2009 SoE monitoring sites ranged from 70 to 152 (Stark 2009). The data in Table 1 suggest that single hand-net samples should be sufficient to detect differences in MCI of 20%.

Trends testing

I am a strong advocate of the use of trends testing for analysing biomonitoring results when sampling has been undertaken on a sufficient number of occasions (Stark & Fowles 2006, Stark 2008). Trends testing allows one to determine whether stream health is improving, deteriorating, or staying the same, and avoids the complications inherent in between-site comparisons that Mr Hamill referred to.

Equivalence tests

Rather than testing for equal means, equivalence testing permits one to test for means that differ by a pre-determined amount. In my view, equivalence testing is the most appropriate and simplest statistical testing to assess compliance with a rule based on a 20% decrease in QMCI.

This is best illustrated by an example.

Assume we collect five replicate Surber samples from an upstream control site (A) and a downstream impact site (B). The mean QMCI at Site A is 4.8 (4.8, 4.7, 4.5, 4.9, 5.1) and the mean for Site B is 4.3 (3.8, 4.2, 4.5, 4.4, 4.6). A conventional *t*-test for independent samples would indicate that the QMCI for Site B is significantly lower than that for Site A (t = 2.887, p = 0.020).

However, if we assume *a priori* that we require the QMCI at Site B to differ by at least 20% from that for the control Site A (i.e., differ by at least 4.8 * 0.2 = 0.96 QMCI units), we can construct a confidence interval about the actual difference (which is 4.8 – 4.3 = 0.5) and determine whether it includes the predetermined tolerance range (i.e., -0.96 – +0.96).

If we assume the data are normally distributed, an interval based on the *t*-distribution can be used. The pooled standard deviation (SD) of the original 10 QMCI values is 0.369, so a 95% confidence interval around the estimated difference is given by the following relationship:-

The actual difference ± critical value of *t* for 2(n-1) degrees of freedom * pooled SD * sqrt (1/n+1/n)

where n is the number of values in each group (5 in this example).

In this case we have 0.5 ± 1.86 * 0.369 * 0.632 = 0.5 ± 0.434

So the confidence interval within which the difference does not exceed 20% is 0.066 – 0.934. Since the pre-determined tolerance range (-0.96 – +0.96) is outside this confidence interval we

can conclude that the QMCI values do not differ by more than 20% in this example - The opposite result to that obtained from a standard *t*-test.

This is an example of the type of statistical test that could be used to test compliance with the 20% difference rule, BUT, this test (assuming it was significant) does not prove that the difference was due to the effects of the discharge.

In my view it is preferable to run equivalence tests than to take the mean QMCI values at face value – clearly the QMCI for Site B was > 80% of the QMCI for Site A (so it's no real surprise that the equivalence test result agrees).

The role of power analysis

Power analysis (which Dr Death advocates along with appropriate statistical testing) is most properly used when designing a sampling programme to determine what effect size could be regularly detected (e.g., 80% of the time). Normally, a pilot survey would be undertaken to obtain estimates of variability that would then be used in power analysis calculations to determine how many replicates should be collected in order to determine an effect size that is deemed to be of practical importance (McBride 2005). Procedures for calculating the power of a performed test (which is what Dr Death is advocating) do exist (e.g., Zar 1996), but are no more informative than the *p*-value obtained already, and they violate the concept of power – which attaches probability to the long-run frequency of given outcomes, not the particular outcome from a given set of data. Power is a consideration in designing studies, not in analysing results, so using power analysis in this way is increasingly being discouraged (McBride 2005).

Conclusions

I can summarise my views as follows:-

- Statistical testing (such as t-tests) could be used to determine whether or not statistically significant differences exist between mean QMCI values (for example) at control and impact sites. However,
    - A non-significant result (in relation to the null hypothesis that there is no difference) tells us nothing except that a significant difference was not proven
    - A statistically significant result does not necessarily indicate that the difference has any ecological significance

- I do not believe that detection of a statistically significant difference between QMCI values (or any other biotic indices) upstream and downstream of a discharge (for example) should automatically be a breach of a consent. In other words, a consent, or the One Plan, should not include a condition or rule whereby there shall not be any statistically significant decrease in QMCI (etc.) between control (upstream) and impact (downstream) sites.
- Given the fact that sampling effort can have a major influence on the ability to detect statistically significant differences, it is important to specify the minimum sampling effort that must be employed in consent biomonitoring programmes. This could be specified in the One Plan, on consent conditions, or (preferably I suspect) covered by a condition that any biomonitoring programme must be approved by HRC. Certainly, I see little need for sampling effort greater than 5 replicate Surber samples per site – more would be onerous on the consent holder. In my view, if biotic indices calculated from 5 replicate samples do not show statistically significant differences, then ecologically significant differences are unlikely.
- In my view, HRC should leave the door open for the use of hand-net sampling and the MCI or SQMCI in compliance monitoring. Kick samples can be collected where it is difficult to sample quantitatively, and these methods permit biomonitoring to be much more cost-effective and less of a financial impost on consent holders. I have no issue with HRC requiring quantitative sampling (5 replicates) and the use of the QMCI in most cases. In other cases, where quantitative monitoring might be ruled out due to cost, I think it would be better to monitor using kick-sampling than not at all.
- Power analysis is a tool for designing sampling programmes. I do not favour its use for analysing the results of a survey that has already been undertaken (unless it is to determine how many replicates should have been collected in order to undertake a better survey!).
- Equivalence tests (which seek to test whether two means differ by a pre-determined amount) are, in my view, one of the best (simple) ways for determining compliance with a rule that requires no more than a 20% decrease in QMCI (for example).
- On long-term biomonitoring programmes, trends testing should be used. By focusing on trends in river health separately at upstream (control) and downstream (impact) sites, one can see whether conditions are getting better or worse (or staying the same). Trends testing avoids the confounding effects associated with comparing one site with another, and allows, for example, one to track improvement in river health downstream of a discharge in response to management initiatives aimed at reducing the impact.
- I support the rule requiring no more than a 20% decrease in QMCI etc. It is practical, can be tested (equivalence testing), and will apply in most situations within the region based on the collection of 5 replicate Surber samples per site. Greater numbers of

replicates would be required where the control site is already grossly polluted (QMCI < 3.06 – see Table 1 above). In such situations, I think it would quite difficult for a discharge to cause a further 20% reduction without being extremely obvious in other ways (e.g., effluent quality monitoring).

- Irrespective of the use of robust methods and statistical analyses, determination of whether or not a significant decrease in river health downstream of a discharge is due entirely to the effects of the discharge (or only partly due to it) may still require the judgment of an experienced freshwater ecologist.
- In my view, biomonitoring programmes based upon comparison with reference condition can be more expensive than traditional approaches. Although references sites (if chosen appropriately) can provide guidance on the stream condition that could be expected in a given situation, they can also be inappropriate and provide unrealistic expectations of what can be achieved. I think there is plenty of scope for aiming to improve stream health at impact sites by comparing with upstream control sites, tracking trends, and evaluating ways to reduce concentrations of harmful substances in the discharges themselves.

Literature Cited

Duggan IC, Scarsbrook MR, Quinn JM 2003. Comparison of coded abundance and fixed count rapid assessment techniques for biomonitoring New Zealand streams. New Zealand Journal of Marine and Freshwater Research 37:23-29.

Karr JR, Chu EW 1999. *Restoring life in running waters: better biological monitoring*, Washington D.C.,Island Press. 206p.

Martinez-Abrain, A 2007. Are there any differences? A non-sensical question in ecology. *Acta Oecologica* 32:203-206.

Martinez-Abrain A 2008. Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica* 34: 9-11.

McBride GB 2005. Using statistical methods for water quality management – issues, problems and solutions. John Wiley & Sons Inc. New York. 313p.

Stark JD 1998. SQMCI: a biotic index for freshwater macroinvertebrate coded-abundance data. *New Zealand Journal of Marine and Freshwater Research* 32:55-66.

Stark JD 1999. Review of the AUSRIVAS trial in the Waikato. Cawthron Reports No. 506. 18p.

Stark JD 2008. Trends in river health of the Manawatu-Wanganui region 2008 with comments on the SoE biomonitoring programme. Stark Environmental Report No. 2008-07. 61p.

Stark JD 2009. Aquatic invertebrate communities of the Manawatu - Wanganui Region - 2009 State of the Environment Report. Stark Environmental Report No. 2009-11: 97p.

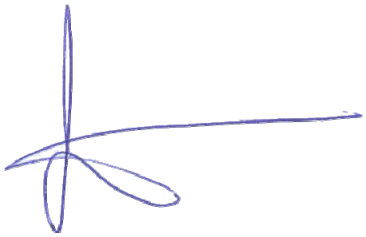
Stark JD 2010. Kapuni macroinvertebrate biomonitoring (18 January 2010). Stark Environmental Report No. 2010-01. 22p.

Stark JD, Fowles CR 2006. An approach to the evaluation of temporal trends in Taranaki state of the environment macroinvertebrate data. Cawthron Report No. 1135. 88p.

Stark JD, Maxted JR 2007. A user guide for the Macroinvertebrate Community Index. Cawthron Report. No. 1166: 58p.[http://www.mfe.govt.nz/publications/water/mci-user-guide-may07/mci-user-guide-may07.pdf]

Zar JH 1996. Biostatistical Analysis. 3rd ed. Prentice-Hall. Upper Saddle River, New Jersey.

I trust you find something useful in the above.



John Stark