3-26-2003

Nonparametric assessment of safety levels in ecological risk assessment (ERA)

Limei Chen
Florida International University

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

NONPARAMETRIC ASSESSMENT OF SAFETY LEVELS

IN ECOLOGICAL RISK ASSESSMENT (ERA)

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Limei Chen

2003

To: Dean Arthur W. Herriott
    College of Arts and Sciences

This thesis, written by Limei Chen, and entitled Nonparametric Assessment of Safety Levels in Ecological Risk Assessment (ERA), having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

<div align="right">

Gary M. Rand

Samuel Shapiro

Hassan Zahedi

Ling Chen, Major Professor

</div>

Date of Defense: March 26, 2003

The thesis of Limei Chen is approved.

<div align="right">

Dean Arthur W. Herriott
College of Arts and Sciences

Dean Douglas Wartzok
University Graduate School

</div>

<div align="center">

Florida International University, 2003

</div>

DEDICATION

I dedicate this thesis to my parents. Without their patience, understanding, support, and love, the completion of this work would not have been possible.

ACKNOWLEDGMENTS

ABSTRACT OF THE THESIS

NONPARAMETRIC ASSESSMENT OF SAFETY LEVELS

IN ECOLOGICAL RISK ASSESSMENT (ERA)

by

Limei Chen

Florida International University, 2003

Miami, Florida

Professor Ling Chen, Major Professor

In ecological risk assessment (ERA), it is important to know whether the exposure that animal species receive from a chemical concentration exceeds the desired safety level. This study examined several statistical methods currently being used in ecological risk assessment and reviewed several statistical procedures related to this subject in the literature. Two large sample nonparametric tests were developed for this study. Monte Carlo study showed that these tests performed well even when the sample size was moderately large. A real data set was used to show that the new methodologies provide a good method for assessing the potential risks of pesticides residues at an investigated site.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

ARAMDG: aquatic risk assessment and mitigation dialogue group

CDF: cumulative density function

EPA: environmental protection agency

ERA: ecological risk assessment

HQ: hazard quotient

JPC: joint probability curves

LCL: lower confidence limit

NACA: national agricultural chemical association

NOECs: no effect concentrations

PAF: potentially affected fraction

PERA: probabilistic ecological risk assessment

PDF: probability density function

SETAC: society of environmental toxicology and chemistry

SSD: species sensitivity distribution

CHAPTER I

INTRODUCTION

This study examines the question of what is the more effective way to determine if the potentially affected fraction (PAF) of a species by a specific toxic substance exceeds 10% in an ecological risk assessment (ERA) study. In such a study, 10% is an established threshold which represents a significant risk. The research develops two nonparametric test procedures to assess if the 90th percentile of an exposure distribution is significantly less than the 10th percentile of a species sensitivity distribution.

An ERA consists of three phases: problem formulation, risk analysis, and risk characterization (U.S. EPA, 1998). The two main components of the risk analysis phase for an ERA are the exposure analysis and the effects analysis. The former looks at the magnitude and duration of an ecological entity or entities' exposure to a stressor while the latter examines the range and extent of possible effects from the stressor. The effects analysis portion for some modern ERAs have relied upon available toxicity data that can be statistically analyzed using one of several statistical distributions such as: the logistic, the normal, the lognormal or the triangular distributions. These "effect" distributions for effects are termed species sensitivity distributions (SSDs) and interpretations of risk have utilized benchmark values for exposure to extrapolate a PAF of species from a SSD. This procedure gives a characterization of potential risks by integrating results from a

distributional analysis of both the effects and exposure in an ERA.

A set of methods was developed for risk assessment of pesticides by a team of scientists, the aquatic risk assessment and mitigation dialogue group (ARAMDG), assembled by society of environmental toxicology and chemistry (SETAC) for the national agricultural chemical association (NACA) and the U.S.EPA (Baker *et al.*, 1994).

The ARAMDG showed that the quotient derived from the $p_1 100^{th}$ percentile of an exposure distribution and the $p_2 100^{th}$ percentile of an effects distribution may be a conservative tool for screening potential hazards, many published risk assessments in North America have utilized this comparison as a portion of their risk characterization (Campbell *et al.*, 2000; Giesy et al., 1999; Hall, Jr. *et al.*, 1998; Hall, Jr. *et al.*, 1999; Hall, Jr. et al., 2000; Klaine *et al.*, 1996; Maund *et al.*, 2001; Solomon *et al.*, 1996).

As discussed by ARAMDG, an ERA that relies solely on the protection of a certain percentage of species (e.g., 10%, 5%, etc.) may not be protective if species of significance are below the specified percentile on a SSD. In choosing a percentile from a distribution of acute or chronic effects, one makes the assumption that protecting a certain amount of species will be protective of the structure and function of an ecosystem and that the available single-species toxicity tests have relevance as representatives of an ecosystem to be protected or the "universe" of species in the environment when deriving quality criteria (Baker *et al.*, 1994). Maund *et al.* (2001) gave supporting evidence for using the tenth percentile of acute distributions based upon ecologically significant

effects observed at higher concentrations in field-type studies.

Some researchers prefer to use log-logistic probability density functions (PDFs) to model species sensitivity. The log-logistic PDFs may be preferred over log-normal PDFs due to their simplicity in calculations, particularly in the risk characterization for pesticide mixtures. The logistic distribution has a heavier tail than the lognormal distribution; hence it may allow a more conservative estimator of hazard concentrations for effects (Wagner & Løkke, 1991).

Data points in SSDs are taken from endpoints in acute or chronic toxicity tests. In acute distributions, data points are normally taken from tests used to derive LC/EC50s. For chronic toxicity distributions, no effect concentrations (NOECs) are commonly used in deriving SSDs. In acute toxicity tests, exposures to a contaminant are generally of short duration (e.g., 24 to 96 hours) while chronic toxicity tests are conducted over a full life cycle or early life stage of an organism. Responses measured for each of these types of tests that are included in SSDs are ones that can be extrapolated to the population level (i.e., survival, growth and reproduction) (ECOFRAM, 1999). Useful criteria for selecting proper test endpoints from ecotoxicity data are stated elsewhere.

Newman et al. (2000) have studied species sensitivity data for a large set of chemicals. They concluded that not all of the usually assumed distributions were supported by the data. Therefore, the use of nonparametric statistical methods in ERA study may be more appropriate.

CHAPTER II

LITERATURE REVIEW

2.1 Statistical Methodology for ERA

Risk estimates can be developed using one or more of the following techniques:

(i) the quotient approach (comparisons of single-point exposure and effects estimates);

(ii) the potentially affected fraction (PAF) approach;

(iii) the probabilistic ecological risk assessment (PERA).

When sufficient data are available to quantify exposure and effects estimates, the simplest and most widely used approach for comparing the estimates is a ratio called hazard quotient (HQ) (Urban and Cook, 1986; Calabrese and Baldwin 1993):

$$Hazard \approx \frac{Exposure\ concentration}{Effect\ concentration} \tag{2.1}$$

where exposure concentration can be an upper confidence limit of mean, an 90th percentile or the maximum concentration for conservatism based on different studies, and effect concentration can be 5th or 10th percentile of a species sensitivity distribution (SSD) depending on the regulation used in different countries(see Figure 1). If the HQ is less than or equal to one, it implies there is a negligible risk. Otherwise, it implies there is a potential risk.

The reciprocal of HQ is defined as:

$$Margin\ of\ Safety \approx \frac{Effect\ Concentration}{Exposure\ Concentration}.$$

(2.2)

Both exposure concentration and effect concentration have sampling distributions (see Figure 2). Some uncertainty measures must be incorporated into the single-point estimates to provide an evaluation of the likelihood that the effects point estimate exceeds the exposure point estimate. However in many studies, the sample sizes are not large enough to obtain good confidence intervals for the endpoints.

Figure 1. Comparison of point estimates
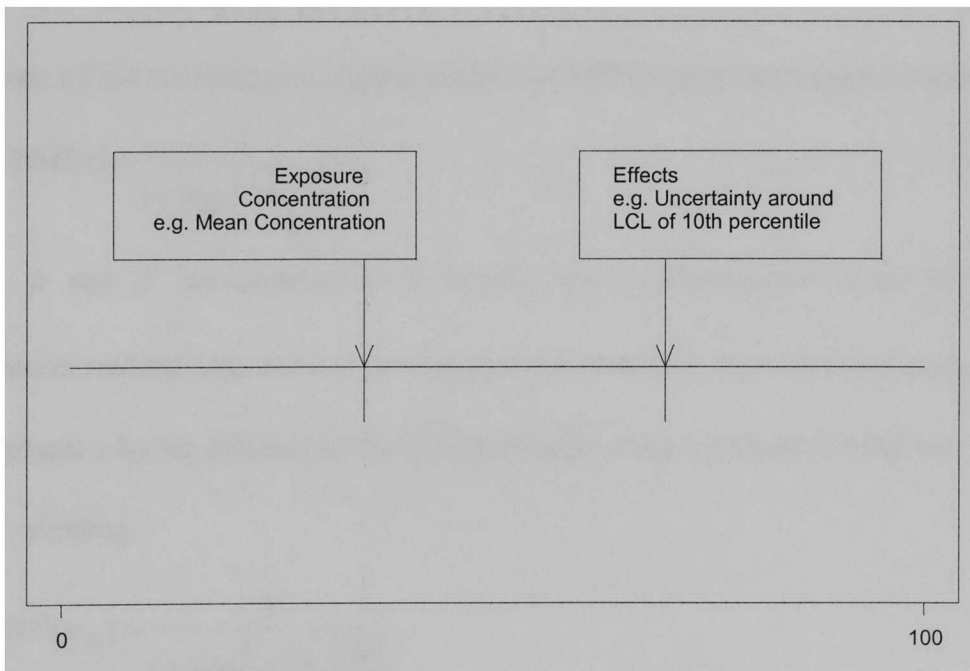
Although a certain ratio of environmental concentration to a target value gives a rough guide concerning risk, this ratio gives little information on environmental impacts other than a ratio <1 implies no or negligible damage. If two substances, however, have the same ratio >1, their environmental impacts may be quite different, and the sum of these ratio has no toxicological meaning.

An alternative method is to utilize benchmark values for the exposure and then extrapolate a potentially affected fraction (PAF) of species from a SSD (Klepper et al. 1997, 1998, 1999). This method may be an improvement over the current quotient methods, since it encompasses the often-observed nonlinearity of species sensitivity and, especially, it allows for aggregating risks over compounds in a mixture. If the log toxicity data from a SSD are fitted to a logistic model, the PAF of certain compound is given by

$$\text{PAF}(x) = \frac{1}{1 + \exp\{-\frac{x - \hat{\mu}}{\hat{\sigma}}\}}, \tag{2.3}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are estimates of the location and scale parameters in the log toxicity distribution respectively, and x is an exposure concentration. For assessing acute effects, one replaces x by the estimate of the 90th percentile of the exposure distribution, denoted by $y_{.90}$, yielding

$$PAF(y_{.90}) = \frac{1}{1 + \exp\left\{-\frac{y_{.90} - \hat{\mu}}{\hat{\sigma}}\right\}}. \tag{2.4}$$

Figure 2. Comparison of point estimation with associated uncertainty
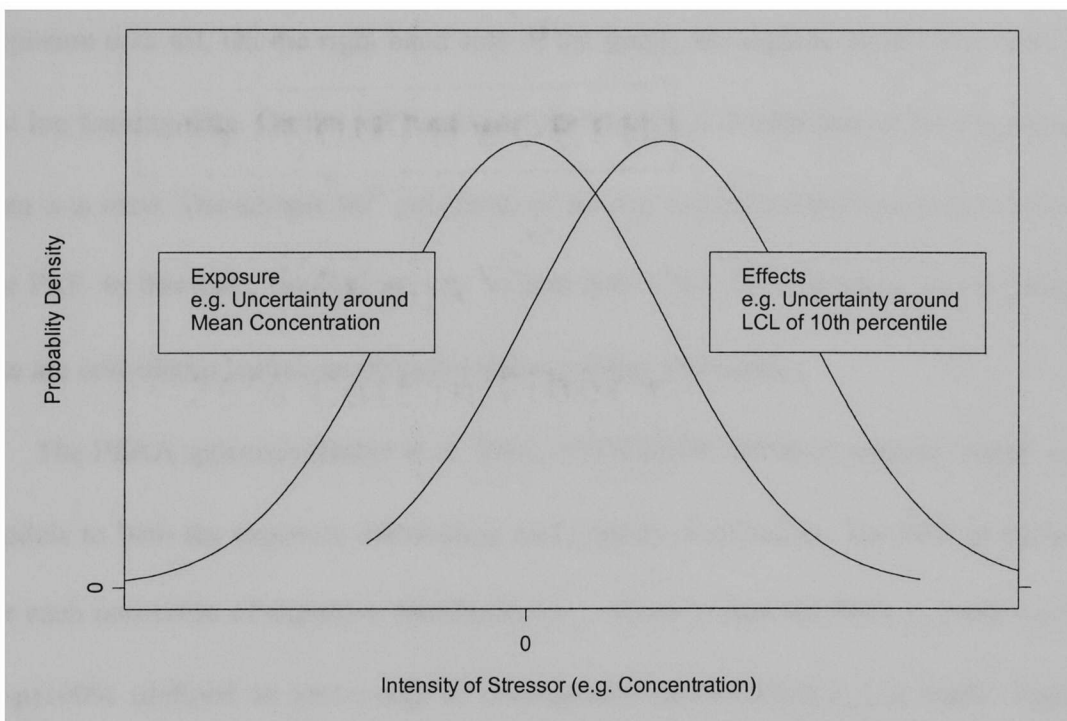
The effects analyses of some modern ERAs require the availability of toxicity data from SSDs. Equation (4) represents an integration of the results from an exposure analysis and an effects analysis in an ERA. It gives a characterization of potential risks. When compared to concentrations causing little or no effects in studies on communities in microcosm under field conditions, the 10th percentile of the toxicity distribution has proven to be a useful metric against which to compare exposure concentrations and is a conservative indicator of the threshold for effects (Solomon et al., 1996; Giesy et al., 1999; Versteeg et al., 1999). Thus, the PAF will be compared to the critical point, 10%. Figure 3 shows how to find a PAF for Atrazine. The points on the graph are

$(x_i, i/(n+1))$ and $(y_j, j/(m+1))$, where x_i is the ith ordered toxicity data and n is the size of the toxicity data set, and y_j is the jth ordered exposure data and m is the size of the exposure data set. On the right hand side of the graph, the logistic model was used to fit the log toxicity data. On the left hand side, the empirical distribution of the log exposure data was used. The sample 90th percentile of the log exposure data was used to calculate the PAF. In this case, the PAF at $y_{.90}$ is less than 10%. The species at the investigated site are considered having no potential risk regarding to Atrazine.

The PERA approach (Baker et al. 1994; ECOFRAM, 1999a,b) requires one to assign models to both the exposure distribution and toxicity distribution. The PAF is computed for each percentile of exposure distribution, y_p, which is denoted PAF(y_p) and a plot of (1-p)100% (defined as percentage of exceedence) against PAF(y_p) is made. Figure 4 illustrates such a plot. The curves shown on Figure 5 are called joint probability curves (JPC). The interpretation of the JPC is central to the PERA method. Each point on the curve represents both the probability that the chosen proportion of species will be affected and also the relative frequency with which the level of the effect would be exceeded. These probabilities are based on the current exposure data so at each point on the line, one can say "under current conditions, x% of species will be effected and that this proportion of species would be affected by y% of the current observations". (See Solomon *et al.* (2000)). JPC describes the probability of a particular set of exposure conditions occurring, relative to the number of taxa that would be expected to be affected.

Figure 3. The PAF approach for risk assessment: the logistic model was used to fit the log toxicity data   .
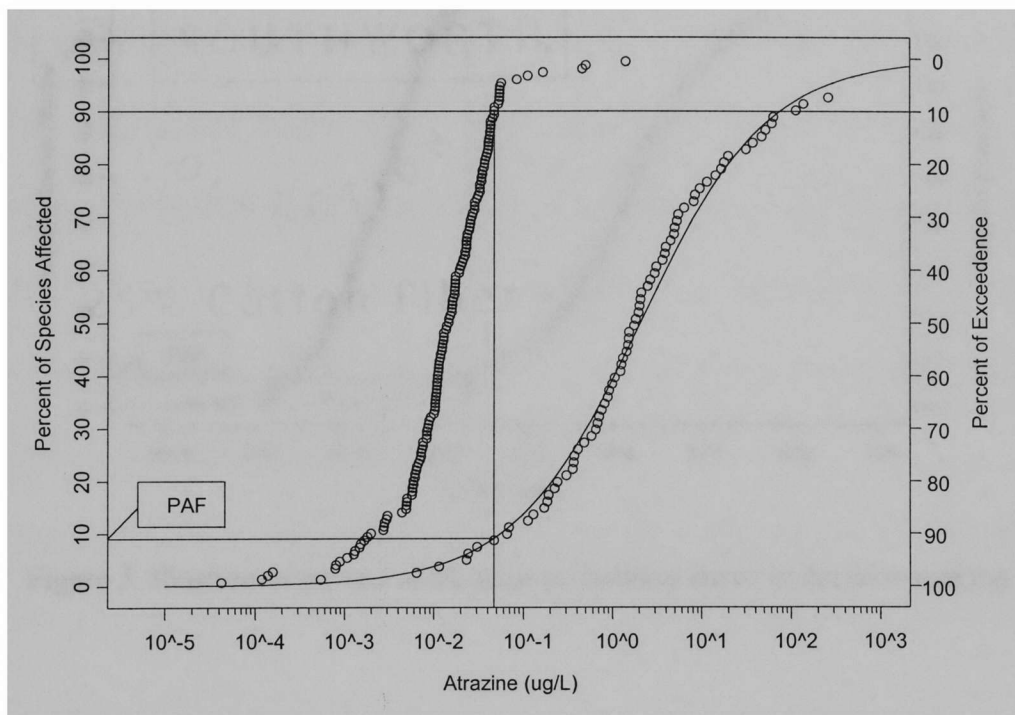
In this type of representation, the closer the joint probability curve is to the axes, the less is the probability of an adverse effect (Figure 5). The allowable exposure concentration can be adjusted until the appropriate level of protection has been achieved. The major advantage of PERA is that it uses all relevant single species toxicity data and, when combined with exposure distributions, allows quantitative estimations of risks. In addition, the data may be revisited, the decision criteria become more robust with additional data, and the method is transparent, producing the same results with the same data sets.

Figure 4. The PERA approach: the logistic model was used to fit the log toxicity data and the log exposure data
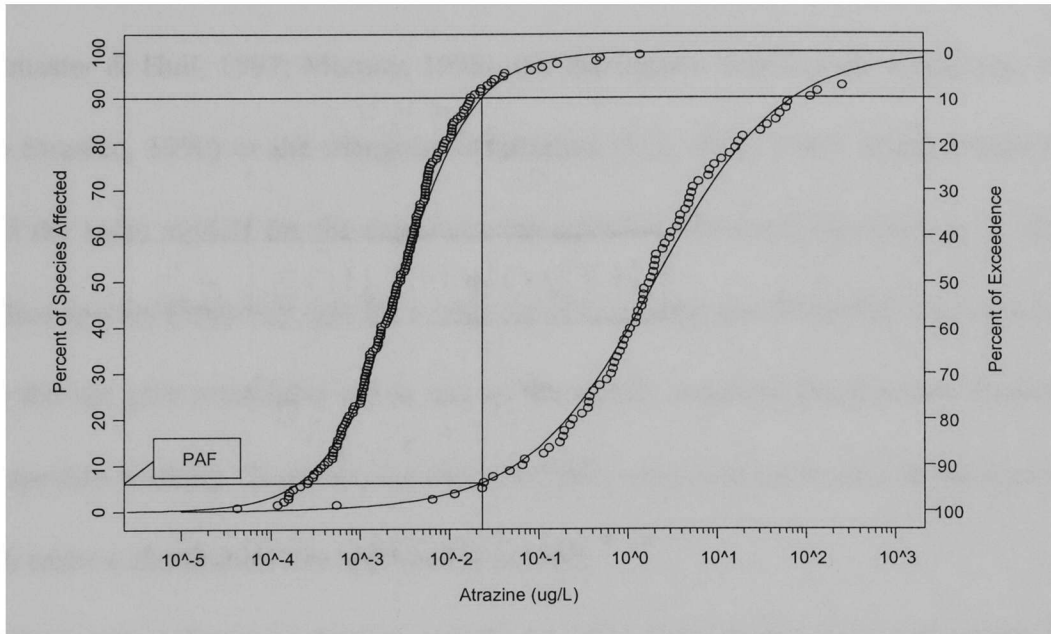
Figure 5. Illustion of the use of the joint probability curve in decision making
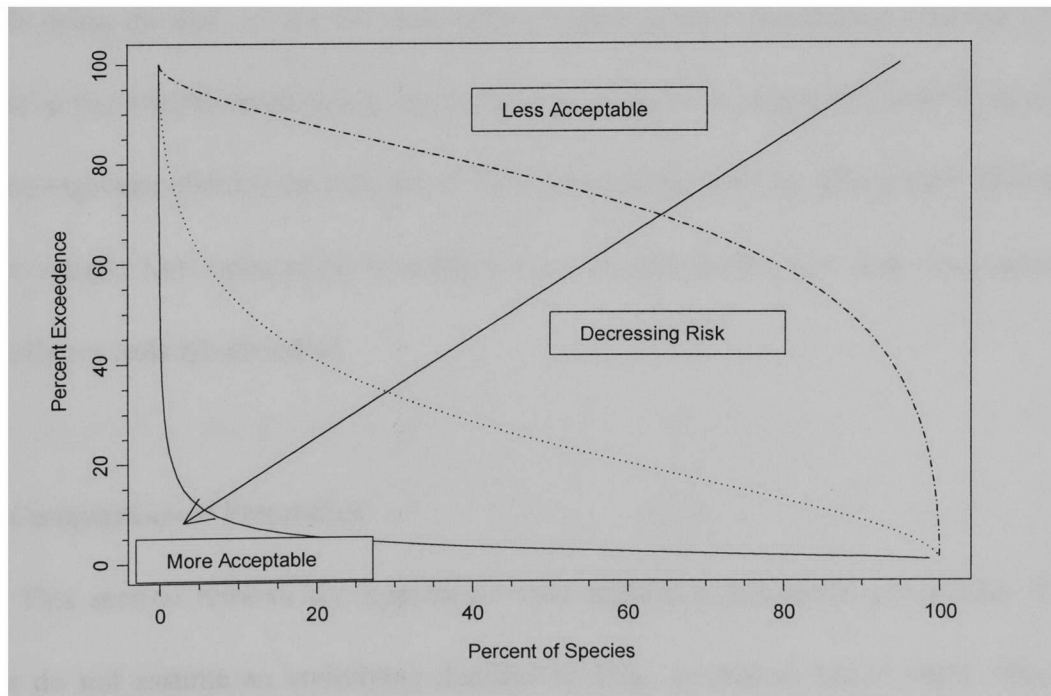
Both the PAF approach and the PERA approach are model related approaches. Most commonly used models for SSDs are the log-normal distribution (Paustenbach, 1995; Burmaster & Hull, 1997; Murphy, 1998), the log-logistic distribution (Kooijman, 1987; Van Straalen, 1990) or the triangular distribution (U.S. EPA, 1985). Many people have used the same models for the exposure concentration. However, Newman et al. (2000) studied species sensitivity data for a large set of chemicals and found that a lot of the data sets did not give reasonable fits to any of the usually assumed distributions. Especially for specific working chemicals, the observed SSD were bimodal or even multi-modal. In such cases a distribution free approach is needed.

Exposure and toxicity data are sample data. The uncertainty associated with the PAF was not taken into account in the risk assessments previously mentioned. Comparison made using the PAF of the 90th percentile of the exposure distribution with the critical point of the toxicity distribution, say 10%, is equivalent to comparing the 90th percentile of the exposure distribution with the 10th percentile of the toxicity distribution. Therefore, large sample tests concerning percentiles were needed. In the next part, three kinds of hypotheses tests are reviewed.

2.2 Comparisons of Percentiles

This section reviews the hypotheses tests regarding population percentiles. These tests do not assume an underlying distribution (e.g., normal or log-normal). They are

often more useful for comparing two population percentiles than those that assume a specific distribution because they make less stringent assumptions.

EPA (2000) indicated that any hypothesis about the pth percentile (ξ_p) can be converted to an equivalent hypothesis about the proportion (p) of the site below a threshold. Then the one sample test for percentile (H_0: the $p100^{th}$ percentile (ξ_p) is ζ_0 or larger $v.s.$ H_a: The $p100^{th}$ percentile (ξ_p) is smaller than ζ_0) can be equivalently written in terms of: $H_0 : p \geq p_0 \; v.s. H_a : p < p_0$. The only assumption required for the one-sample proportion test is the assumption of a random sample. To verify this assumption, one need only review the procedures and documentation used to select the sampling points and ascertain that proper randomization has been used. The test statistic is defined as:

$$z = \frac{\hat{p} - \dfrac{0.5}{n} - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{2.5}$$

where $\hat{p}$ denote the proportion of X's that do not exceed ζ_0 for given a random sample $X_1, X_2, ..., X_n$ of measurements from the population. i.e. $\hat{p} = k/n$, where the number (k) of sample points that are less than or equal to ζ_0, and n is the sample size.

If $z < z_\alpha$, $\Phi(z_\alpha) = \alpha$, the null hypothesis may be rejected and one concludes that the proportion is less than p_0, i.e. the $p100^{th}$ percentile is smaller than ζ_0.

Similarly, one might use the following test to compare two proportions or percentiles. Let p_1 represent the true proportion for population 1, and p_2 represent the true proportion of population 2. Let ξ_{p_1} be the p_1100^{th} percentile for population 1 and ζ_{p_2} be the

$p_2 100^{th}$ percentile for population 2. The hypothesis $H_0 : \xi_{p_1} \geq \zeta_{p_2}$ v.s. $H_a : \xi_{p_1} < \zeta_{p_2}$ can

be equivalently written in terms of $H_0 : p_1 \geq p_2$ v.s. $H_a : p_1 < p_2$. The test statistic is defined

as:

$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{\hat{p}(1-\hat{p})(1/m + 1/n)} \qquad (2.6)$$

where $\hat{p} = (k_1 + k_2)/(m+n)$, m is the sample size of the sample $X_1, X_2, ..., X_m$ from the

first population, k_1 is the number of points from the first sample which exceed some

threshold C. Likewise n is the sample size of the sample $Y_1, Y_2, ..., Y_n$ from the second

population, k_2 is the number of points from the second sample which exceed the same

threshold C. The sample proportions are $\hat{p}_1 = k_1 / m$ and $\hat{p}_2 = k_2 / n$.

If $z < z_\alpha$, the null hypothesis may be rejected and one concludes that the $p_1 100^{th}$

percentile for population one is smaller than the $p_2 100^{th}$ percentile for population two.

Rohatgi (1984) provided another one-sample test concerning the percentiles. A

quantile of order p of a random variable with distribution function F is a number ξ_p that

satisfies

$$P(X \leq \xi_p) \leq p \text{ and } P(X \geq \xi_p) \leq 1-p, \, 0 < p < 1 \qquad (2.7)$$

So that ξ_p is a solution of

$$p + P\{X = x\} \leq F(x) \leq p \qquad (2.8)$$

If $P(X = x) = 0$, as is the case, when X is a continuous random variable, then ξ_p is a

solution of the equation $F(x) = p$. Hence

$$P\left(X_{(r)} \le \xi_p\right) = \sum_{j=r}^{n} \binom{n}{j} p^j (1-p)^{n-j} \tag{2.9}$$

Thus one uses a test of the hypothesis: $H_0 : \xi_p \ge \zeta_0, versus\ H_a : \xi_p < \zeta_0$.

If the level of significance α is specified, then the largest r will be chosen such that

$$\alpha \ge P\left\{X_{(r)} < \zeta_0 \mid H_0\right\} = \sum_{j=r}^{n} \binom{n}{j} p^j (1-p)^{n-j} = 1 - \sum_{j=0}^{r-1} \binom{n}{j} p^j (1-p)^{n-j} \tag{2.10}$$

where $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$ are the order statistics for a random sample $X_1, X_2, ..., X_n$.

Then H_0 will be rejected if the observed value of $X_{(r)}$ is less than ζ_0.

Kosorok (1999) indicated that there are several promising nonparametric two-sample median comparison procedures for censored survival data. The earliest of these, proposed by Wang & Hettmansperger (1990), requires either a two-sample shift model be assumed or requires the estimation of the involved densities. To avoid density estimation, Su & Wei (1993) developed a minimum dispersion statistic based on the Kaplan-Meier estimator; see also Basawa & Koul (1988). The fact that Su & Wei's statistic is easily computed and asymptotically chi-squared is appealing, but their analytical approach cannot be directly applied to group sequential clinical trials with staggered patient entry. Keaney & Wei (1994) manage to solve this difficult problem by using an interesting extension of the resampling procedure of Parzen, Wei & Ying (1994). So Kosorok(1999) developed a multivariate two-sample quantile tests for equality of a given collection of quantiles which can be applied to a variety of empirical distribution functions, including both the Kaplan-Meier estimator, Turnbull's (1974) self-consistent survival estimator for

doubly-censored data and an estimator for repeated measured data.

Let F be the distribution function for the fist sample, and let G be the distribution function for the second sample. Assume $\hat{F}$ and $\hat{G}$ are the usual empirical distribution function estimators for independent and identically distributed observations. Let m be the sample size associated with $\hat{F}$, and let n be the sample size associated with $\hat{G}$. If $F, \hat{F}_j, G$ and $\hat{G}_j$ satisfy the following conditions:

(1) Distribution function F has a density f in a neighborhood of t such that f is continuous at t and $0 < f(t) < \infty$.

(2) As $n \to \infty, c_n \left\{ \hat{F}(t) - F(t) \right\}$ converges in distribution to a bounded random variable with continuous distribution function.

(3) For every $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} pr \left\{ \sup_{s:|t-s|<\delta} c_n \mid \hat{F}(t) - F(t) - \hat{F}(s) + F(s) \mid > \varepsilon \right\} = 0$$

Then to test $H_0 : F^{-1}(p_j) = G^{-1}(p_j)(j = 1, 2, ..., J)$.

$$\hat{f}_j \equiv \int_{\Re} n_1^{1/5} \hat{Q}_F^{-1} K \left\{ \frac{\hat{F}_j^{-1}(p_j) - x}{n_1^{1/5} \hat{Q}_F^{-1}} \right\} d\hat{F}(x) \tag{2.11}$$

$$\hat{g}_j \equiv \int_{\Re} n_1^{1/5} \hat{Q}_G^{-1} K \left\{ \frac{\hat{G}_j^{-1}(p_j) - x}{n_1^{1/5} \hat{Q}_G^{-1}} \right\} d\hat{G}(x) \tag{2.12}$$

where $\hat{Q}_F$ and $\hat{Q}_G$ are twice the estimated interquartile ranges of F and G, respectively and where the kernel is triangular:

$$K(x) = \begin{cases} x+1 & \text{if } x \in [-1, 0], \\ 1-x & \text{if } x \in (0, 1], \\ 0 & \text{if } |x| > 1. \end{cases} \qquad (2.13)$$

To estimate the densities, either $O_p\left(n^{-1/2}\right)$ window estimators or optimal-order estimators such as the $O_p\left(n^{-1/5}\right)$ window estimators can be used, then an asymptotically chi-squared statistic with J degrees of freedom for testing H_0.

Although the above tests are for percentiles, the two involved percentiles are for the same percentage. The ERA study needs the tests for the difference between $100p_1$ percentile of a species sensitivity distribution and the $100p_2$ percentile of an exposure distribution ($p_1 \neq p_2$).

CHAPTER III

LARGE SAMPLE TESTS FOR THE DIFFERENCE

BETWEEN TWO PERCENTILES:

In this section, the following two cases will be considered.

(1) The critical point is assumed for a risk assessment, that is, either one assumes that the true SSD to a toxicant of interest is known or that the 10th percentile of the distribution (or the 5th percentile used in most European countries) has been estimated conservatively [for example, use of the lower confidence limit (LCL) proposed by Van der Hoeven (2001) or the minimum risk estimator proposed by Chen (2003)]. In such a case, the sample exposure data are taken from the investigated site. The test of interest is H_0: the exposure concentration is greater than or equal to the critical point $v.s.$ H_a: the exposure concentration is less than the critical point. The exposure concentration is defined as the p100th percentile of the exposure distribution, where $0 < p < 1$. Commonly used percentiles in an exposure distribution in risk assessment are the median or the 90th percentile.

(2) It is assumed that for a certain compound both the true p_1 100th percentile of the exposure distribution and the p_2 100th percentile of the true SSD to the toxic concentration (a critical point) are unknown, where $0 < p_i < 1$, $i = 1, 2$. The test of interest is H_0: the p_1 100th percentile of the exposure distribution is greater than or equal to the

p_2 100th percentile of the SSD to the toxic concentration *v.s.* H_a : the p_1 100th percentile

of the exposure distribution is less than the p_2 100th percentile of the SSD to the toxic

concentration.

3.1 One Sample Test for Testing the Exposure Concentration is Less than the Desired Safety Concentration Level

Ecological risk assessment requires estimation of the ambient concentration of a

compound at a contaminated site or the concentration predicted to result from a proposed

use. For the derivation of environmental quality criteria, a cutoff p100% is chosen, where

$0 < p < 1$ (to protect $(1 - p)$ 100% of species), and desired "safe" concentration is defined

as HC_P, the p100th percentile of the SSD. If a threshold for significant risk has been

identified by policy (e.g., effects on more than 5% or 10% of species are unacceptable),

any concentration higher than the desired "safe" concentration $HC_{0.05}$ or $HC_{0.10}$ can be

considered to pose a significant risk. If the variance or uncertainty is estimated, the risk

may be defined as the probability of harming more than 5% or 10% of the species based

on a specified application.

Van der Hoeven (2001) proposed a nonparametric method for estimating the 5th

percentile of a SSD. The article showed that this proved methodology performed better

than using the 95% LCL based on a log-logistic model or a lognormal model proposed by

Van Straalen and Denneman (1989) and Wagner and Lökke (1991) respectively in the

sense of having the same confidence level but being less conservative.

Chen (2003) proposed a minimum risk estimator for the 5th percentile of SSD based on an asymmetric loss function. The parameters in the asymmetric loss function can vary based on the degree of the conservatism.

Suppose a threshold for significant risk is defined by $\zeta_{.05} \left(= HC_{0.05} \right)$, and the critical point of the risk assessment has been chosen by using either Van der Hoeven's LCL or Chen's method for $HC_{0.05}$ as $\zeta_0 \left(= \hat{\zeta}_{.05} \right)$. To determine if a particular chemical concentration at a site would affect more than 5% of a species, a test $H_0 : \xi_p \geq \zeta_0$ versus $H_a : \xi_p < \zeta_0$ is desired, where ξ_p is the p100th percentile of the exposure distribution. The proposed test statistic is as follows:

$$Z_1 = \frac{\sqrt{n}\left(\hat{\xi}_p - \zeta_0\right)}{\sqrt{p(1-p)/\hat{f}^2\left(\hat{\xi}_p\right)}} \tag{3.1}$$

where $\hat{\xi}_p = X_{(k)} + \left\{ p(n+1) - k \right\}\left(X_{(k+1)} - X_{(k)} \right),$ (3.2)

where $k = \left[p(n+1) \right]$, and [w] stands for the integral part of w, $X_{(r)}$ denotes the rth order statistic, and $\hat{f}$ is a consistent estimator of the PDF of the exposure distribution. According to following known result (Theorem 2.1 stated below, see example Rohatgi & Saleh, 2001), asymptotically, the test statistic Z_1 has a standard normal distribution under certain assumptions.

Theorem 3.1. Let $f(x)$ and $F(x)$ be PDF and CDF of X respectively, and $X_1, X_2, ..., X_n$ be a simple random sample from $F(x)$. Assume that

1. for a fixed p $(0 < p < 1)$, $F(\xi_p) = p$;

2. $f(x)$ is continuous at $x = \xi_p$ and $f(\xi_p) > 0$, then

$$\sqrt{n}(\hat{\xi}_p - \xi_p) \xrightarrow{L} N\left(0, \sigma_p^2\right) \tag{3.3}$$

where $\sigma_p^2 = p(1-p)/f^2(\xi_p)$ and

$$\hat{\xi}_p = X_{([(n+1)p])} + \left\{ p(n+1) - \left[(n+1)p\right] \right\} \left(X_{([(n+1)p]+1)} - X_{([(n+1)p])} \right),$$

where [w] stands for the integer part of w and X_r denotes the rth order statistic.

It is easy to see that under assumptions in Theorem 3.1 and H_0, we have

$$\frac{\sqrt{n}\left(\hat{\xi}_p - \zeta_0\right)}{\sqrt{p(1-p)/f^2(\xi_p)}} \xrightarrow{L} Z, \, as \, n \to \infty, \tag{3.4}$$

where Z is the standard normal random variable.

Notice that

$$\frac{\sqrt{n}\left(\hat{\xi}_p - \zeta_0\right)}{\sqrt{p(1-p)/\hat{f}^2(\hat{\xi}_p)}} = \frac{\sqrt{n}\left(\hat{\xi}_p - \zeta_0\right)}{\sqrt{p(1-p)/f^2(\xi_p)}} \cdot \frac{\hat{f}(\hat{\xi}_p)}{f(\xi_p)}. \tag{3.5}$$

Obviously, if one chooses a density estimator $\hat{f}(\hat{\xi}_p)$ such that $\hat{f}(\hat{\xi}_p) \xrightarrow{P} f(\xi_p)$, as $n \to \infty$. then $Z_1 \xrightarrow{L} Z, \, as \, n \to \infty$. The choice of the density estimator will be discussed in Section 3.3.

3.2 Two Sample Test for Testing the Exposure Concentrate is Less than Desired Safety Concentration Level

In practice, the true SSD is unknown. Toxicity data is based on a sample from the SSD. Instead of using Van der Hoeven's method or Chen's method to find a critical point for the one sample test proposed in Section 3.1, one can test $H_0 : \xi_{p_1} \geq \zeta_{p_2}$ v.s. $H_a : \xi_{p_1} < \zeta_{p_2}$, where ξ_{p_1} and ζ_{p_2} are the p_1 100th percentile of the exposure distribution and the p_2 100th percentile of the SSD, respectively. The proposed test statistic is as follows:

$$Z_2 = \frac{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}{s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}} \tag{3.6}$$

where

$$s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}} = \sqrt{\frac{p_1(1-p_1)}{n\hat{f}_x^2(\hat{\xi}_{p_1})} + \frac{p_2(1-p_2)}{m\hat{f}_y^2(\hat{\zeta}_{p_2})}}. \tag{3.7}$$

where n and m are sample sizes, and $\hat{f}_x$ and $\hat{f}_y$ are the density estimators of an exposure distribution and a SSD (or toxicity distribution), respectively. Since samples from an exposure distribution and a SSD are independent, applying Theorem 3.1 and assuming H_0 is true,

$$\frac{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}{\sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}} \xrightarrow{L} Z, \text{ as } n, m \to \infty, \tag{3.8}$$

where

$$\sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}} = \sqrt{\frac{p_1(1-p_1)}{nf_x^2(\xi_{p_1})} + \frac{p_2(1-p_2)}{mf_y^2(\zeta_{p_2})}}. \tag{3.9}$$

Notice that

$$Z_2 = \frac{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}{\sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}} \cdot \frac{\sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}}{s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}}. \tag{3.10}$$

If $s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}} \xrightarrow{p} \sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}$ as $n, m \to \infty$, $Z_2 \xrightarrow{L} Z$, as $n, m \to \infty$.

In fact, under certain condition, $\hat{f}_x(\hat{\xi}_{p_1})$ and $\hat{f}_y(\hat{\zeta}_{p_2})$ are strongly consistent estimators of $f_x(\xi_{p_1})$ and $f_y(\zeta_{p_2})$. Hence, $s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}} \xrightarrow{a.s} \sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}$ as $n, m \to \infty$. It implies

$s_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}} \xrightarrow{p} \sigma_{\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}}$ as $n, m \to \infty$.

3.3 Choice of the density estimators

Let $k(u)$ be a kernel function defined on the real line R such that

$(i) \sup_{u \in R} |k(u)| < \infty,$

$(ii) \lim_{u \to \pm\infty} |uk(u)| = 0, and$

$(iii) \int_{-\infty}^{\infty} k(u)du = 1.$

And let $\{h_n\}$ be a sequence of nonnegative constants satisfy $h_n = O(n^{-\delta}), 0 < \delta < \frac{1}{4}$. A kernel estimator of $f(x)$ for a given x can be defined as

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right). \tag{3.11}$$

This estimator was proposed bye Rosenblatt (1956) and later studied by Parzen (1962) and Nadaraya(1965), among others.

Lin & Wu (1980) proven $\hat{f}(\hat{\xi}_p)$ is a strongly consistent estimator for $f(\xi_p)$.

Theorem 3.2. Let $X_1, X_2, ..., X_n$ be a simple random sample from F. Assume that

(A.1) $F'(\xi_p) = f(\xi_p) > 0$, and

(A.2) F'' exists and is bounded in a neighborhood N_p of ξ_p

Let k(u) be a kernel function of bounded variation and assume that for any $\varepsilon > 0$,

$\sum_{n=1}^{\infty} \exp(-\varepsilon n h_n^2) < \infty$. Then for any p, 0<p<1,

$$\hat{f}(\hat{\xi}_p) \xrightarrow{a.s.} f(\xi_p) \text{ as } n \to \infty.$$

In practice, we recommend to use the following kernel density estimator:

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right). \tag{3.12}$$

where $h_n = O(n^{-\delta})$ and $K(u) = \frac{3}{20\sqrt{5}}(5 - u^2), when |u| \le a, otherwise K(u) = 0$

Another choice of a density estimator is the k-NN density estimator. Let k_n be an integer between 1 and n. Given any x, Let

$$a_n(x) = a_n(x; x_1, ..., x_n)$$
$$= \min\{a \mid \#k_n \text{ of } x_1, ..., x_n \in [x - a, x + a]\}$$

then the k-NN density estimator is defined as

$$\hat{f}_n(x) = k_n / (2na_n(x)). \tag{3.13}$$

By using the strong consistency of k-NN density estimator proved by Wagner (1973), we have

Theorem 3.3. Let $X_1, X_2, ..., X_n$ be a simple random sample from F with density f satisfying assumptions (A.1) and (A.2) in theorem 2, Assume k_n satisfies

$(1) k_n / n \to 0$

$(2) \sum e^{-ck_n} < \infty, \text{ for any } c > 0$

Then, for any p, 0<p<1,

$$\hat{f}\left(\hat{\xi}_p\right) \xrightarrow{a.s.} f\left(\xi_p\right) \text{ as } n \to \infty.$$

Proof: It is clear that

$$|\hat{f}\left(\hat{\xi}_p\right) - f\left(\xi_p\right)| \leq |\hat{f}\left(\hat{\xi}_p\right) - f\left(\hat{\xi}_p\right)| + |f\left(\hat{\xi}_p\right) - f\left(\xi_p\right)|$$

$$\leq \sup_{x \in R} |\hat{f}(x) - f(x)| + |\hat{\xi}_p - \xi_p| \cdot \sup_{X \in N_p} |f'(x)|$$

$$= A + B, say.$$

As $n \to \infty, A \xrightarrow{a.s.} 0$, by Wagner (1973), and $B \xrightarrow{a.s.} 0$, since $\hat{\xi}_p \xrightarrow{a.s.} \xi_p$ and

(A.2) is assumed.

Note that an important case for satisfying (2) in Theorem 3.3 is

$\lim_{n \to \infty} \log(n) / k_n = 0.$ therefore, one can choose a k-NN density estimator with $k_n = \sqrt{n}$.

The theoretical background of these choices can be found in Chen (1984).

3.4 Power of the tests

Let $f(x)$ be the *pdf* of an exposure distribution and $d = \xi_p - \zeta_0$. Then

$$Power_{z_1} = P_{H_a}\left(Z_1 < z_\alpha\right) = P\left(Z < z_\alpha - \frac{\sqrt{n} \cdot d \cdot \hat{f}\left(\hat{\xi}_p\right)}{\sqrt{p(1-p)}}\right)$$

$$= P\left(Z < z_\alpha - \frac{\sqrt{n} \cdot d \cdot f\left(\xi_p\right)}{\sqrt{p(1-p)}} \cdot \frac{\hat{f}\left(\hat{\xi}_p\right)}{f\left(\xi_p\right)}\right),$$

(3.14)

where z_α satisfies $\Phi(z_\alpha) = \alpha$, and $\Phi(\cdot)$ is *cdf* of the standard normal random variable.

If one chooses one of the density estimators recommended in Section 3.3,

$\dfrac{\hat{f}(\hat{\xi}_p)}{f(\xi_p)} \xrightarrow{a.s.} 1$, as $n \to \infty$. Hence, when n is large,

$$Power_{z_1} \approx \Phi\left(z_\alpha - \frac{\sqrt{n} \cdot d \cdot f(\xi_p)}{\sqrt{p(1-p)}} \right) \qquad (3.15)$$

It is easy to see that the power of the Z_1 test is not only related to p, the difference between ξ_p and ζ_0, and the sample size n but is also related to the value of the density function at ξ_p. The logistic distribution is a commonly used model in ecological risk assessment.

Figure 6. The approximate power of the Z_1 test when the logistic

model is used (p=0.9, $\mu = 2.0$, $\sigma = 1.5$)   .

25
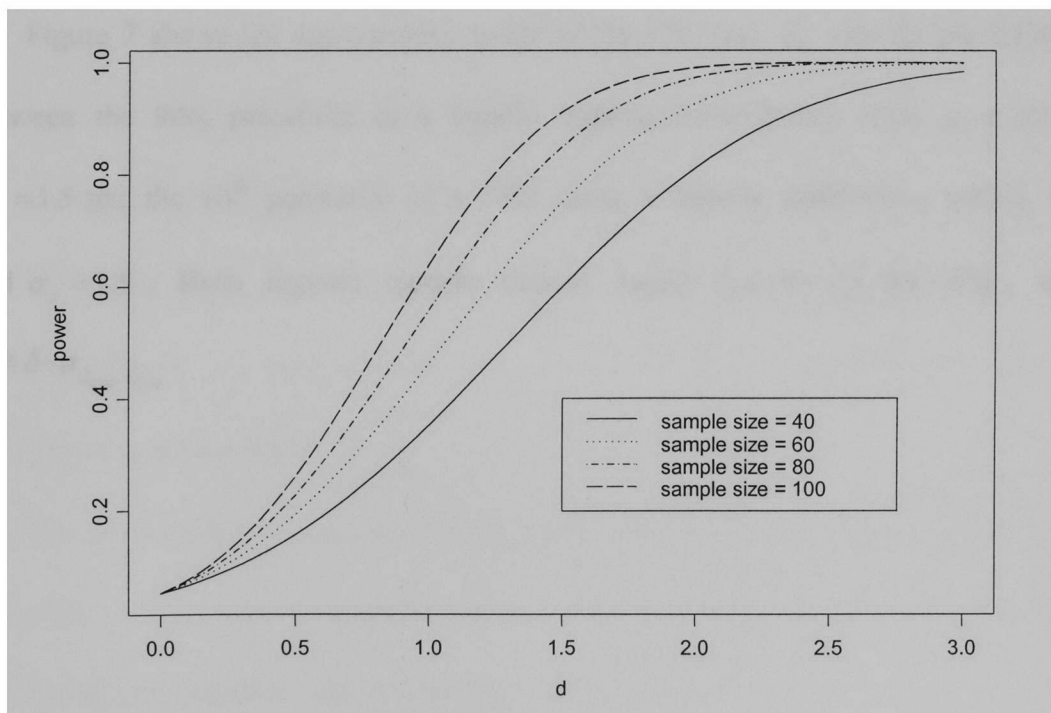
Figure 6 is power curves for the Z_1 test with $\alpha = 0.05$ for varying values of $d = \delta \cdot \sigma_{\hat{\xi}_p}$ and 4 different sample sizes. The curves are for a test of $\xi_{0.90}$ based on a logistic model with $\mu = 2.0$ and $\sigma = 1.5$. This graph yields the power property of the new nonparametric test if samples are truly from a logistic distribution.

For the power of Z_2 test, using the notation in Section 3.2, define $d = \xi_{p_1} - \zeta_{p_2}$. Similar to the derivation of the power of the Z_1 test, when both n and m are large,

$$Power_{z_2} \approx \Phi\left(z_\alpha - \frac{d}{\sqrt{p_1(1-p_1)/\left[n \cdot f_x^2\left(\xi_{p_1}\right)\right] + p_2(1-p_2)/\left[m \cdot f_y^2\left(\zeta_{p_2}\right)\right]}} \right) \qquad (3.16)$$

Assuming that $p_2 = 1 - p_1$ and $f_x(x) = f_y(-(x-d))$, then $f_y\left(\zeta_{p_2}\right) = f_x\left(\xi_{p_1}\right)$ and

$$Power_{z_2} \approx \Phi\left(z_\alpha - \frac{d \cdot f_x\left(\xi_{p_1}\right)}{\sqrt{p_1(1-p_1)(1/n + 1/m)}} \right). \qquad (3.17)$$

Figure 7 shows the approximate power of the 5% level Z_2 test for the difference between the 90th percentile in a logistic exposure distribution with $\mu_x = 2.0$ and $\sigma_x = 1.5$ and the 10th percentile of a SSD using a logistic distribution with $\mu_y = 5.0$ and $\sigma_y = 1.5$. Both logistic models chosen satisfy $f_x(x) = f_y(-(x-d))$, where $d = \delta \cdot \sigma_{\hat{\xi}_{0.90} - \hat{\zeta}_{0.10}}$.

Figure 7. The approximate power of the Z_2 test when the logistic model is used for F_x and F_y ($\sigma_x = \sigma_y = 1.5$)
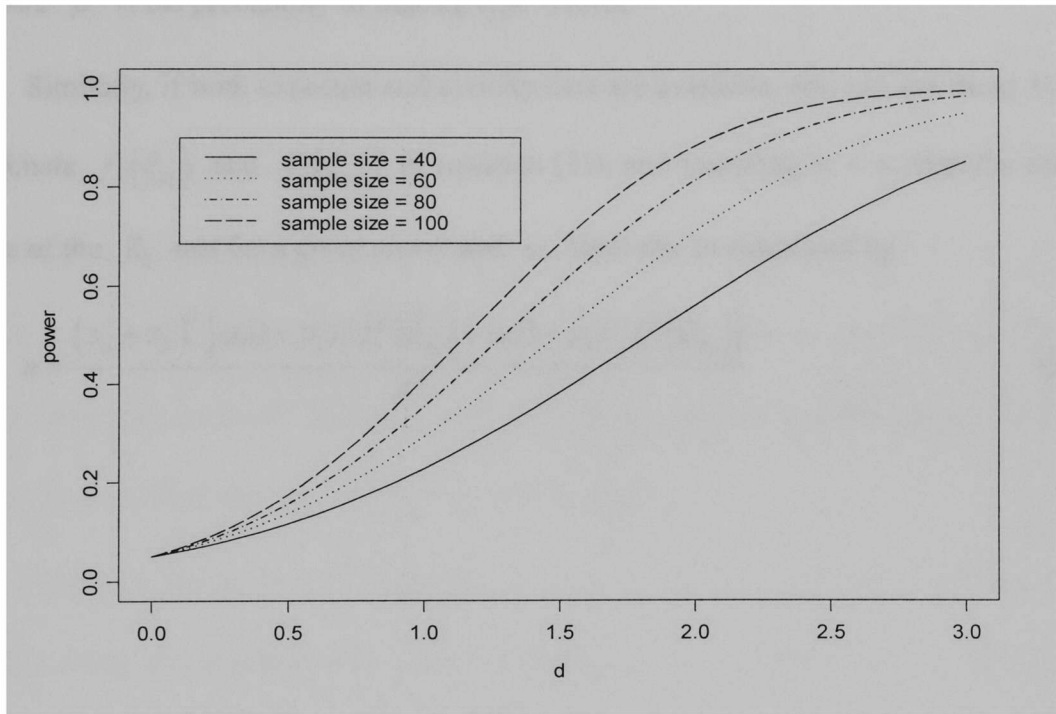
3.5 Sample Size Calculation

Under the assumptions for F_x and F_y in Section 3.3, the approximate powers of both the Z_1 and Z_2 tests are related to the unknown density functions used to model the exposure and the SSD.

When there are exposure data available from previous studies, one can use these data to estimate $f_x(\xi_p)$ in equation (10), then the sample size to yield a given power for the Z_1 test of a specified α can be calculated as follows:

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 p(1-p)}{d^2 \hat{f}_x^2\left(\hat{\xi}_p\right)} \tag{3.18}$$

where β is the probability of making type II error.

Similarly, if both exposure and toxicity data are available, one can use these data to estimate $f_x\left(\xi_{p_1}\right)$ and $f_y\left(\zeta_{p_2}\right)$ in equation (11), and assuming $m = n$; then the sample size of the Z_2 test for a given power and α level can be calculated by

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \left\{p_1(1-p_1)/\hat{f}_x^2\left(\hat{\xi}_{p_1}\right) + p_2(1-p_2)/\hat{f}_y^2\left(\hat{\zeta}_{p_2}\right)\right\}}{d^2} \tag{3.19}$$

CHAPTER IV

RESULTS

4.1 Results of Monte Carlo Study:

The tests developed in section 3 are large sample tests. To ensure the test can also be used for moderately large sample cases, a Monte Carlo study was conducted. The log-normal and the log-logistic distributions were used, since they have been previous used in ERA. For each simulation condition, random samples of sample size n=40, 60, 100 were generated via FORTRAN. To reduce the simulation error, the Monte Carlo sample size of 100,000 was chosen in the simulation.

Based on the study of evaluating the potential risks of pesticides residues in the C-111 canal, the means for the log toxicity data have about "2.00" (i.e. $\mu_2 = 2$) and the standard deviation was approximately "1.5" (i.e. $\sigma_2 = 1.5$). So in the simulations, the random data were generated from distributions with $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1.5$.

Silverman (1986) showed the ideal window width (h_{opt}) can be obtained by simple calculus:

$$h_{opt} = k_2^{-2/5} \left\{ \int K(t)^2 \, dt \right\}^{1/5} \left\{ \int f''(x)^2 \, dx \right\}^{-1/5} n^{-1/5} \qquad (4.1)$$

For the normal distribution with a variance σ^2, using Epanechnikov kernel, the window width obtained from (15) would be $h_{opt} \approx 1.05\sigma n^{-1/5}$ (4.2). For the logistic distribution with shape parameter σ, using Epanechnikov kernel, the window width obtained from (15) would be $h_{opt} \approx 1.62\sigma^{2/5} n^{-1/5}$ (4.3).

Table 1

Power of Two Samples' Test Based on Data from Log-Logistic Distributions using Kernel Density Estimate (where $\kappa *$ indicates how many times of the standard deviations for the test statistic)

alpha=0.05					
$\kappa *$	0	0.71	1.06	2.12	2.83
n=40	0.0597	0.1734	0.2591	0.5569	0.7322
n=60	0.0604	0.1766	0.2647	0.5824	0.7636
n=100	0.0595	0.1812	0.2734	0.6105	0.7959
alpha=0.10					
$\kappa *$	0	0.71	1.06	2.12	2.83
n=40	0.0949	0.2484	0.3493	0.6618	0.8174
n=60	0.098	0.2554	0.3623	0.6903	0.8461
n=100	0.0997	0.2662	0.3774	0.7212	0.8733

Table 2

Power of Two Samples' Test Based on Data from Log-Normal Distributions using Kernel Density Estimate (where $\kappa *$ indicates how many times of the standard deviations for the test statistic)

alpha=0.05					
$\kappa *$	0	0.71	1.06	2.12	2.83
n=40	0.0526	0.1617	0.2479	0.5716	0.766
n=60	0.055	0.1676	0.2577	0.5978	0.7924
n=100	0.0549	0.1733	0.2669	0.6216	0.8176
alpha=0.10					
$\kappa *$	0	0.71	1.06	2.12	2.83
n=40	0.0896	0.2409	0.3478	0.6873	0.8519
n=60	0.0931	0.2516	0.3613	0.713	0.874
n=100	0.0966	0.2619	0.3775	0.7375	0.8936

The results in this paper are only based on the data generated from log-normal distributions and log-logistic distributions. Type I errors and power simulation results are summarized in Table 1 and Table 2.

When α =0.10, the test appears to give good results. When the distance between two percentiles is equal to 0, the simulation type I error is very close to the nominal α, and when the distance of the two true percentiles is equal to 2.83 standard deviations, the power of the test is about 0.8~0.9. However, when α =0.05, there is slightly inflated type I error. Also notice that the power changed little when the sample sizes increasing. The deficiency might be due to the density estimator used in the test statistic. One way to improve these tests is to find a better estimator for the variance for the sample percentile. This is under further investigation.

4.2. An Example

In the study for evaluating the potential risks of pesticides residues in Biscayne Bay and Florida Bay areas, Acute (LC/EC50) laboratory toxicity data for water exposures from Atrazine were collected and analyzed. All acute endpoints were derived from the AQUIRE database. As discussed by ARAMDG (Baker et al., 1994), an ERA that relies solely on the protection of a certain percentage of species (e.g., 10%, 5%, etc.) may not be protective if species of significance are below the specified percentile of a SSD. And Maund et al. (2001) gave supporting evidence for using the 10th percentile of acute

31

distributions based upon ecologically significant effects observed at higher concentrations in field-type studies.

This laboratory Atrazine data had 72 observations from 0.97 ug/L to 76000 ug/L. After taking a log10 transformation, the data ranged -0.0132 to 4.8808. Using a = 1:0 and b = 1:5 in Chen's method (2003), the minimum risk estimate of the 10th percentile of the SSD is 3.1163. Suppose this estimate is chosen as a critical point in this risk assessment. Water exposure data for Atrazine were taken from one of the investigated sites. A sample of size 62 was available and the estimated 90th percentile of log10 transformed data was 1.9481. Let $\xi_{.90}$ be the true 90th percentile of the exposure distribution. In order to decide if there is Atrazine residue risk in the fresh water at this site, a test of $H_0 : \xi_{.90} \geq 3.1163$ versus $H_a : \xi_{.90} < 3.1163$,at $\alpha = 0.05$ was used. $\xi_{.90}$ in this case equals 1.9481 and $\hat{f}(1.9481) = 0.1503$. The observed test statistic is $Z_1 = -4.6073$. Since the p-value of the test is less than 0.0001, we reject H_0 and conclude there will be no more than 10% species potentially affected by Atrazine residues at this site.

Using this information and the formula in equation (13), the sample size needed for a 5% level Z_1 test with an approximate power greater than 0.70 and $d = 2\hat{\sigma}_{\xi_p}$ in a future study should be at least 73.

Suppose there was no policy critical point available for the study. Then a 5% level Z_2 test would be performed on the same data sets. That is, using the data sets used in the Z_1 test, test $H_0 : \xi_{.90} \geq \zeta_{.10}$ versus $H_a : \xi_{.90} < \zeta_{.10}$: Based on the toxicity data, we have

$\hat{\zeta}_{.10} = 3.2508$ and $\hat{f}_y(3.2508) = 0.0894$. Hence, the observed test statistics is $Z_2 = -2.5736$.

Since the p-value of the test is 0.0050, the same conclusion as that from the Z_1 test can be made.

Using the above information and assuming n = m, the sample size needed for a 5% level Z_2 test with an approximate power greater than 0.70 and $d = 2 \cdot \hat{\sigma}_{\hat{\zeta}_{0.90} - \hat{\zeta}_{0.10}}$ in a future study should be at least 81.

CHAPTER V

DISCUSSION

5.1. Confidence Intervals Concerning Percentiles

The proposed test procedures can be easily extended to confidence interval procedures. The $(1-\alpha)100\%$ confidence interval for ξ_p is

$$\hat{\xi}_p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n\hat{f}_x^2\left(\hat{\xi}_p\right)}} \tag{5.1}$$

and the $(1-\alpha)100\%$ confidence interval for the difference between ξ_{p_1} and ζ_{p_2} is

$$\hat{\xi}_{p_1} - \hat{\zeta}_{p_2} \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n\hat{f}_x^2\left(\hat{\xi}_{p_1}\right)} + \frac{p_2(1-p_2)}{m\hat{f}_y^2\left(\hat{\zeta}_{p_2}\right)}} \tag{5.2}$$

These confidence intervals can be compared with the results of other researchers summarized as follows:

Albers & Löhnberg (1984) provided the confidence interval for the difference between quantiles:

$$\mid d - \left(\hat{\xi}_{p_1} - \hat{\zeta}_{p_2}\right) \mid \le \frac{1}{4}\left\{\left(X_{mk_m^+} - X_{mk_m^-}\right)^2 + \left(Y_{nl_n^+} - Y_{nl_n^-}\right)^2\right\}^{\frac{1}{2}}. \tag{5.3}$$

Where $\hat{\xi}_p = F^{-1}(p)$ is the sample p^{th} quantile. Let $X_1, \ldots, X_m$ be a sample from a CDF F.

$\hat{\zeta}_p = G^{-1}(p)$ is the second sample p^{th} quantile. Let $Y_1, \ldots, Y_n$ be a sample from a CDF G.

$d = \xi_p - \zeta_p$. Let $\xi_p = F^{-1}(p)$ is the pth quantile of F. $\zeta_p = G^{-1}(p)$ is the pth quantile of G.

And $\left\{mk_m^-\right\}$ and $\left\{mk_m^+\right\}$ be sequences of integers such that

$$\frac{k_m^{\pm}}{m} = p \pm K_\alpha \left\{ \frac{p(1-p)}{m} \right\}^{\frac{1}{2}} + o\left(m^{-\frac{1}{2}} \right) \tag{5.4}$$

where K_α is the $(1-\alpha)^{\text{th}}$ quantile of the standard normal distribution. Similarly, $\{nl_n^-\}$ and $\{nl_n^+\}$ be sequences of integers such that

$$\frac{l_n^{\pm}}{n} = p \pm K_\alpha \left\{ \frac{p(1-p)}{n} \right\}^{\frac{1}{2}} + o\left(n^{-\frac{1}{2}} \right) \tag{5.5}$$

where K_α is the $(1-\alpha)$'th quantile of the standard normal distribution

Steinberg & Davis(1985) summarized several methods to construct the confidence interval for quantile in small samples: First, Harrel & Davis (1982) propose A $100(1-\alpha)\%$ confidence interval for ξ_p is given by

$$\hat{\xi}_p \pm \phi^{-1}\left(1 - \alpha/2\right) S\left(\hat{\xi}_p\right) \tag{5.6}$$

where

$$\hat{\xi}_p = \sum_{i=1}^{n} W_{n,i} X_{(i)}, W_{n,i} = I_{i/n}\{p(n+1),(1-p)(n+1)\} - I_{\frac{i-1}{n}}\{p(n+1),(1-p)(n+1)\}$$

$$S^2\left(\hat{\xi}_p\right) = \frac{n-1}{n} \sum_{i=1}^{n} \left(S_j - \bar{S}\right)^2, S_j = s_j' \underline{x}, \left(s_j\right)_i = I(i \neq j)\phi(w), \phi(w) = w_{h-1, i - I\{i > j\}}$$

$$\bar{S} = n^{-1} \sum_{j=1}^{n} S_j$$

Kaigh & Lagenbruch (1982) present:

$$\hat{\xi}_p \pm t_{n-k, 1-\alpha/2} S\left(K_p, k\right) \tag{5.7}$$

where $\hat{\xi}_p = \sum_{i=r}^{n+r-k} u_{n,k,i} X_{(i)}, u_{n,k,i} = \binom{j-1}{r-1}\binom{n-j}{k-r}\binom{n}{k}^{-1}, r = [(k+1)p], S^2\left(K_{p,k}\right)$

is jackknife variance estimator. $t_{n-k, 1-\alpha/2}$ is the $1-\alpha/2$ percentage point of the t distribution with n-k degrees of freedom.

Efron (1979) provided the bootstrap interval for the median is

$$X_{(m)} \pm \phi^{-1}\left(1-\alpha/2\right)\left(E_*^2(R)\right)^2,\qquad(5.8)$$

$X_{(m)}$ is the sample median from a sample of size n=2m-1, and

$$E_*\left(R^*\right)^2 = \sum_{l=1}^{n}\left[X_{(l)}-X_{(n)}\right]^2\left[P\left(B_m\left(n,\frac{l-1}{n}\right)\le m-1\right)-P\left(B_m\left(n,\frac{l}{n}\right)\le m-1\right)\right]\quad(5.9)$$

Greenberg & Sarhan (1962) provided a large sample approach produces a confidence

interval

$$\hat{\xi}_p \pm \phi^{-1}\left(1-\alpha/2\right)S\left(\hat{\xi}_p\right),\qquad(5.10)$$

$$where\,\hat{\xi}_p = \frac{n\left(\bar{X}-X_{(1)}\right)\ln\left(1-p\right)}{n-1}\,an\,appropriate\,parametric\,estimator\,for\,a\,quantile$$

$$of\,the\,one-parameter\,\exp onential\left(\lambda=1/\theta\right)distribution,$$

$$S^2\left(\hat{\xi}_p\right)=\frac{n\left(\bar{X}-X_{(1)}\right)^2\left(\ln\left(1-p\right)\right)^2}{\left(n-1\right)^2}$$

Steinberg & Davis (1985) also provided a large sample approach produces a

confidence interval

$$\hat{\xi}_p \pm \phi^{-1}\left(1-\alpha/2\right)S\left(\hat{\xi}_p\right),\qquad(5.11)$$

$$where\,\hat{\xi}_p = p\frac{n}{n-1}X\left(n\right)an\,appropriate\,parametric\,estimator\,for\,a\,quantile$$

$$of\,an\,underlying\,uniform\,distribution\,on\,i\,nterval\left(0,\theta\right),$$

$$S^2\left(\hat{\xi}_p\right)=\left[\frac{p\dfrac{n+1}{n}X_{(n)}}{\sqrt{n\left(n+2\right)}}\right]^2$$

5.2. Test for Difference between Exposure Mean Concentration and the Desired Safety Concentration Level

In some studies, the exposure concentration is defined as the mean concentration. If the analysis is based on the original data set, then by using the central limit theorem, a similar Z test can be developed to test $H_0: \mu_x \geq \zeta_p$ versus $H_a: \mu_x < \zeta_p$; where μ_x is the mean exposure concentration and ζ_p is the HC_p. However, the same procedure can not be used for the log transformed data, since $\log(\bar{X}) \neq \sum \log(X_i)/n$.

5.3. Summary

This study has proposed two distribution free procedures. In these procedures the standard error of a sample percentile is related to the PDF of the underlying distribution and hence the powers of the proposed tests are related to the underlying distribution. Since the percentiles of interest in this study are in the tail of the distribution, for distributions that have similar tail behavior, the density at the percentile of interest should not greatly affect the power of the test.

Censored data are often encountered in the effects analysis and exposure analysis of an ERA for toxic compounds. ECOFRAM (1999) formalized guidelines for censored data in exposure and effects distributions. According to ECOFRAM, chemical exposure data that is below detection limit is assigned a dummy value of zero and used for the calculation of rank preceding the lowest detected concentration. For effects distributions,

point values or values preceded by a greater than (>) that exceed a compound's water solubility for insensitive species or toxicological test response endpoints are used to calculate rank after the highest effect concentration below a compound's. The tests proposed in this study are interested in the right tail of the exposure distribution and the left tail of the toxicity distribution, rather than the censoring sides. Therefore the large tests are fine to censored data sets in ERA.

REFERENCES

Baker, J.L., Barefoot, A.C., Beasley, L.E., Burns, L.A., Caulkins, P.P., Clark, J.E., et al. (1994) Aquatic dialogue group: pesticide risk assessment and mitigation. Pensacola, FL. SETAC Press.

Basawa, I.V., & Koul, H.L. (1988). Large-sample statistics based on quadratic dispersion. *Int. Statist. Rev, 56,* 199-219.

Campbell K.R., Bartell, S.M., & Shaw, J.L. (2000). Characterizing aquatic ecological risks from pesticides using a diquat dibromide .case study. II. Approaches using quotients and distributions. *Environmental Toxicology and Chemistry, 19,* 760-774.

Chen, Ling (in press). A conservative, nonparametric estimator for the 5th percentile of the species sensitivity distribution. *Journal of Statistical Planning and Inference.*

Chen, X.R. (1984). Nonparametric statistics (Lecture note). Univeristy of Science and Technology of China, Hefei, Anhui, China.

Efron, B., (1979). *Order Statistics. 2d ed.,* New York, John Wiley.

EPA (1998) Guideline for ecological risk assessment. *Risk Assessment Forum.* Retrieved October 20, 2002, from http://www.epa.gov

Giddings, J.M., Hall, L.W. Jr., & Solomon, K.R. (2000). Ecological risks of Diazinon from agricultural use in the Sacramento-San Joaquin river basins, California. *Risk Analysis, 20,* 545-572.

Giesy, J.P., Solomon, K.R., Coats, J.R., Dixon, K.R., Giddings, J.M., & Kenaga, E.E. (1999). Chlorpyrifos, ecological risk assessment in North American aquatic environments. *Reviews in Environmental Contamination and Toxicology, 160,* 1-129

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reihold, New York.

Greenberg, E.C., & Sarhan, A.E. (1962). Exponential distribution: best linear unbiased estimates. *Contributions to Order Statistics,* New York, John Wiley, 352-360.

Hall, L.W., Scott, Jr., M.C., & Killen, W.D. (1998). Ecological risk assessment of copper and cadmium in surgace waters of Chesaperake Bay watershed. *Environmental Toxicology and Chemistry, 17,* 172-189.

Hall, L.W., Giddings, Jr., J.M., Solomon, K.R., & Balcomb, R. (1999). An ecological risk assessment of the use of Irgarol 1051 as an algaecide for antifoulant paints. Critical Reviews in *Toxicology, 29*, 367-437.

Hall, L.W., Scott, Jr., M.C., Killen, W.D., & Unger, M.A. (2000). A probability ecological risk assessment of tributyltin in surface waters of Chesapeake Bay watershed. *Human and Ecological Risk Assessment, 6,* 141-179.

Kaigh, W.D., & Lachenbruch, P.A. (1982). A generalized quantile estimator. *Cummun. Statist. –Theor. Meth., A11,* 2217-2238.

Keaney, K.M., & Wei, L.J. (1994). Interim analyses based on median survival times. *Biometrika, 81,* 279-86.

Klaine, S.J., Cobb, G.P., Dickerson, R.L., Dixon, K.R., Kendal, R.J., Smith, E.E., & Solomon, K.R. (1996). An ecological risk assessment for the use of the biocide dibromoni-trilopropionamide (DBNPA) in industrial cooling systems. *Environmental Toxicology and Chemistry, 15,* 21-30.

Kroll, Charles N., & Stedinger, Jery R. (1996) Estimation of moments and quantiles using censored data. *Water Resources Research, 32, No. 4,* 1005-1012.

Kooijman, S.A.L.M. (1987) A safety factor for LC50 values allowing for differences among species. *Water Res., 21,* 269-276

Kosorok, Michael R. (1999) Two-sample quantile tests under general conditions. *Biometrika, 86,* 909-921.

Lin, P.E., & Wu, K.T. (1980). Asymptotic joint distribution of sample quantiles and sample mean with applications. *Commun. Statist. – Theor. Meth., A9(1),* 51-60

Maund, S.J., Travis, K.Z., Hendley, P., Giddings, J.M. & Solomon, K.R. (2001). Probabilistic risk assessment of cotton pyrethroids: V. Combining landscape-level exposures and ecotoxicological effects data to characterize risks. *Environmental Toxicology and Chemistry, 20,* 687-692.

Nadaray, E.A. (1965). On nonparametric estimates of density function and regression curve. *Theory Prob. Appli., 10,* 186-190.

Newman, M.C., Ownby, D.R., Mezin, L.C.A., Powell, D.C., Christensen, T.R.L., Lerberg, S.B., & Anderson,B.-A. (2000). Applying species sensitivity distributions in ecological risk assessment, assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry, 19,* 508-515.

Office of Environmental Information (July, 2000), *Guidance for Data Quality Assessment – Practical Methods for Data Analysis, EPA QA/G-9,QA00UPDATE, EPA/600/R-96/084,* United States Environmental Protection Agency, Washington, DC. 20460. Retrieved December 24, 2002, from http://www.epa.gov/ swerust1/cat/ epaqag9.pdf

Parzen, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist., 33,* 1065-1076.

Rohatgi, Vijay K. (1984) *Statistical Inference.* John Wiley & Sons, Inc.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of density function, *Ann., Math. Statist., 27,* 832-837

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* London, Chapman and Hall.

Solomon, K.R., Baker, D.B., Richards, R.P., Dixon, K.R., Klaine, S.J., La Point, T.W., R.J. et al. (1996). Ecological risk assessment for atrazine in North American surface waters. *Environmental Toxicology and Chemistry, 15,* 31-76.

Solomon, K., Giesy, J., & Jones, P. (2000). Probabilistic risk assessment of agrochemicals in the environment. *Crop Protection, 19,* 649-655.

Steinberg, S.M., & Davis, C.E. (1985). Distribution-free confidence intervals for quantiles in small samples. *Commun. Statist. –Theor. Meth., 14(4),* 979-990.

Su, J.Q & Wei, L.J. (1993). Nonparametric estimation of the difference or ratio of median failure times. *Biometrics,49,* 603-7.

Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J.Am. Statist. Assoc. 69,* 169-73.

Urban, D.J., & Cook, J.N. (1986). *Ecological risk assessment: Hazard Evaluation Division standard procedure.* Washington, D.C.: Office of Pesticide Programs, U.S. Environmental Proctection Agency. EPA-54019-83-001

Van Straalen, N.M. & Denneman, C.A.J. (1989). Ecoloxicological evaluation of soil quality criteria. *Ecotoxical. Envirn. Saf. 18*, 241-251

Wang, J.L. & Hettmansperger, T.P. (1990). Two-sample inference for median survival times based on one-sample procedures for censored survival data. *J.Am. Statist. Assoc. 85*, 529-36.

Wagner, C. and H. Løkke. (1991). Estimation of ecotoxicological protection levels from NOEC toxicity data. *Water Research, 25*, 1237-1242.

APPENDICES

1. Proof Related to Simulation:

$x, y \sim \log istic$

$\xi_1 = \mu_1 + \beta_1 \cdot L_1 \quad where \ L_1 = \log[p_1 /(1-p_1)]$

$\zeta_2 = \mu_2 + \beta_2 \cdot L_2 \quad where \ L_2 = \log[p_2 /(1-p_2)]$

$\mu_2 \ given, \zeta_2 = \mu_2 + \beta_2 \cdot L_2$

$\zeta_2 - \xi_1 = \delta \cdot \sqrt{Var(Y_2)} \Rightarrow \xi_1 = \zeta_2 - \delta \cdot \sqrt{Var(Y_2)}$

$Var(Y_2) = \dfrac{p_2(1-p_2)}{n_2 f_2^2(\zeta_2)}, Var(X_1) = \dfrac{p_1(1-p_1)}{n_1 f_1^2(\xi_1)}$

$assume \ \beta_1 = \beta_2 = \sigma$

$\zeta_2 - \xi_1 = \mu_2 - \mu_1 + \sigma \cdot (L_2 - L_1) \Rightarrow \mu_1 = \mu_2 - (\zeta_2 - \xi_1) + \sigma \cdot (L_2 - L_1)$

$\left. \begin{array}{c} \mu_1 = \mu_2 - (\zeta_2 - \xi_1) + \sigma \cdot (L_2 - L_1) \\ \delta \cdot \sqrt{Var(Y_2)} = \zeta_2 - \xi_1 \\ \zeta_2 = \mu_2 + \sigma \cdot L_2 \end{array} \right\} \Rightarrow \mu_1 = \zeta_2 - \delta \cdot \sqrt{Var(Y_2)} - \sigma \cdot L_1$

$$f_1(\xi_1) = \frac{1}{\sigma} \frac{\exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)}{\left[1 + \exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)\right]^2}, \ f_2(\zeta_2) = \frac{1}{\sigma} \frac{\exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right)}{\left[1 + \exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right)\right]^2}$$

$$\delta* = \frac{\zeta_2 - \xi_1}{\sqrt{Var(Y_2) + Var(X_1)}} = \frac{\delta \sqrt{Var(Y_2)}}{\sqrt{Var(Y_2) + Var(X_1)}} = \frac{\delta}{\sqrt{1 + \dfrac{Var(X_1)}{Var(Y_2)}}} = \frac{\delta}{\sqrt{1 + ratio^2}}$$

$$ratio^2 = \frac{Var(X_1)}{Var(Y_2)} = \frac{\dfrac{p_1(1-p_1)}{n_1 f_1^2(\xi_1)}}{\dfrac{p_2(1-p_2)}{n_2 f_2^2(\zeta_2)}} = \frac{f_2^2(\zeta_2)}{f_1^2(\xi_1)} \ [assume \ p_1(1-p_1) = p_2(1-p_2) \ and \ n_1 = n_2]$$

$$ratio = \frac{f_2(\zeta_2)}{f_1(\xi_1)} = \frac{\exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right) \cdot \sigma \cdot \left[1 + \exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)\right]^2}{\sigma \cdot \left[1 + \exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right)\right]^2 \cdot \exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)} = \frac{\exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right) \cdot \left[1 + \exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)\right]^2}{\left[1 + \exp\left(\dfrac{\zeta_2 - \mu_2}{\sigma}\right)\right]^2 \cdot \exp\left(\dfrac{\xi_1 - \mu_1}{\sigma}\right)}$$

x , y ~ normal

$$\xi_1 = \mu_1 + \sigma_1 \cdot Z_1 \ \ where \ Z_1 = \Phi_{nor}^{-1}(p_1), \ \Phi_{nor}(z) = \int_{-\infty}^{z} (1/\sqrt{2 \cdot \pi}) \cdot \exp(-w^2/2) \cdot dw$$

$$\zeta_2 = \mu_2 + \sigma_2 \cdot Z_2 \ \ where \ Z_2 = \Phi_{nor}^{-1}(p_2)$$

$$\mu_2 \ given, \zeta_2 = \mu_2 + \sigma_2 \cdot Z_2$$

$$\zeta_2 - \xi_1 = \delta \cdot \sqrt{Var(Y_2)} \Rightarrow \xi_1 = \zeta_2 - \delta \cdot \sqrt{Var(Y_2)}$$

$$Var(Y_2) = \frac{p_2(1-p_2)}{n_2 f_2^2(\zeta_2)}, Var(X_1) = \frac{p_1(1-p_1)}{n_1 f_1^2(\xi_1)}$$

assume $\sigma_1 = \sigma_2 = \sigma$

$$\zeta_2 - \xi_1 = \mu_2 - \mu_1 + \sigma \cdot (Z_2 - Z_1) \Rightarrow \mu_1 = \mu_2 - (\zeta_2 - \xi_1) + \sigma \cdot (Z_2 - Z_1)$$

$$\left. \begin{array}{c} \mu_1 = \mu_2 - (\zeta_2 - \xi_1) + \sigma \cdot (Z_2 - Z_1) \\ \delta \cdot \sqrt{Var(Y_2)} = \zeta_2 - \xi_1 \\ \zeta_2 = \mu_2 + \sigma \cdot Z_2 \end{array} \right\} \Rightarrow \mu_1 = \zeta_2 - \delta \cdot \sqrt{Var(Y_2)} - \sigma \cdot Z_1$$

$$f_1(\xi_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\xi_1 - \mu_1)^2}{2\sigma^2}\right], f_2(\zeta_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\zeta_2 - \mu_2)^2}{2\sigma^2}\right]$$

$$\delta^* = \frac{\zeta_2 - \xi_1}{\sqrt{Var(Y_2) + Var(X_1)}} = \frac{\delta\sqrt{Var(Y_2)}}{\sqrt{Var(Y_2) + Var(X_1)}} = \frac{\delta}{\sqrt{1 + \dfrac{Var(X_1)}{Var(Y_2)}}} = \frac{\delta}{\sqrt{1 + ratio^2}}$$

$$ratio^2 = \frac{Var(X_1)}{Var(Y_2)} = \frac{\dfrac{p_1(1-p_1)}{n_1 f_1^2(\xi_1)}}{\dfrac{p_2(1-p_2)}{n_2 f_2^2(\zeta_2)}} = \frac{f_2^2(\zeta_2)}{f_1^2(\xi_1)} \ [assume \ p_1(1-p_1) = p_2(1-p_2) \ and \ n_1 = n_2]$$

$$ratio = \frac{f_2(\zeta_2)}{f_1(\xi_1)} = \frac{\dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(\zeta_2 - \mu_2)^2}{2\sigma^2}\right]}{\dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(\xi_1 - \mu_1)^2}{2\sigma^2}\right]} = \frac{\exp\left[-\dfrac{(\zeta_2 - \mu_2)^2}{2\sigma^2}\right]}{\exp\left[-\dfrac{(\xi_1 - \mu_1)^2}{2\sigma^2}\right]} = \frac{\exp\left[\dfrac{(\xi_1 - \mu_1)^2}{2\sigma^2}\right]}{\exp\left[\dfrac{(\zeta_2 - \mu_2)^2}{2\sigma^2}\right]}$$

2. FORTRAN Codes for Simulation:

Two Sample Simulation Using Kernel Density Estimate:

```fortran
!    using subroutine to simulate two normal case using kernel density estimate
     INTEGER nmax,nmay,ln,lm,ly,ld,lp,ls
     REAL px,py
     Parameter(nmax=1000,nmay=1000,ln=3,lm=100000,ld=5,px=0.9,py=0.1,      &
          ly=1,lp=3,ls=1)
     REAL x(nmax),y(nmay), fhatx,fhaty,mux,muy,orpecy,orpecx,pecx,pecy,op(lp),   &
          fm,p,fn, oosigma(ls),pxinv,pyinv,ratio,deltastar(ld),tsigma,ss,power(ld),   &
          delta(ld),ormuy(ly),osigma
     EXTERNAL RNSET,RNUN,RNUNF,SVRGN,ANORIN,ANORDF,RNNOF
     INTEGER iseed(lm),ip,im,inn,iy,id,an(ln),is
     DATA op/0.01,0.05,0.1/
     DATA an/40, 60, 100/
     DATA delta/0.0, 1.0, 1.5, 3.0,4.0/
     DATA ormuy/2.0/
     DATA oosigma/1.5/
     open (1, FILE='kerres1.dat', STATUS='unknown')
     fm=lm+0.0
     pi=2*ASIN(1.0)
     pxinv=ANORIN(px)
     pyinv=ANORIN(py)
     do 999 ip=1,lp
         p=op(ip)
         crpoint=ANORIN(p)
         write (1,*) 'p=',p
         write (1,*) 'sd=     0.00      0.71     1.06     2.12     2.83'
         call RNSET(13300345)
          do im=1, lm
             iseed(im)=RNUNF()*(10.0**8.0)
          end do
         do 998 inn=1,ln
            n=an(inn)
            fn=n+0.0
            write (1,*) 'n=',n
            do 997 iy=1,ly
                muy=ormuy(iy)
                do 996 is=1,ls
```

```fortran
        osigma=oosigma(is)
        orpecy=muy+osigma*pyinv
        do 995 id=1,ld
            gy=(EXP(-(((orpecy-muy)**2)/                          &
                (2*(osigma**2)))))/(osigma*(SQRT(2*pi)))
            vary=(py*(1-py))/(fn*(gy**2))
            dif=delta(id)*SQRT(vary)
            mux=orpecy-dif-osigma*pxinv
            orpecx=mux+osigma*pxinv
            ratio=EXP((orpecx-mux)**2-(orpecy-muy)**2)
            deltastar(id)=delta(id)/(SQRT(1+ratio**2))
            ss=0.0
            write (1,*)'delta=',delta(id), 'deltastar=',deltastar(id)
            do 994 im=1,lm
                call RNSET(iseed(im))
                call RNORMN(n,x,mux,osigma)
                call RNORMN(n,y,muy,osigma)
                call PERC(n,x,px,pecx)
                call PERC(n,y,py,pecy)
                call FKERN(n,x,px,pecx,fhatx)
                call FKERN(n,y,py,pecy,fhaty)
                tsigma=SQRT((px*(1-px))/(fn*(fhatx**2))              &
                        +(py*(1-py))/(fn*(fhaty**2)))
                ts=(pecx-pecy)/tsigma
                if (ts .lt. crpoint) then
                    ss=ss+1.0
                end if
994             continue
                power(id)=ss/fm
!                write (1,*) power(id)
995         continue
            write (1, 1000) (power(id), id=1, ld)
996       continue
997     continue
998   continue
999 continue
1000  FORMAT(1x, 5(f8.4))
      close (1)
      stop
      end
```

46

```fortran
!
!          the following are subroutine
!
!          generate the normal variable
!
           subroutine RNORMN(n,x,mu,sigma)
           INTEGER n,i
           REAL x(n),mu,sigma,temp
           EXTERNAL RNNOF
             do i=1,n
                 temp=RNNOF()
                 x(i)=mu+sigma*temp
             end do
           return
           end subroutine
!
!          calculate the percentile
!
           subroutine PERC(n,x,px,pecx)
           INTEGER n
           REAL x(n),px,pecx,fn,pix
           EXTERNAL SVRGN
             fn=n+0.0
             call SVRGN(n, x, x)
             pix=px*(fn+1.0)
             pecx=x(INT(pix))+(pix-INT(pix))*(x(INT(pix)+1)-x(INT(pix)))
           return
           end subroutine
!
!          estimate the f function for normal distribution
!
           subroutine FKERN(n,x,px,pecx,fhat)
           INTEGER n,i
           REAL x(n), px,fhat,sumx,sumx2,sdf,kn,pecx,xu(n),xk(n),sumxk
             sumxk=0.0
             fn=n+0.0
!*        calculate standard deviation
             sumx=0.0
             sumx2=0.0
             do i=1,n
```

```fortran
          sumx=sumx+x(i)
          sumx2=sumx2+(x(i))**2
       end do
       sdf=SQRT((sumx2-((sumx)**2)/fn)/(n-1))
!*     sd
       kn=1.05*sdf*fn**(-(1.0/5.0))
       do i=1,n
          xu(i)=(pecx-x(i))/kn
          if ( ABS(xu(i)) .le. SQRT(5.0) ) then
               xk(i)=3*(1-(xu(i))**2/5)/(4*SQRT(5.0))
            else
               xk(i)=0
          end if
          sumxk=sumxk+xk(i)
       end do
       fhat=sumxk/(fn*kn)
    return
    end subroutine
```