## Appendix 4. Analysis primer

While a comprehensive review of the tools used in analysis is not necessary to the management of risks, in this appendix we briefly discuss types of data, data distributions (especially as those relate to biota transfers), reliability and fault-tree analysis, and the evaluation of complex adaptive systems such as those characteristic of the exporting (Missouri River system), importing (Red River system), or engineering controls (e.g., water treatment and distribution system) considered in the current investigation within the context of risk reduction. For a more extensive treatment of any of the analytical tools the reader is referred to the references.

### 4.1 Types of Data: Categorical data and measurement data

Categorical data reflect objects being grouped into categories based on some qualitative trait, and the resulting data are merely labels (Figure 1). Common day examples of categorical data are hair color, flower colors, sex, and in our current investigation, species occurrence data (present/absent data or more precisely, found/not found data). A simple review of even these common day

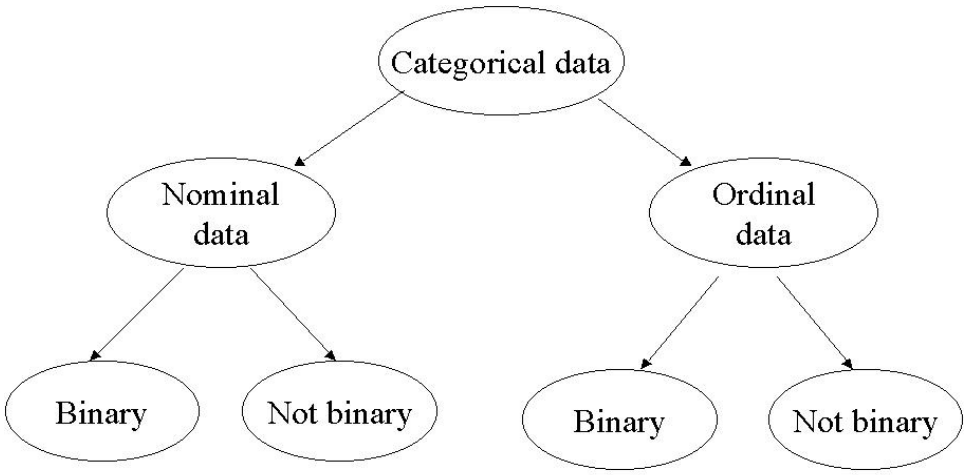



**Figure 1**. Types of categorical data.

examples indicates that categorical data can also be classified based upon the number of categories that are potentially characteristic of all members of the population. Categorical data are classified as being nominal, ordinal, or binary (dichotomous) in character. Nominal data are a type of categorical data in which objects fall into unordered categories (e.g., flower colors). In contrast, ordinal data are categorical data in which order is important, e.g., developmental stages of some invertebrates are an ordered set referred to as eggs, larvae, juveniles, and adults or pathological states such as morbidity may be scored as none, mild, moderate, and severe. Binary or dichotomous data are categorical data that occur as one of two possible states; that is, there are only two independent categories, e.g., species occurrence (e.g., present/absent). Binary data can either be nominal or ordinal.

Measurement data are those that are measured, based on some quantitative trait and the resulting data are set of numbers, e.g., height, weight, age, number of organisms in a region, or stream velocity (Figure 2). Measurement data are classified as discrete or continuous, where discrete

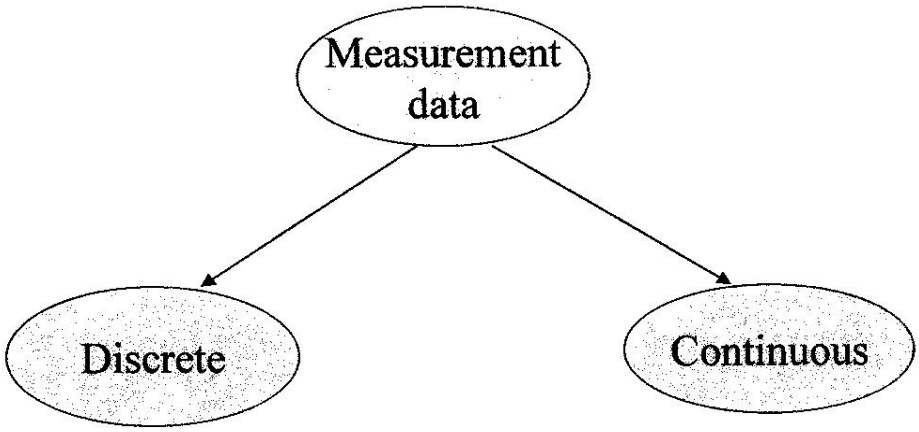



**Figure 2**. Types of measurement data.

measurement data occur as only certain values; that is, there are gaps between the values. Values for discrete data are generally whole numbers and occur at count data, e.g, population counts such as number of fish in a pond. In contrast to discrete measurement data, continuous measurement data may occur as any whole number plus take on any value in the interval between whole numbers, e.g., distance, height, and age. Categorical data are commonly summarized using

"percentages" (or "proportions"), and measurement data are typically summarized using "averages" (or "means") or some descriptive statistic that characterizes a particular attribute of a sample of numbers taken from a population of interest.

### 4.2 Data distributions encountered in the analysis of biota transfer

In data mining operations such as those implemented for the current investigation, an understanding of data and their characteristic distributions are necessary to conduct an analysis of risks, especially for probabilistic analyses (see, e.g, Bedford and Cooke 2001). Predicted or forecasted outcomes of risk scenarios, be those characterized by qualitative or quantitative methods, that capture the concerns of stakeholders reflect issues incorporated into conceptual models of alternative events (such as biota transfers yielding species invasions or shifts in metapopluations). Inevitably, data mining and the evaluation of encountered data has been completed in the absence of a fully characterized distribution of data, which is common in studies such as ours, in part, owing to dependence on diffuse data sources collected across multiple publications across a wide range of time. Our current work, however, frequently requires assumptions of data distributions likely characteristic of these data compiled during the course of the study; hence, a brief overview of frequently encountered data distributions and their interrelationships is included in this appendix in order to better characterize risks, and in particular uncertainties associated with these risks (see standard references and online sources such as Weisstein (1999), e.g., http://mathworld.wolfram.com/about/mathworld.html for source materials for this portion of Appendix 4 and additional detail on data distributions).

**Bernoulli Distribution.** The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by $n = 0$ and $n = 1$ in which $n = 1$ ("success") occurs with probability $p$ and $n = 0$ ("failure") occurs with probability $q \equiv 1 - p$, where $0 < p < 1$ (Figure 3; see, e.g., Evans, et al 2000; Balakrishnan and Nevzorov, 2003). As such, the distribution has probability function:

$$P(n) = \begin{cases} 1-p & \text{for } n = 0 \\ p & \text{for } n = 1, \end{cases}$$

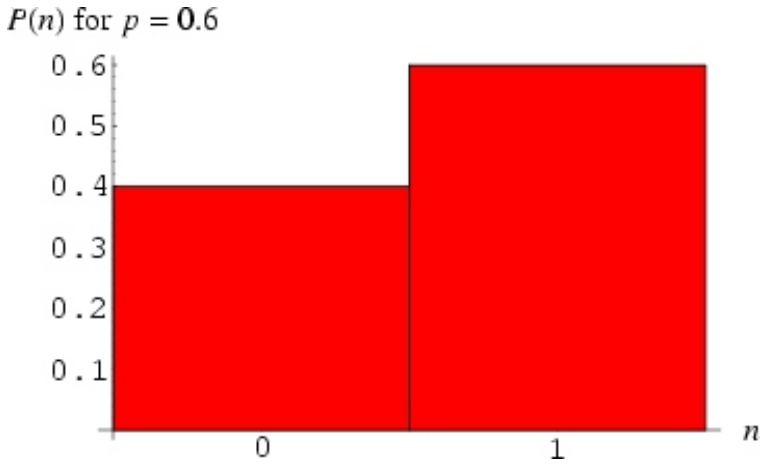



**Figure 3.** Bernoulli distribution.

which can also be written

$$P(n) = p^n(1-p)^{1-n}.$$

The corresponding distribution function is

$$D(n) = \begin{cases} 1-p & \text{for } n = 0 \\ 1 & \text{for } n = 1. \end{cases}$$

The performance of a fixed number of trials with fixed probability of success on each trial is known as a Bernoulli trial, which is an experiment in which *s* trials are made of an event with probability *p* of success in any given trial.

The distribution of heads and tails in coin tossing is an example of a Bernoulli distribution with $p = q = 1/2$. The Bernoulli distribution is the simplest discrete distribution and is the building block for other more complicated discrete distributions. The distributions of a number of variate types are based on sequences of independent Bernoulli trials that are constrained in some way, e.g., the binomial distribution is characterized by the number of successes in *n* trials (Evans et al 2000; Balakrishnan and Nevzorov 2003).

The characteristic Bernoulli function is

$$\phi(t) = 1 + p(e^{it} - 1),$$

and mean , variance , skewness , and kurtosis are then

$$\mu = p$$

$$\sigma^2 = p(1-p)$$

$$\gamma_1 = \frac{1-2p}{\sqrt{p(1-p)}}$$

$$\gamma_2 = \frac{6p^2-6p+1}{p(1-p)}.$$

To find an estimator $\hat{p}$ for the mean of a Bernoulli population with population mean $p$, let $N$ be the sample size and suppose $n$ successes are obtained from the $N$ trials. Assume an estimator given by

$$\hat{p} \equiv \frac{n}{N},$$

so that the probability of obtaining the observed $n$ successes in $N$ trials is then

$$\binom{N}{n} p^n (1-p)^{N-n}.$$

The expectation value of the estimator $\hat{p}$ is therefore given by

$$\langle \hat{p} \rangle = \sum_{n=0}^{N} p \binom{N}{n} p^n (1-p)^{N-n}$$

$$= (1-p)^N \left( \frac{1}{1-p} \right)^N p = p,$$

so $\hat{p}$ is indeed an unbiased estimator for the population mean $p$.

**Binomial Distribution.** The binomial distribution gives the discrete probability distribution $P_p(n|N)$ of obtaining exactly *n* successes out of *N* Bernoulli trials, where the result of each Bernoulli trial is true with probability *p* and false with probability $q = 1 - p$ (see, e.g., Evans et al 2000; Balakrishnan and Nevzorov 2003). The binomial distribution is therefore given by:

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

$$= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n},$$

where $\binom{N}{n}$ is a binomial coefficient . The following plot (Figure 4) shows the distribution of *n* successes out of $N = 20$ trials with $p = q = 1/2$.

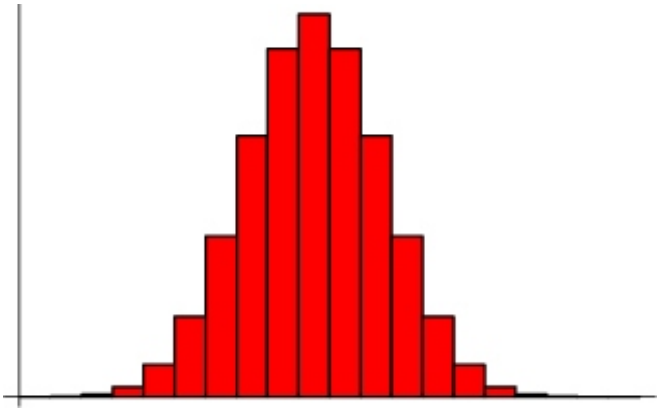


**Figure 4.** Binomial distribution.

The probability of obtaining *more* successes than the *n* observed in a binomial distribution is

$$P = \sum_{k=n+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = I_p(n+1, N-n),$$

where

$$I_x(a,b) \equiv \frac{B(x;a,b)}{B(a,b)},$$

$B(a,b)$ is the beta function , and $B(x;a,b)$ is the incomplete beta function.

The characteristic function for the binomial distribution is

$$\phi(t) = (q + pe^{it})^n$$

(see, e.g., Evans et al 2000; Balakrishnan and Nevzorov 2003), and the skewness and kurtosis are

$$\begin{aligned}\gamma_1 &= \frac{1-2p}{\sqrt{Np(1-p)}} \\ &= \frac{q-p}{\sqrt{Npq}} \\ \gamma_2 &= \frac{6p^2-6p+1}{Np(1-p)} \\ &= \frac{1-6pq}{Npq}.\end{aligned}$$

The mean deviation is given by

$$\mathrm{MD} = \sum_{k=0}^{N} |k - Np| \binom{N}{k} p^k (1-p)^{N-k}.$$

For the special case $p = q = 1/2$, this is equal to

$$\mathrm{MD} = 2^{-N} \sum_{k=0}^{N} \binom{N}{k} \left|k - \tfrac{1}{2}N\right|$$

$$= \begin{cases} \frac{N!!}{2(N-1)!!} & \text{for } N \text{ odd} \\ \frac{(N-1)!!}{2(N-2)!!} & \text{for } N \text{ even,} \end{cases}$$

where $N!!$ is a double factorial . For $N$ = 1, 2, ..., the first few values are therefore 1/2, 1/2, 3/4, 3/4, 15/16, 15/16, ... A complete derivation is not included here. However, treating the distribution as continuous,

$$\lim_{N\to\infty} \sum_{n=0}^{N} P(n) \approx \int P(n)\, dn = \int_{-\infty}^{\infty} P(\tilde{n}+\eta)\, d\eta = 1.$$

Since each term is of order $1/N \sim 1/\sigma^2$ smaller than the previous, we can ignore terms higher than $B_2$, so

$$P(n) = P(\tilde{n}) e^{-|B_2|\eta^2/2}.$$

The probability must be normalized, so

$$\int_{-\infty}^{\infty} P(\tilde{n}) e^{-|B_2|\eta^2/2}\, d\eta = P(\tilde{n}) \sqrt{\frac{2\pi}{|B_2|}} = 1,$$

and

$$P(n) = \sqrt{\frac{|B_2|}{2\pi}} e^{-|B_2|(n-\tilde{n})^2/2}$$

$$= \frac{1}{\sqrt{2\pi Npq}} \exp\left[-\frac{(n-Np)^2}{2Npq}\right].$$

Defining $\sigma^2 \equiv Npq$,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$P(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(n-\tilde{n})^2}{2\sigma^2}\right],$$

which is a normal distribution. For $p \ll 1$, a different approximation procedure shows that the binomial distribution approaches the Poisson distribution (see Haight 1967).

**Normal distribution and the Central Limit Theorem.** A normal distribution in a variate $X$ with mean $\mu$ and variance $\sigma^2$ has probability function on the domain $x \in (-\infty, \infty)$. The term "normal distribution" or "Gaussian distribution" are commonly used in reference to this distribution, and because of its curved flaring shape, social scientists refer to it as the "bell curve" (Figure 5; see Patel and Read, 1982).

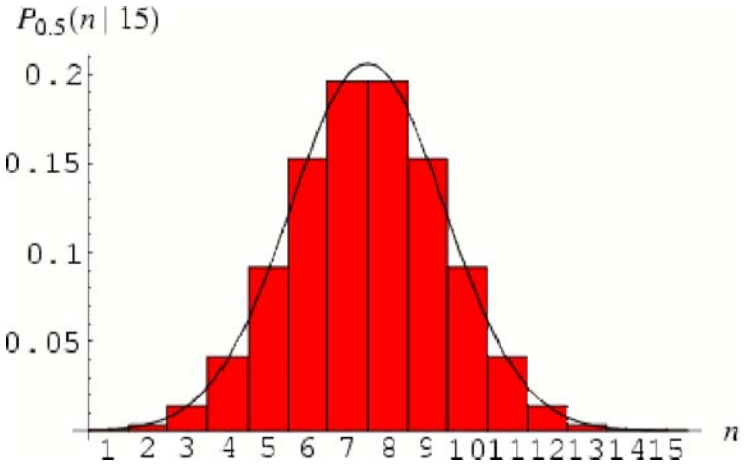



**Figure 5.** Normal distribution.

The so-called "standard normal distribution" is given by taking $\mu = 0$ and $\sigma^2 = 1$ in a general normal distribution. An arbitrary normal distribution can be converted to a standard normal

distribution by changing variables to $Z \equiv (X - \mu)/\sigma$, so $dz = dx/\sigma$, yielding:

$$P(x)\,dx = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz.$$

The normal distribution function $\Phi(z)$ gives the probability that a standard normal variate assumes a value in the interval $[0, z]$,

$$\Phi(z) \equiv \frac{1}{\sqrt{2\pi}} \int_0^z e^{-x^2/2}\,dx = \tfrac{1}{2}\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right),$$

where erf is a function sometimes called the error function; neither $\Phi(z)$ nor erf can be expressed in terms of finite additions, subtractions, multiplications, and root extractions. Consequently, both must be either computed numerically or otherwise approximated.

The normal distribution (Figure 5) is the limiting case of a discrete binomial distribution $P_p(n|N)$ as the sample size *N* becomes large, in which case $P_p(n|N)$ is normal with mean and variance

$$\mu = Np$$

$$\sigma^2 = Npq,$$

respectively, when $q \equiv 1 - p$.

The distribution *P*(*x*) is properly normalized since

$$\int_{-\infty}^{\infty} P(x)\,dx = 1.$$

The cumulative distribution function , which gives the probability that a variate will assume a value $\leq x$, is then the integral of the normal distribution,

$$D(x) \equiv \int_{-\infty}^{x} P(x')\,dx'$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x'-\mu)^2/(2\sigma^2)}\,dx'.$$

$$= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right],$$

where erf is again called the error function.

Normal distributions have many convenient properties, so random variates with unknown distributions are often assumed to be normal. Although this can be a dangerous assumption, it is often a good approximation due to a surprising result known as the central limit theorem . This theorem states that the mean of any set of variates with any distribution having a finite mean and variance tends to the normal distribution. Many common attributes conform to a normal distribution, with few members at the high and low ends and many in the middle. Because the normal distribution occurs frequently, there is a tendency to invoke assumptions of normality in situations where they may not be applicable: "Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation" (Whittaker and Robinson 1967).

The unbiased estimator for the variance of a normal distribution is given by

$$\sigma^2 = \frac{N}{N-1} s^2,$$

where

$$s \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2,$$

so

$$\mathrm{var}(\bar{x}) = \frac{s^2}{N-1}.$$

The characteristic function for the normal distribution is

$$\phi(t) = e^{i\mu t - \sigma^2 t^2/2},$$

and the variance , skewness , and kurtosis excess are given by

$$\mathrm{var}(x) = \sigma^2$$

$$\gamma_1 = 0$$

$$\gamma_2 = 0.$$

The variance of the sample variance $s^2$ for a general distribution is given by

$$\mathrm{var}(s^2) = \frac{(N-1)[(N-1)\mu_4 - (N-3)\mu_2^2]}{N^3},$$

which simplifies in the case of a normal distribution to

$$\mathrm{var}(s^2) = \frac{2\sigma^4(N-1)}{N^2}$$

(Kenney and Keeping 1951). If $P(x)$ is a normal distribution, then

$$D(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right],$$

so variates $X_i$ with a normal distribution can be generated from variates $Y_i$ having a uniform distribution in (0,1) via

$$X_i = \sigma\sqrt{2}\,\mathrm{erf}^{-1}(2Y_i - 1) + \mu.$$

The normal distribution is also a special case of the chi-squared distribution , since making the substitution

$$\tfrac{1}{2}z \equiv \frac{(x-\mu)^2}{2\sigma^2}$$

gives

$$d(\tfrac{1}{2}z) = \frac{(x-\mu)}{\sigma^2}\,dx = \frac{\sqrt{z}}{\sigma}\,dx.$$

Now, the real line $x \in (-\infty, \infty)$ is mapped onto the half-infinite interval $z \in [0, \infty)$ by this transformation, so an extra factor of 2 must be added to $d(z/2)$, transforming $P(x)\,dx$ into

$$P(z)\,dz = \frac{1}{\sigma\sqrt{2\pi}} e^{-z/2} \frac{\sigma}{\sqrt{z}}\, 2(\tfrac{1}{2}\,dz)$$

$$= \frac{e^{-z/2} z^{-1/2}}{2^{1/2}\Gamma(\frac{1}{2})}\,dz$$

(Kenney and Keeping 1951), where use has been made of the identity $\Gamma(1/2) = \sqrt{\pi}$.

**Poisson distribution and rare events.** A Poisson process is one that satisfies the following properties:

- The numbers of changes in nonoverlapping intervals are independent for all

intervals.

- The probability of exactly one change in a sufficiently small interval $h \equiv 1/n$ is

$$P = \nu h \equiv \nu / n,$$

where $\nu$ is the probability of one change and *n* is the number of trials.

- The probability of two or more changes in a sufficiently small interval *h* is essentially 0.

In the limit of the number of trials becoming large, the resulting distribution is called a Poisson distribution (Figure 6; see Haight 1967).

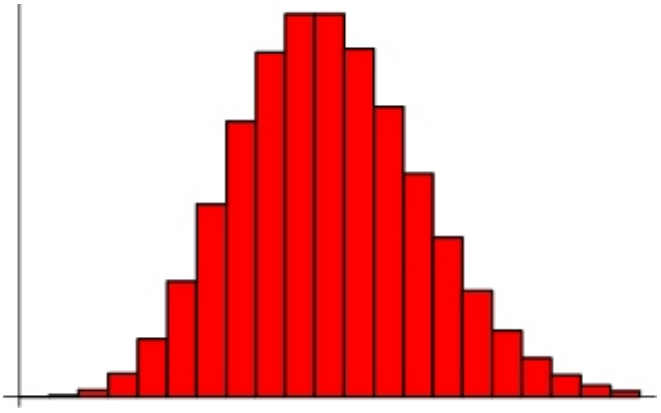


**Figure 6.** Poisson distribution.

Given a Poisson process, the probability of obtaining exactly *n* successes in *N* trials is given by the limit of a binomial distribution

$$P_p(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}.$$

Viewing the distribution as a function of the expected number of successes $\nu \equiv Np$ instead of the sample size *N* for fixed *p*, the equation then becomes

$$P_{\nu/N}(n|N) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n},$$

Letting the sample size *N* become large, the distribution then approaches:

$$P_\nu(n) = \lim_{N\to\infty} P_B(n)$$

$$= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{n!}\frac{\nu^n}{N^n}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n}$$

$$= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{N^n}\frac{\nu^n}{n!}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n}$$

$$= 1\cdot\frac{\nu^n}{n!}\cdot e^{-\nu}\cdot 1$$

$$= \frac{\nu^n e^{-\nu}}{n!},$$

which is known as the Poisson distribution (see, e.g., Haight 1967; Papoulis 1984; Pfeiffer and Schum 1973). Note that the sample size *N* has completely dropped out of the probability function, which has the same functional form for all values of $\nu$.

The Poisson distribution is normalized so that the sum of probabilities equals 1, since

$$\sum_{n=0}^{\infty} P_\nu(n) = e^{-\nu}\sum_{n=0}^{\infty}\frac{\nu^n}{n!} = e^{-\nu}e^{\nu} = 1.$$

The mean , variance , skewness , and kurtosis are

$$\mu = \nu$$

$$\sigma^2 = \nu$$

$$\gamma_1 \equiv \frac{\mu_3}{\sigma^3} = \frac{\nu}{\nu^{3/2}} = \nu^{-1/2}$$

$$\gamma_2 \equiv \frac{\mu_4}{\sigma^4} - 3 = \frac{\nu(1+3\nu)}{\nu^2} - 3$$

$$= \frac{\nu + 3\nu^2 - 3\nu^2}{\nu^2} = \nu^{-1}.$$

The characteristic function for the Poisson distribution is

$$\phi(t) = e^{\nu(e^{it}-1)}$$

(Haight 1967; Papoulis 1984), and the cumulative function is

$$K(h) = \nu(e^h - 1) = \nu(h + \tfrac{1}{2!}h^2 + \tfrac{1}{3!}h^3 + \ldots),$$

so

$$\kappa_r = \nu.$$

The Poisson distribution can also be expressed in terms of

$$\lambda \equiv \frac{\nu}{x},$$

the rate of changes, so that

$$P_\nu(n) = \frac{(\lambda x)^n e^{-\lambda x}}{n!}.$$

**Biological implications of data types and data distributions.** While the underlying mathematical principals excerpted from http://mathworld.wolfram.com/about/mathworld.html (see Weisstein (1999), last accessed November 16, 2004) and briefly summarized in this appendix are infrequently considered in the quantitative assessment of biological functions and processes,

the risk analysis and subsequent interpretation of risks must be completed and interpreted with an awareness of what these principals are and what they mean within the context of managing risks characterized by investigations such as the current effort dependent on existing data and available information from a wide variety of sources.

Species presence and absence data, or more appropriately stated, data that characterize species as found or not found, are binary or Bernoulli variables. As noted in the preceding sections of the primer, a Bernoulli random variable is formally described as a variable that results from an experiment in which *s* trials are made of an event, with probability *p* of success (i.e., found) in any given trial. By definition, a failure (i.e., not found) in any given trial is characterized as $q = 1 - p$. The classical example of a Bernoulli trial is a coin toss, preferably with a "fair coin," one where the probability of a head or tail is equal on any given trial. A binomial distribution will characterize the outcomes of repeated Bernoulli trials, where the binomial distribution gives the discrete probability distribution of obtaining exactly *n* successes out of *N* Bernoulli trials where the result of each Bernoulli trial is characterized by a probability $p + q = 1$. The binomial distribution is therefore given by

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

$$= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n},$$

where $\binom{N}{n}$ is a binomial coefficient. While the intricacies of interrelationships between differing statistical distributions will not be developed further (see Evans et al 2000 and Balakrishnan and Nevzorov 2003 for additional details), there are two important distributions, the normal

distribution as specified,

$$P(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(n-\tilde{n})^2}{2\sigma^2}\right],$$

and the Poisson distribution as specified,

$$P_\nu(n) = \frac{\nu^n e^{-\nu}}{n!},$$

which appear as the limits for sequences of binomial distributions. The differences between these limiting distributions reflects the contrasting asymptotic behaviors of alternative sequences of binomial random variables and reflect the influence of sample size and *p* values on system behavior (e.g., distributions of binomial random variables reflecting $p << 1$ approach limits captured by the Poisson distribution). Applications of these analytical tools to problems such as those of interest in the present investigation reflect these asymptotic behaviors and their boundary conditions.

The balance of this section will focus on an overview of how these distributional properties influence the work completed as part of this investigation focused on potential biota transfers between the Missouri River and Red River basins.

### 4.3 Reliabililty and fault-tree analysis: Fault-Probability Trees (FPTs)

Complex interactive systems, be those engineered systems designed and constructed following industry standards or biological systems at any level of organization (e.g., molecular, cellular, tissues and organs, organismic, populations, communities, or ecosystems), are subject to inevitable events commonly referred to as "failures." These failures potentially compromise the system's performance for various time periods, ranging from the inconsequential events to catastrophic terminal events. Failure analysis, especially within the context of biological systems and their relationships to alternative engineering systems, was a primary tool in the evaluation of

risks of biota transfers associated with water diversions between the Missouri River and Red River basin.

Failures range from the inconsequential to the catastrophic. But, from the point of view of assessing system reliability, catastrophic failures are handled no differently from failures that occur when a key parameter of a system of interdependent components drifts slightly out of specification. Regardless of the systems complexity, departures from nominal structure or function call for an unscheduled maintenance action or recovery process in engineering or ecological systems, respectively. Consequences associated with failure events vary widely, since the restoration of a system's performance is a function of magnitude of departure (e.g., more than one component fails) and the sensitivity of the system to failure of its various components (e.g., not all components may be equally sensitive to failure and some components may be more critical to system performance than others at various periods in a system's lifetime).

For the current investigation, problem formulation identified the "failures" that were the major concerns of Reclamation and Technical Team. From a system perspective, failures were considered to be biota transfers (both species invasions and shifts in metapopulations) that were variably affected by alternative control systems interjected into the water diversion to the attendant reduce risks. Failure analysis, then, was critical to the evaluation of risks, since the biological or ecological "failures" (e.g., a species invasion) associated with interbasin water transfers were influenced by "failures"in the alternative technologies incorporated into the proposed water distribution system linking Missouri River sources with importing areas in the Red River basin.

As background to the current investigation and to encourage future iterations of this analysis consider the critical interactions between biological and ecological systems and the role that engineering systems play in reducing risks, a brief overview of failure analysis follows. For more comprehensive technical guidance on failure analysis and its potential value in evaluating risks and consequences, the reader is referred to Barlow (1998), Blischke and Parbhakar Murthy (2000),

and NIST/SEMATECH (2004).

**Repairable and non-repairable systems and lifetime distribution models.** A repairable system is one which can be restored to satisfactory operation following some scheduled or unscheduled action to remedy a departure from acceptable performance (a failure), e.g., control systems involving water filtration will have a routine maintenance schedule to reduce risks of failure in treatment system, or ecosystems may recover following unsuccessful species invasions. When discussing the rate at which failures occur during system operation (and are then repaired), an engineer will define a "Rate Of Occurrence Of Failure" (ROCF) or "repair rate" which would be roughly equivalent to the restoration ecologist's term of "recovery rate." While the engineer actively develops corrective action plans (e.g., scheduled maintenance), restoration ecologists may assume active or passive roles in the recovery process (see, e.g., Jordan et al 1987; Manci 1989; FISRWG, 1998). For engineering systems, "failure rates" or "hazard rates" are terms applied to the first failure times for a population of non-repairable components or to non-repairable systems. Biological analogs of non-repairable components or non-repairable systems would be characterized as aging-related events (e.g., decreased fecundity) commonly measured as changes in survivorship (for example) in life-table analysis. A non-repairable population is one for which individual items that fail are removed permanently from the population. While the system may be repaired by replacing failed units from either a similar or a different population, the members of the original population dwindle over time until all have eventually failed. The comparison to cohorts and their passage through the population ecologist's life table are clearly evident (see, e.g., Caswell 2001).

**Tools for evaluating non-repairable populations.** In general, population models used to describe unit lifetimes are known as lifetime distribution models regardless of whether the populations of interest are biological or engineering in origin. A population is generally considered to be all of the possible unit lifetimes for all of the units, and a random sample of size $n$ from this population is the collection of failure times observed for a randomly selected group of $n$ units. A lifetime distribution model can be any probability density function (or PDF), $f(t)$, defined

over the range of time from $t = 0$ to $t =$ infinity. The corresponding cumulative distribution function (or CDF), $F(t)$, characterizes the probability that a randomly selected unit will fail by time $t$. Figure 7 that follows illustrates the relationship between $f(t)$ and $F(t)$. The lifetime CDF may be characterized by $F(t)$ as (1) $F(t)$ = the area under the PDF $f(t)$ to the left of $t$; (2) $F(t)$ = the probability that a single randomly chosen new unit will fail by time $t$; and (3) $F(t)$ = the proportion of the entire population that fails by time $t$.

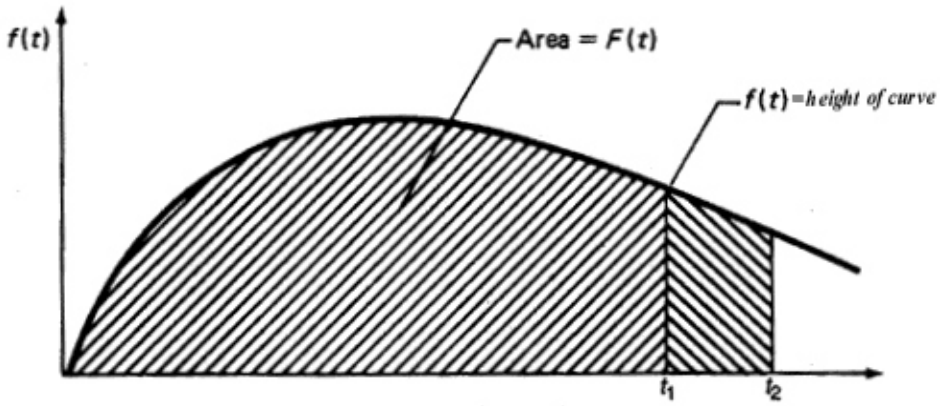



Figure 7. Cumulative distribution function for lifetime model.

The figure above also shows a shaded area under $f(t)$ between the two times $t_1$ and $t_2$. This area is $[F(t_2) - F(t_1)]$ and represents the proportion of the population that fails between times $t_1$ and $t_2$ (or the probability that a brand new randomly chosen unit will survive to time $t_1$ but fail before time $t_2$). It is worthy to note that the PDF $f(t)$ has only non-negative values and eventually either becomes 0 as $t$ increases or decreases towards the origin. Ideally, the CDF $F(t)$ is monotonically increasing and goes from 0 to 1 as $t$ approaches infinity. In other words, the total area under the curve is always 1.

A good example of a life distribution model is the 2-parameter Weibull distribution for $F(t)$. It has the CDF and PDF equations given by:

$$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}, \quad f(t) = \frac{\gamma}{t}\left(\frac{t}{\alpha}\right)^{\gamma} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$$

where $\gamma$ is the "shape" parameter and $\alpha$ is a scale parameter called the characteristic life.

Survival is the complementary event to failure, and the reliability function, *R(t)*, also known as the survival function, *S(t)*, is defined by:

*R(t)* = *S(t)* = the probability a unit survives beyond time *t*.

Since a unit either fails or survives, and one of these two mutually exclusive alternatives must occur, we have

*R(t) = 1 - F(t), F(t) = 1 - R(t)*

Calculations using *R(t)* often occur when building up from single components to subsystems with many components. The reliability of a system is the product of the reliability functions of the components since both must survive in order for the system to survive. Building up to a "system" from the individual components is referred to as the "bottom-up" method. The bottom-up method is guided by the general rule: to calculate the reliability of a system of independent components, multiply the reliability functions of all the components.

**Failure (or hazard) rate.** The failure rate is the rate at which the population survivors at any given instant are "falling over the cliff," that is the failure rate is defined for non-repairable populations as the (instantaneous) rate of failure for the survivors to time *t* during the next instant of time. It is a rate per unit of time, and it represents a "snapshot" in time, since the next instant the failure rate may change and the units that have already failed play no further role since only the survivors count. The failure rate (or hazard rate) is denoted by *h(t)* and calculated from

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)} = \text{the instantaneous (conditional) failure rate.}$$

The failure rate is sometimes called a "conditional failure rate" since the denominator 1 - *F(t)* (i.e., the population survivors) converts the expression into a conditional rate, given survival past some time, *t*. Since *h(t)* is equal to the negative of the derivative of ln{*R(t)*}, we have the useful identity:

$$F(t) = 1 - \exp\left\{ -\int_0^t h(t)dt \right\}$$

If we let

$$H(t) = \int_0^t h(t)dt$$

be the cumulative hazard function, we then have $F(t) = 1 - e^{-H(t)}$. Two other useful identities that follow are:

$$h(t) = -\frac{d \ln R(t)}{dt}$$

$$H(t) = -\ln R(t).$$

A failure rate over any interval ($T_1 \rightarrow T_2$) characterizes an "average" failure rate for the interval and is denoted by $AFR(T_1, T_2)$. AFR's are calculated:

$$AFR(T_2 - T_1) = \frac{\left( \int_{T_1}^{T_2} h(t)dt \right)}{T_2 - T_1} = \frac{H(T_2) - H(T_1)}{T_2 - T_1} = \frac{\ln R(T_1) - \ln R(T_2)}{T_2 - T_1}$$

$$AFR(0, T) = AFR(T) = \frac{H(T)}{T} = \frac{-\ln R(T)}{T}$$

**Graphical depictions of failure rates: "Bathtub" curve.** A plot of the failure rate over time yields a curve that looks like a drawing of a bathtub (at least to an engineer; Figure 8). If enough units from a given population are observed operating and failing over time, it is relatively easy to compute estimates of the failure rate *h(t)*.

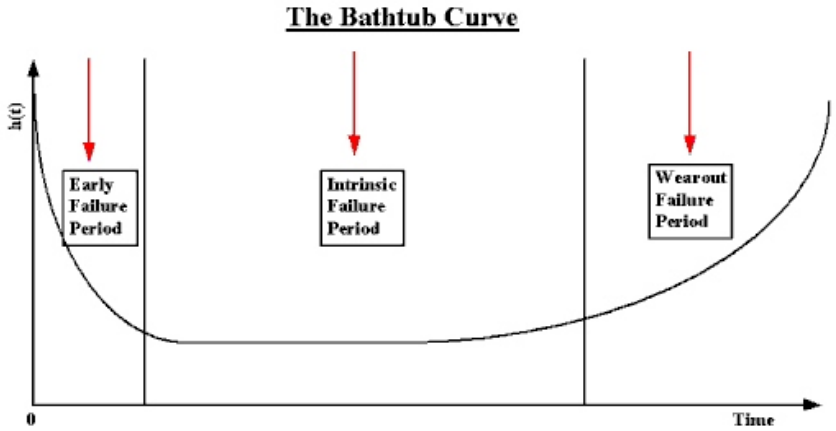



Figure 8. Typical "bathtub" curve of the reliability engineer.

In an idealized bathtub curve, the initial region begins at time zero ($t_0$) when a system's operation commences (which is analogous to birth in life table analysis). The system is initially characterized by a high but rapidly decreasing failure rate (e.g., early failure period for an engineering system, infant mortality period for biological populations and actuaries), with the decreasing failure rate typically lasting several weeks to a few months depending on the system. Following the initial, frequently transitory high failure rate, the failure rate levels off and remains roughly constant for throughout "useful life of the system." This long period of a relative constant failure rate is known as the intrinsic failure period or the stable failure period. The constant failure rate level during this period is referred to as the intrinsic failure rate. Most systems function most of their lifetimes in this flat portion of the bathtub curve. If units from the population remain in use long enough, the failure rate begins to increase as materials wear out and degradation failures occur at an ever increasing rate. This is the "wearout failure period."

Based on empirical observations, the bathtub curve also applies to repairable systems, but in this

instance, a "repair rate" or the "rate of occurrence of failures" (ROCOF) characterizes the ordinate of Figure 8. A different approach is used for modeling the repair rates for a repairable system, since failures occur at given system ages and the system, once repaired, be the same as new, or better, or worse than the original system. Frequency of repairs may be increasing, decreasing, or staying at a roughly constant rate, and may be characteristic of a given system.

Let *N(t)* be a counting function that keeps track of the cumulative number of failures a given system has had from $t_0$, $t_1$, $t_2$,. . .$t_n$, $t_{n+1}$. Then, *N(t)*is a step function that jumps up one every time a failure occurs and stays at the new level until the next failure. Every system will have its own observed *N(t)* function over time. If we observed the *N(t)* curves for a large number of similar systems and "averaged" these curves, we would have an estimate of *M(t)* = the expected number (average number) of cumulative failures by time *t* for these systems. Repair rate is the mean rate of failures per unit time, and the derivative of *M(t)*, denoted *m(t)*, is defined as the repair rate at time, *t*.

**Lifetime distribution models.** A handful of lifetime distribution models are commonly applied to investigations where data mining provides "starter sets" for an analysis. While empirical data sets developed as a direct result of observational or designed studies have contributed much to the literature for use in the current investigation focused on biota transfers, the inevitable stochastic character of the invasion process leads the analysis of risks to distribution models that have enjoyed great practical success in past investigations. There are a handful of distribution models that have successfully served as population models for lifetime distributions and failure times arising from a wide range of applications (e.g., engineering, biological, and ecological) and failure mechanisms. Sometimes there are probabilistic arguments based on the physics of the failure mode that tend to justify the choice of model. At other times the model is used solely because of its empirical success in fitting actual failure data. Six models frequently used are described in this appendix: Exponential, Weibull, Extreme Value, Lognormal, Gamma, and Proportional Hazards.

**Exponential distribution.** The exponential model with only one unknown parameter is the

simplest of all distribution models. The key equations for the exponential distribution are listed below, with the failure rate reducing to the constant $\lambda$ for any time. As a consequence, another name for the exponential mean is the "mean time to fail" (MTTF) = $1/\lambda$. The exponential distribution is the only distribution to have a constant failure rate. The Cum Hazard function for the exponential is just the integral of the failure rate or $H(t) = \lambda t$. The PDF and CDF for the exponential have the familiar shapes shown below (Figure 9 and Figure 10, respectively).

$$\text{CDF: } F(t) = 1 - e^{-\lambda t}$$

$$\text{RELIABILITY: } R(t) = e^{-\lambda t}$$

$$\text{PDF: } f(t) = \lambda e^{-\lambda t}$$

$$\text{MEAN: } \frac{1}{\lambda}$$

$$\text{MEDIAN: } \frac{\ln 2}{\lambda} \cong \frac{.693}{\lambda}$$

$$\text{VARIANCE: } \frac{1}{\lambda^2}$$

$$\text{FAILURE RATE: } h(t) = \lambda$$

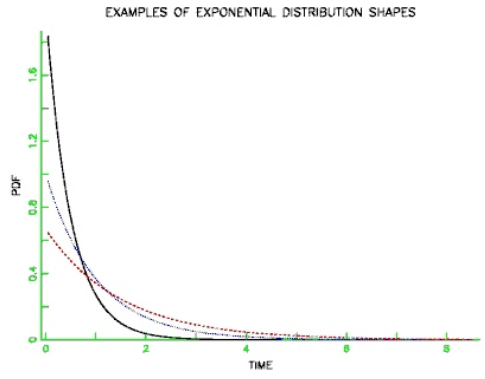



**Figure 9.** PDF for exponential distribution.

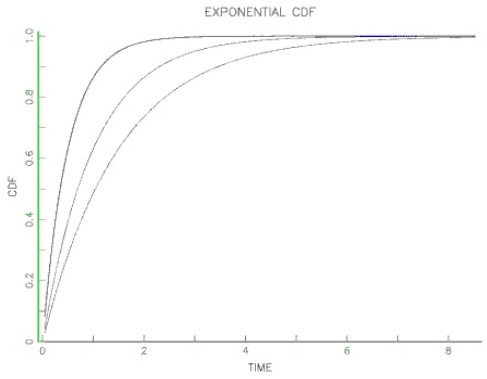



**Figure 10.** CDF for exponential distribution.

The exponential distribution models the flat portion of the "bathtub" curve, because of its

constant failure rate property. Since most components and systems spend most of their lifetimes in this portion of the Bathtub Curve, this justifies frequent use of the exponential distribution when early failures or wear out is not a concern.

**Weibull distribution.** The Weibull distribution is a very flexible life distribution model with two parameters, and has CDF and PDF and other key formulas given by:

$$\text{CDF: } F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$$

$$\text{RELIABILITY: } e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$$

$$\text{PDF: } f(t) = \frac{\gamma}{t}\left(\frac{t}{\alpha}\right)^{\gamma} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$$

$$\text{FAILURE RATE: } \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1}$$

$$\text{MEAN: } \alpha\Gamma\left(1+\frac{1}{\gamma}\right)$$

$$\text{MEDIAN: } \alpha(\ln 2)^{\frac{1}{\gamma}}$$

$$\text{VARIANCE: } \alpha^2\Gamma\left(1+\frac{2}{\gamma}\right) - \left[\alpha\Gamma\left(1+\frac{1}{\gamma}\right)\right]^2$$

with $\alpha$ the scale parameter (the characteristic life), $\gamma$ (gamma) the shape parameter, and $\Gamma$ is the Gamma function with $\Gamma(N) = (N\text{-}1)!$ for integer $N$. The Cum Hazard function for the Weibull is the integral of the failure rate or

$$H(t) = \left(\frac{t}{\alpha}\right)^{\gamma}$$

A more general 3-parameter form of the Weibull includes an additional waiting time parameter μ (sometimes called a shift or location parameter). The formulas for the 3-parameter Weibull are easily obtained from the above formulas by replacing $t$ by $(t - \mu)$ wherever $t$ appears. No failure can occur before μ hours, so the time scale starts at μ, and not 0. If a shift parameter μ is known (based, perhaps, on the physics of the failure mode), then all you have to do is subtract μ from all the observed failure times and/or readout times and analyze the resulting shifted data with a 2-parameter Weibull. When $\gamma = 1$, the Weibull reduces to the exponential model , with $\alpha = 1/\lambda$ = the "mean time to fail" (MTTF). Depending on the value of the shape parameter $\lambda$ , the

Weibull model can empirically fit a wide range of data histogram shapes as illustrated below (Figure 11).

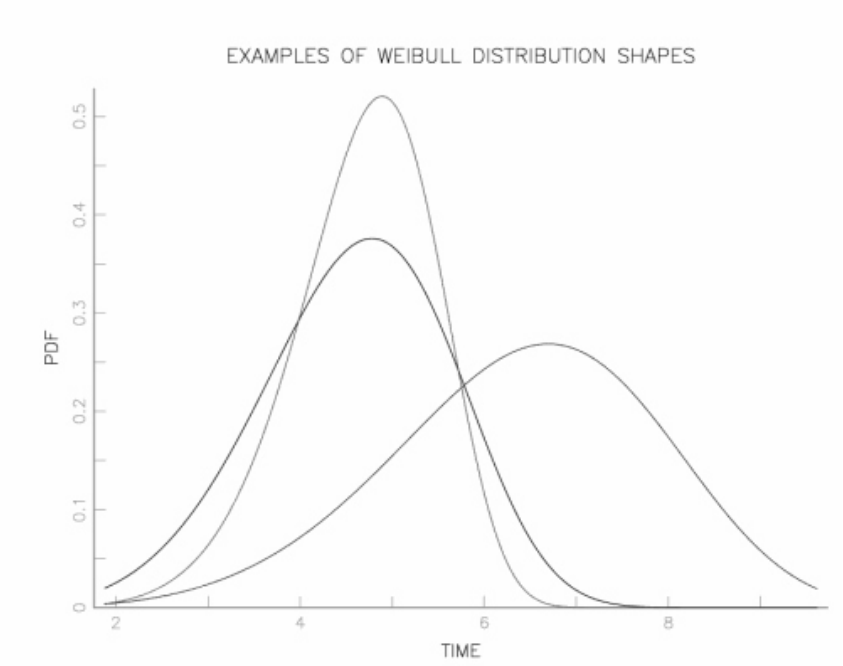



**Figure 11.** Illustrations of Weibull distribution.

As a failure rate model, the Weibull is a natural extension of the constant failure rate exponential model since the Weibull has a polynomial failure rate with exponent $\{\gamma - 1\}$. The Weibull has been applied to many failure analyses because of its flexible shape and ability to model a wide range of failure rates across a wide range of physical and biological systems.

**Extreme value distributions.** Extreme value distributions are the limiting distributions for the minimum or the maximum of a very large collection of random observations from the same arbitrary distribution (see Castillo et al, 2005). Gumbel (1958) showed that for any well-behaved initial distribution (i.e., *F(x)*is continuous and has an inverse), only a few models are needed, depending on whether you are interested in the maximum or the minimum, and also if the observations are bounded above or below. In the context of reliability modeling, extreme value distributions for the minimum are frequently encountered, e.g., if a system consists of *n* identical components in series, and the system fails when the first of these components fails, then system

failure times are the minimum of *n* random component failure times. Extreme value theory says that, independent of the choice of component model, the system model will approach a Weibull as *n* becomes large. The same reasoning can also be applied at a component level, if the component failure occurs when the first of many similar competing failure processes reaches a critical level.

The distribution often referred to as the extreme value distribution is the limiting distribution of the minimum of a large number of unbounded identically distributed random variables. The PDF and CDF are given by:

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}}, \quad -\infty < x < \infty, \beta > 0$$

$$F(x) = 1 - e^{-e^{\frac{x-\mu}{\beta}}}, \quad -\infty < x < \infty, \beta > 0$$

If the *x* values are bounded below (as is the case with times of failure) then the limiting distribution is the Weibull. PDF shapes for the (minimum) extreme value distribution are illustrated in Figure 12.

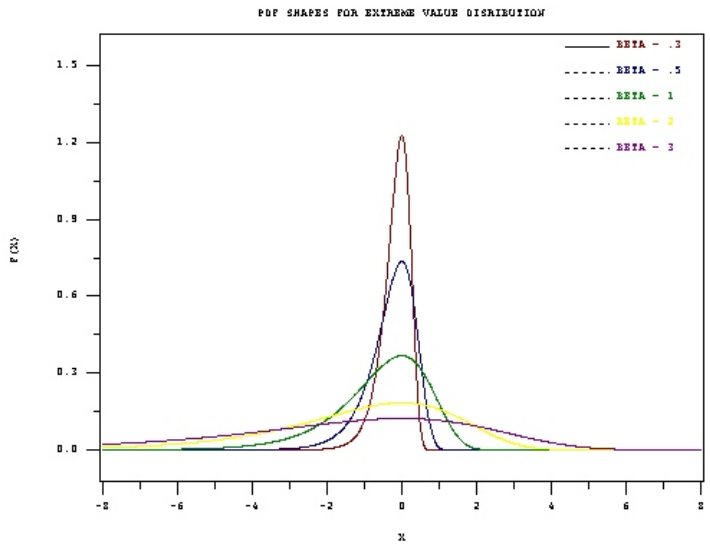



**Figure 12.** Illustrations of various Extreme Value distributions.

The Weibull distribution and the extreme value distribution have a useful mathematical relationship. If $t_1, t_2, \ldots, t_n$ are a sample of random times of failure from a Weibull distribution, then $\ln t_1, \ln t_2, \ldots, \ln t_n$ are random observations from the extreme value distribution. In other words, the natural log of a Weibull random time is an extreme value random observation. Because of this relationship, computer programs and graph papers designed for the extreme value distribution can be used to analyze Weibull data which is similar to using normal distribution programs to analyze lognormal data, after first taking natural logarithms of the data points.

**Lognormal distribution.** The lognormal life distribution, like the Weibull, is a very flexible model that can empirically fit many types of failure data. The two parameter form has parameters $\sigma$ = the shape parameter and $T_{50}$ = the median (a scale parameter). If time to failure, $t_f$, has a lognormal distribution, then the (natural) logarithm of time to failure has a normal distribution with mean $\mu = \ln T_{50}$ and standard deviation $\sigma$. This makes lognormal data convenient to work with; just take natural logarithms of all the failure times and censoring times and analyze the resulting normal data. Later on, convert back to real time and lognormal parameters using $\sigma$ as the lognormal shape and $T_{50} = e^{\mu}$ as the (median) scale parameter. Below is a summary of the key formulas for the lognormal.

$$\text{PDF: } f(t) = \frac{1}{\sigma t \sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln T_{50})^2}$$

$$\text{CDF: } F(T) = \int_0^T \frac{1}{\sigma t \sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln T_{50})^2} dt = \Phi\left(\frac{\ln t - \ln T_{50}}{\sigma}\right)$$

with $\Phi(z)$ denoting the standard Normal CDF

$$\text{RELIABILITY: } R(T) = 1 - F(t)$$

$$\text{FAILURE RATE: } h(t) = \frac{f(t)}{R(t)}$$

$$\text{MEAN: } T_{50} e^{\frac{\sigma^2}{2}}$$

$$\text{MEDIAN: } T_{50}$$

$$\text{VARIANCE: } T_{50}^2 e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$$

A more general 3-parameter form of the lognormal includes an additional waiting time parameter $\theta$ (sometimes called a shift or location parameter). The formulas for the 3-parameter

lognormal are easily obtained from the above formulas by replacing *t* by (*t* - $\theta$) wherever *t* appears. No failure can occur before $\theta$ hours, so the time scale starts at $\theta$ and not 0. If a shift parameter $\theta$ is known (based, perhaps, on the physics of the failure mode), then all you have to do is subtract $\theta$ from all the observed failure times and/or readout times and analyze the resulting shifted data with a 2-parameter lognormal.

Examples of lognormal PDF and failure rate plots are shown in Figure 13 and Figure 14, respectively. Observe that lognormal shapes for small sigmas are very similar to Weibull shapes when the shape parameter $\gamma$ is large and large sigmas give plots similar to small Weibull $\gamma$'s. Both distributions are very flexible and it is often difficult to choose which to use based on empirical fits to small samples of (possibly censored) data.

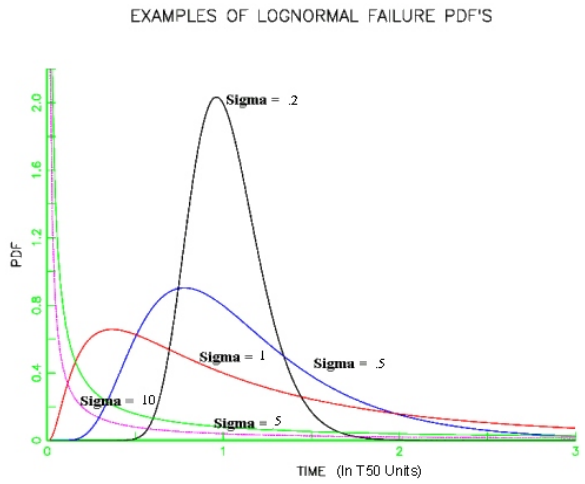



**Figure 13.** Illustrations of the lognormal distribution PDFs.

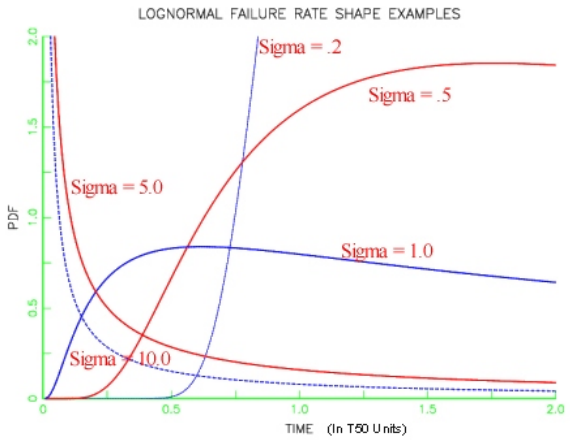
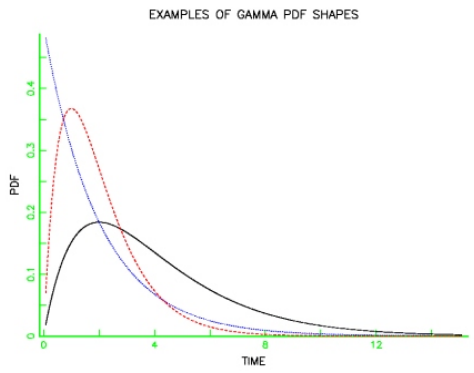



**Figure 14.** Illustrations of lognormal failure rates.

As suggested by the preceding plots, lognormal PDF and failure rate shapes are flexible enough to make the lognormal a very useful empirical model. Lognormal models can be theoretically derived under assumptions matching many common failure processes, which does not mean that the lognormal is always the correct model for these mechanisms, but it does perhaps explain why it has been empirically successful in so many cases.

**Gamma distribution.** In the literature, the gamma distribution is commonly presented in one of two forms, and different authors use different symbols for the shape and scale parameters. Below we show three ways of writing the gamma, with $a = \alpha = \gamma$, the "shape" parameter, and $b = 1/\beta$, the scale parameter. The exponential is a special case of the gamma when $a = 1$, the gamma reduces to an exponential distribution with $b = \lambda$. Another well-known statistical distribution, the Chi-Square, is also a special case of the gamma, where a Chi-Square distribution with $n$ degrees of freedom is the same as a gamma with $a = n/2$ and $b = 0.5$ (or $\beta = 2$). Figure 15 illustrates of gamma PDFs, CDFs, and failure rate shapes.



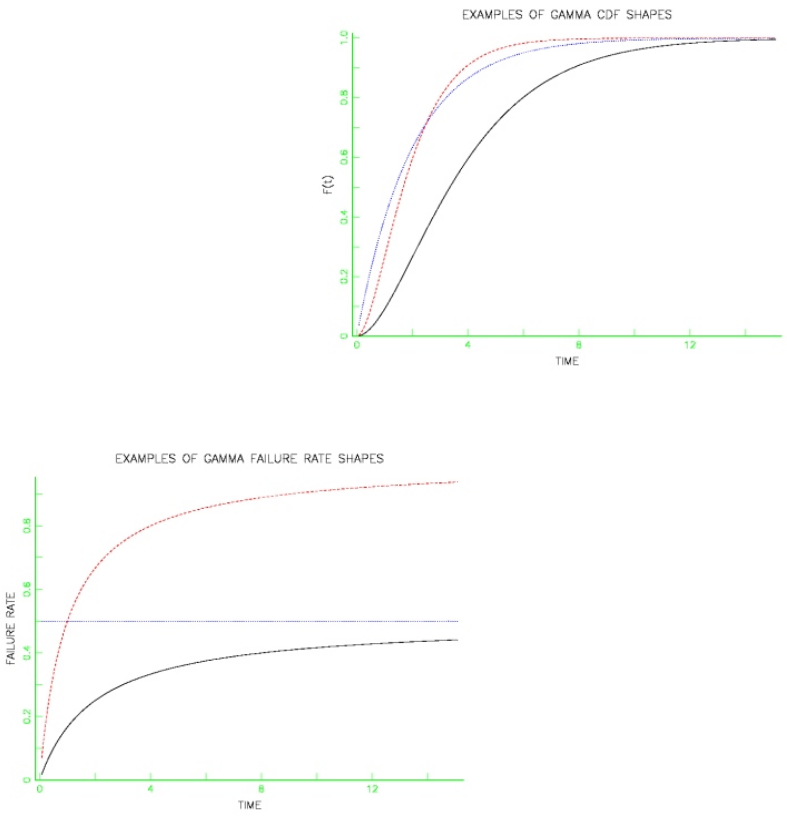



Figure 15. Illustrations of PDFs (top left), CDFs (top right), and failure rates (bottom) for gamma distribution.

The gamma distribution is commonly used for Bayesian reliability analysis, since it is a flexible life distribution model and frequently provides a good fit for failure data.

**Proportional hazards model.** The proportional hazards model is often used in survival analysis, but infrequently with engineering data. Cox's proportional hazards model (Cox 1972) has been used primarily to evaluate survival when secondary variables are likely exerting effects on the system. Its strength lies in its ability to model and test many inferences about survival without making any specific assumptions about the form of the life distribution model.

Proportional hazards model is based on an assumption that there are one or more explanatory variables (continuous, categorical, or binary) that affect lifetime. The hazard rate for a nominal (or baseline) set $z_0 = (x_0, y_0, \ldots)$ of these variables be given by $h_0(t)$, with $h_0(t)$ denoting legitimate hazard function (failure rate) for some unspecified life distribution model. The proportional hazard model assumes changing a stress variable (or explanatory variable) has the effect of multiplying the hazard rate by a constant. The proportional hazards model assumes we can write the changed hazard function for a new value of $z$ as:

$$h_z(t) = g(z)h_0(t)$$

In other words, changing $z$, the explanatory variable vector, results in a new hazard function that is proportional to the nominal hazard function, and the proportionality constant is a function of $z$, $g(z)$, independent of the time variable $t$. A common and useful form for $f(z)$ is the log-linear model which has the equation: $g(x) = e^{ax}$ for one variable, $g(x,y) = e^{ax + by}$ for two variables.

The proportional hazards model is equivalent to the acceleration factor concept if and only if the life distribution model is a Weibull (which includes the exponential model, as a special case). For a Weibull with shape parameter $\gamma$, and an acceleration factor $AF$ between nominal use fail time $t_0$ and high stress fail time $t_s$ (with $t_0 = AFt_s$) we have $g(s) = AF^{\gamma}$. In other words, $h_s(t) = {}_{AFh0}(t)$. Under a log-linear model assumption for $g(z)$ without any further assumptions about the life

distribution model, it is possible to analyze experimental data and compute maximum likelihood estimates and use likelihood ratio tests to determine which explanatory variables are highly significant. More details on the theory and applications of the proportional hazards model may be found in Kalbfleisch and Prentice (2002) and Lawless (2003).

**Data limitations and failure analysis.** The more reliable a system is, the more difficult it is to gather failure data to predict its failure. Two closely related problems that are typical of reliability data and related types of statistical data encountered in invasion biology (e.g., species distribution data where "species not found" may be misinterpreted as "species absence"). First, data are generally censored (e.g., when an observation period ends, but not all units have failed). Failure data may be "right censored" or "left censored," depending on the way the data were collected (e.g., testing period of fixed time or fixed number of failures defines testing period, respectively; see Kalbfleisch and Prentice (2002), Lawless (2003), and Meeker and Escobar (1998) for a comprehensive review of the role of data censoring in limiting failure analysis). Data may also be "multicensored," since different studies may record observations differently for identical systems being considered, e.g., failure may be identified as a run-time endpoint, if the unit did not fail while under observation, or failure may be identified as an exact failure time, or failure may be identified as an interval of time during which the unit failed. Many statistical methods can be used to fit models and estimate failure rates even with censored data (e.g., probability plotting, maximum likelihood estimation; see Meeker and Escobar 1998).

Second, observed failures may be few in number or completely absent, if the system is highly reliable or inadequately sampled. Independently or in combination, these data limitations influence the uncertainty associated with analyzing failure data, particularly as those tools apply to evaluations for risk. Although serving as sources of uncertainty, solutions to these data limitations generally mean making additional assumptions in developing risk scenarios and using "best guess" models for characterizing failure events and their role in modifying risks (e.g., increasing or decreasing risk estimates).

**Distinguishing Failure Modes.** Failures are a generally a coarse measurement endpoint, and may result from several different failure modes (e.g., root cause of failure may differ from one occurrence to the next), and in the current investigation the discrimination between species invasions linked to interbasin water transfers and those linked with other pathways are considered within the context of competing risks in Section 4.

In general, the analysis of competing risks, regardless of the focus being on biological and ecological systems, or on engineering systems, revolves about failure mechanisms that are assumed to be independent, with the first "failure mode" that occurs causes the system to fail. For example, if a species invasion is considered a failure, then each of $k$ different failure modes or ways a failure can occur are competing (e.g., for species invasions, different pathways may be interpreted as different failure modes), and underlying each failure mode is a failure mechanism (for a given pathway, each mode will have one to many different failure mechanisms).

In evaluating competing risks, a system's reliability is considered as a "build up" model, based on evaluations of the reliability of each failure mode. Three assumptions are generally specified in such an analysis of competing risks: (1) each failure mechanism leading to a particular type of failure (i.e., failure mode) proceeds independently of every other one at least until a failure occurs; (2) a failure event occurs when the first of all the competing failure mechanisms reaches a failed state; and (3) each of the $k$ failure modes has a known life distribution model $F_i(t)$.

Quantitatively, the competing risk model is best applied when all three assumptions hold. If $R_c(t)$, $F_c(t)$, and $h_c(t)$ denote the reliability, CDF and failure rate for the component, respectively, and $R_i(t)$, $F_i(t)$and $h_i(t)$ are the reliability, CDF and failure rate for the $i$-th failure mode, respectively, then the competing risk model formulas are:

$$R_c(t) = \prod_{i=1}^{k} R_i(t)$$

$$F_c(t) = 1 - \prod_{i=1}^{k} \{1 - F_i(t)\}$$

$$h_c(t) = \sum_{i=1}^{k} h_i(t)$$

**Multiply reliabilities and add failure rates.** For evaluating competing risks, consider all failure mechanisms are racing to see which can reach failure first, e.g., which competing risk is most likely to yield a species invasion. If the failure mechanisms are assumed independent, then the component reliability is the product of the failure mode reliabilities and the component failure rate is the sum of the failure rates. This algorithm holds for any arbitrary life distribution model, as long as "independence" and "first mechanism failure causes the component to fail" assumptions are not violated.

Alternative "rules" associated with calculating risks for different types of systems are briefly reviewed below.

**Failures in series models.** The series model is used to go from individual components to the entire system, assuming the system fails when the first component fails and all components fail or survive independently of one another. The series model is a "build up" model where components are constructed to yield sub-assemblies and systems, and only applies to non-replaceable populations (or first failures of populations of systems). The assumptions and formulas for the series model are identical to those for the competing risk model, with the *k* failure modes within a component replaced by the *n* components within a system. In Figure 16, the entire system has *n* components in series, and the system operates when all components function or fails when at least one component fails. Each component is independent, but failure in one component means the system fais. Simplified, the system of 5 components in series may be represented by an equivalent

system (as far as reliability is concerned) with only one component.

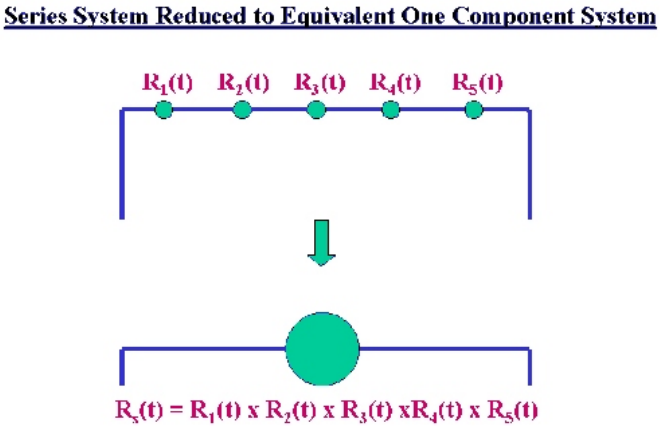



**Figure 16.** Illustration of a series system.

**Failures in parallel or redundant systems.** In parallel systems, all n components that make up the system operate independently and the system works as long as at least one component still works. Parallel systems are the opposite of a system operating in series in which the first component failure causes the system to fail. In a parallel system, all the components have to fail before the system fails. If there are *n* components, any (*n-1*) of them may be considered redundant to the remaining one (even if the components are all different). When the system is turned on, all the components operate until they fail. The system fails at the time of the last component failure.

In contrast to a system operating in series, the assumptions for a parallel model are: (1) all components operate independently of one another, as far as reliability is concerned; (2) the system operates as long as at least one component is still operating, and system failure only occurs at the time of the last component failure; and (3) the CDF for each component is known. For a system operating in parallel, the CDF $F_s(t)$ for the system is just the product of the CDF's $F_i(t)$ for the components or

$$F_s(t) = \prod_{i=1}^{n} F_i(t)$$

$R_s(t)$ and $h_s(t)$ can be evaluated using basic definitions, once we have $F_s(t)$. Figure 17 represents a parallel system with 5 components and the (reliability) equivalent 1 component system with a CDF $F_s$ equal to the product of the 5 component CDFs.

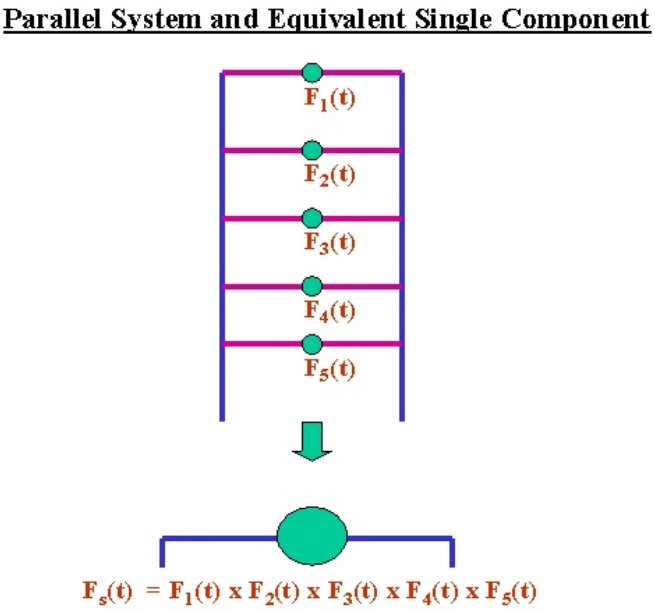



**Figure 17.** An illustration of a parallel system.

**R out of N model.** An "*r* out of *n*" system survives when at least *r* of its components are working (any *r*). An "*r* out of *n*" system contains includes the series system and the parallel system as special cases. The system has *n* components that operate or fail independently of one another and as long as at least *r* of these components (any *r*) survive, the system survives. System failure occurs when the [*n* - *(r+1)*] component failure occurs. When $r = n$, the *r* out of *n* model reduces to the series model, and when $r = 1$, the *r* out of *n* model becomes the parallel model. When all the components of the system (1) are identical and have the identical reliability function *R(t)*; (2) operate independently of one another (as far as failure is concerned); (3) the system can survive any (*n-r*) of the components failing, but fails upon the [(*n* - *(r+1)*] component failure, then system reliability is given by adding the probability of exactly *r* components surviving to time *t* to the probability of exactly (*r+1*) components surviving, and so on up to the probability of all

components surviving to time *t*. These are binomial probabilities (with $p = R(t)$), so the system reliability is given by:

$$R_s(t) = \sum_{i=r}^{n} \binom{n}{i} [R(t)]^i [1 - R(t)]^{n-i}$$

If all the components are not identical, then $R_s(t)$ would be the sum of probabilities evaluated for all possible terms that could be formed by picking at least *r* survivors and the corresponding failures. The probability for each term is evaluated as a product of *R(t)*'s and *F(t)*'s. For example, for $n = 4$ and $r = 2$, the system reliability would be (abbreviating the notation for *R(t)* and *F(t)* by using only *R* and *F)*:

$$R_s = R_1R_2F_3F_4 + R_1R_3F_2F_4 + R_1R_4F_2F_3 + R_2R_3F_1F_4 + R_2R_4F_1F_3 + R_3R_4F_1F2 + R_1R_2R_3F_4 + R_1R_3R_4F_2 + R_1R_2R_4F_3 + R_2R_3R_4F_1 + R_1R_2R_3R_4$$

**Complex systems.** For complex systems, reliability can be evaluated by successive applications of series and parallel models. Many complex systems can be diagramed as combinations of series components, parallel components, and *R* out of *N* components (see, e.g., Miller and Escobar 1998; Thompson 2000; Borgelt and Kruse 2002; Huzurbazar 2005; Banerjee et al 2004; Salthe 1985; Puccia and Levins 1985). While many engineering analyses, and indeed many evaluations of ecological systems, seek to reduce their complexity to "equivalent" simple systems, many systems with marked interdependence and interconnectedness, or with systems characterized by complicated operational logic structure, alternative tools such as event trees, Boolean representations, coherent structures, cut sets and decompositions may be involved. The reader is referred to those authors listed above for more comprehensive treatment of complex systems analysis[1].

[1] Graphics and excerpts from NIST/SEMATECH (2004), e-Handbook of Statistical Methods (available at http://www.itl.nist.gov/div898/handbook/) have been relied upon for peer-reviewed technical summaries incorporated into this overview of reliability analysis in this appendix.

**References**

Balakrishnan, N., and V.B. Nevzorov, 2003, A primer on statistical distributions, John Wiley & Sons, Inc., New York, 305pp.

Banerjee, S., B.P. Carlin, and A.E. Gelfand, 2004, Hierarchical modeling and analysis of spatial data, Chapman & Hall/CRC, Boca Raton, Florida, 452pp.

Barlow, R.E., 1998, Engineering reliability, American Statistical Society, Alexandria, Virginia, and Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 199pp.

Bedford, T., and R. Cooke, 2001, Probabilistic risk analysis, Cambridge University Press, Cambridge, UK, 393pp.

Blischke, W.R., and D.N. Parbhakar Murthy, 2000, Reliability, John Wiley & Sons, Inc., New York, 812pp.

Borgelt, C., and R. Kruse, 2002, Graphical models, John Wiley & Sons, Ltd., Chichester, UK, 358pp.

Castillo, E., A.S. Hadi, N. Balakrishnan, and J.M. Sarabia, 2005, Extreme value and related models with applications in engineering and science, John Wiley & Sons, Inc., New York, 362.

Caswell, H., 2001, Matrix population models, Second Edition, Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 722pp.

Cox, D. R., 1972, Regression models with life tables, J. Royal Statistical Society, 34, 187-220.

Evans, M.; Hastings, N.; and Peacock, B., 2000, Statistical Distributions, 3rd edition, Wiley &

Sons, Inc., New York, 221pp.

FISRWG, 1998, Stream Corridor Restoration: Principles, Processes, and Practices. By the Federal Interagency Stream Restoration Working Group (FISRWG; 15 Federal agencies of the US government, GPO Item No. 0120-A; SuDocs No. A 57.6/2:EN 3/PT.653, ISBN-0-934213-59-3.

Gumbel, E.J., 1958, Statistics of extremes, Columbia University Press, New York, 375pp.

Haight, F.A., 1967, Handbook of the Poisson distribution, John Wiley & Sons, Inc., New York, 168pp.

Huzurbazar, A.V., 2005, Flowgraph models for multistate time-to-event data, John Wiley & Sons, Inc., New York, 270pp.

Jordan, III, W.R., M.E. Gilpin, and J.D. Aber (editors), 1987, Restoration ecology, Cambridge University Press, Cambridge, UK, 342pp.

Kalbfleisch, J.D., and R.L. Prentice, 2002, The statistical analysis of failure time data, John Wiley & Sons, Inc., New York, 439pp.

Kenney, J. F. and Keeping, E. S., 1951, Mathematics of Statistics, Pt. 2, Second Edition, Van Nostrand, Princeton, New Jersey, 202pp.

Lawless, J.F., 2003, Statistical models and methods for lifetime data, John Wiley & Sons, Inc., New York, 630pp.

Manci, Karen M. 1989. Riparian ecosystem creation and restoration: A literature summary. U.S. Fish and Wildlife Service Biological Report 89(20):1-59, Northern Prairie Wildlife Research

Center Home Page, Jamestown, North Dakota, available and last accessed at http://www.npwrc.usgs.gov/resource/literatr/ripareco/ripareco.htm (Version 16JUL97).

Meeker, W.Q., and L.A. Escobar, 1998, Statistical methods for reliability data, John Wiley & Sons, Inc., New York, 680pp.

National Institute of Standards and Technology (NIST)/SEMATECH, 2004, e-Handbook of Statistical Methods, available at http://www.itl.nist.gov/div898/handbook/, last accessed November 11, 2004.

Papoulis, A., 1984, Probability, Random Variables, and Stochastic Processes, 2nd edition, McGraw-Hill Publishers, Inc., New York, 576pp.

Patel, J. K. and Read, C. B., 1982, Handbook of the Normal Distribution, Dekker, New York, 351pp.

Pfeiffer, P. E. and Schum, D. A., 1973, Introduction to Applied Probability, Academic Press, Inc., 403pp.

Puccia, C.J., and R. Levins, 1985, Qualitative modeling of complex systems, Harvard University Press, Cambridge, Massachusetts, 259pp.

Salthe, S.N., 1985, Evolving hierarchical systems, Columbia University Press, New York, 343pp.

Thompson, J.R., 2000, Simulation, John Wiley & Sons, Inc., New York, 297pp.

Weisstein, E.W.. 1999, MathWorld – A Wolfram Web Resource. Access on World-wide Web at http://mathworld.wolfram.com. © 1999 CRC Press LLC, © 1999-2004 Wolfram Research, Inc.

Whittaker, E. T. and Robinson, G., 1967, Normal Frequency Distribution, In The Calculus of Observations: A Treatise on Numerical Mathematics, 4th Edition, Dover Books, New York, pp. 164-208, 1967.