5. DATA ANALYSIS

Statistics are a standard tool for the biologist working within the area of inventory or monitoring. Reliable data are necessary for managing natural resources and statistical inference allows a biologist to make informed decisions about a larger population based on a collection of representative samples. Despite the frequency of their use, many people forget that statistics are only a tool to manipulate data in an effort to tease out meaningful patterns. More specifically, it is not uncommon for educated people to adopt a statistical test result as the definitive answer to a question, without any consideration for the limitations and assumptions upon which this answer is based. Because of this, it is important that statistical analysis be well considered and appropriately applied to each project. Given the potentially misguided faith placed in statistical tests, a poor choice of test may produce an inappropriate management strategy.

The discussion which follows is meant to provide useful guidance for analysis of general species data at the presence/not detected, relative abundance, and absolute abundance levels of intensity. Species-specific data analysis will be discussed within species manuals. (Although initial versions of the CBCB species manuals focused more on data collection rather than analysis, data analysis is a high priority for subsequent versions.)

A small and final point: in the interest of maintaining a logical flow, the chapter on Data Analysis has been placed at the end this manual. However, although data analysis is performed after data are collected, the frameworks, assumptions, and constraints of various statistical tests must be considered well before this, during project planning.
5.1 Hypothesis Testing

When investigating population trends or comparing population estimates between areas, the traditional model for hypothesis testing should be followed:

1. Identify the parameters to be compared and the level of statistical significance.
2. Identify a null and alternative hypothesis(es).
3. Select an appropriate statistical procedure to test the null hypothesis.
4. Determine if the observed value of the statistic has a probability of occurrence less than a pre-chosen level of significance.
5. If it does, reject the null hypothesis.

This process results in a definitive decision as to whether to reject the null hypothesis. This decision carries with it the potential for two types of error. A Type I error occurs when the null hypothesis is rejected when it should not be. The probability of this occurring is designated alpha (\( \alpha \)) which should normally be set at 0.10 (but may vary depending on objectives), thus ensuring that if the difference is not significant, the risk of a mistaken conclusion is no more than 10%. The hazards associated with Type I error are generally well recognized and carefully controlled. A Type II error occurs when the null hypothesis is retained when it should be rejected; the probability of this occurring is designated beta (\( \beta \)). Traditionally \( \beta \) has been ignored, although the recent trend in ecological reporting is to give it much greater consideration. Typically \( \beta \) should be set at 0.20, but can also vary depending on objectives.

Generally, a conservative approach is recommended when setting the levels of \( \alpha \) or \( \beta \). Depending on how the hypothesis is worded, the risk of Type I or Type II error may be adjusted in an attempt to avoid irresponsible errors. For example, if a null hypothesis is stated:

\[
H_0: \text{There is no significant difference between the abundance of Species X in pristine and severely disturbed Study Areas.}
\]

The consequences of making a Type II error based on this type of hypothesis could result in potentially irresponsible management practices for Species X. It is very important that error be minimal when evaluating such a hypothesis. In such a case, a conservative approach would be to reduce the possibility of making such an error by reducing the level of \( \beta \), possibly below 0.20 if this was appropriate. The intent is not to manipulate statistics to produce a specific result, but rather, to ensure that experiments are designed to minimize the risk of making errors with potentially damaging impacts. It is wise to test hypotheses in a manner which is conservative with regard to management implications, and ecological sense should always take precedence over statistical testing.

Finally, it is important when selecting the appropriate statistical procedure to determine whether parametric or non-parametric methods should be used. Parametric methods of statistical analysis assume that the population has a normal distribution. Ecological variables often have skewed distributions which violate this assumption. Traditionally, non-parametric methods of data analysis have often been used as these do not require the population distribution to be normal and are not subject to restrictions in terms of variance or form (Krebs 1989). However, non-parametric methods have recently been criticized for having lower power, and hidden restrictive assumptions regarding the distribution of data. For example most non-parametric methods still require a symmetrical distribution and are not necessarily robust to highly skewed data. Many times parametric methods are as robust to non-normal distributions as non-parametric techniques (as long as N is large enough), and also have significantly higher power (See Day and Quinn 1989). Currently, generalized linear models (poisson regression, logistic regression, White...
and Bennetts (1996) negative binomial test) are being used as a method to accommodate non-normal distributions.