Methods for Assessing Distributional Assumptions in One and Two Sample Problems

Vijayan N. Nair and *Anne E. Freeny*
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

This is an expanded version of a chapter in the edited volume *Probabilistic and Statistical Methods in the Physical Sciences*, which is part of the series on *Methods in Experimental Physics.* The volume is edited by J. Stanford and S. Vardeman and is to be published by Academic Press.

Efficient methods of estimation for different parametric models are discussed in other chapters in this volume. The use of these methods, however, requires knowing which parametric model is appropriate for a given situation. Sometimes, prior knowledge or subject matter expertise can be used to select the right model. More often, however, such information is unavailable or incomplete, and empirical methods are needed for assessing goodness-of-fit and for identifying appropriate models. This is the topic of the present paper. Graphical and formal goodness-of-fit methods are considered for situations with complete as well as various types of censored data. The focus here is on assessing distributional assumptions in one and two sample problems. Extensions to more complex situations are discussed briefly.

Methods for Assessing Distributional Assumptions in One and Two Sample Problems

Vijayan N. Nair and *Anne E. Freeny*
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

1 Introduction

1.1 Motivation

There are a variety of reasons why parametric models, and hence methods for assessing their fit, are needed.

- Nonparametric methods for estimating the underlying distribution are discussed in Section 4. The nonparametric estimators, however, can be highly variable when the sample size is small or if there is a lot of censoring. For more efficient inference, estimators based on a parametric model must be used.

- Parametric models are also used to obtain estimates or predictions beyond the range of the observed data. In many applications, interest centers on quantities associated with the tails of distributions. The nonparametric estimator cannot be extrapolated to yield estimates in extreme tails. If there is a high degree of censoring, even quantities such as the median may not be estimable using nonparametric methods. We will see examples of this later. Thus, one typically fits a parametric model to the observed data and extrapolates the smooth curve to get the required estimates/predictions. Another kind of extrapolation arises in regression problems where a parametric regression model is used for the functional relationship between the response (dependent) variable and explanatory variables. The fitted regression model is extrapolated to predict responses at different settings of the explanatory variables.

- Estimation is often not the primary goal of scientific investigations. It is more important to gain insights into the underlying behavior of the process or system. In reliability applications, for example, reliability engineers are interested in the underlying hazard rate, whether it is constant, decreasing, increasing or U-shaped. In studies on emission of radioactive particles, a question of interest might be whether the rate of decay is constant or varies with time. In experimental design and regression problems, the primary goal is to understand the relationship between the response and explanatory variables. These types of information can be obtained by fitting a parsimonious parametric model to the data and interpreting the properties of the fitted model.

We will discuss several real examples where some of the above modeling issues arise.

1.2 Scope

The focus here is on assessing distributional assumptions in one and two sample problems. In a one-sample situation, we have data from a population of interest, and the goal is to model the underlying cumulative distribution function (*cdf*) $F(x)$. The simplest goodness-of-fit problem in this situation is assessing the null hypothesis

$$F(x) = F_0(x) \quad \text{for all} \quad x \ , \tag{1}$$

where $F_0(x)$ is a completely specified *cdf*. This is called a simple hypothesis because $F_0(x)$ is completely known. A more common problem is assessing whether

$$F(x) = F_0(x; \boldsymbol{\theta}) \quad \text{for some} \quad \boldsymbol{\theta} \ , \tag{2}$$

where $F_0(x; \boldsymbol{\theta})$ is a specified parametric family, such as the normal, Weibull or gamma. This is called a composite hypothesis because the *cdf* $F(x)$ is not completely specified under the null hypothesis. Of particular interest are parametric location-scale families

$$F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) \ , \tag{3}$$

where μ is a location parameter and σ is a scale parameter. Examples are the normal location-scale family and the exponential scale family.

In a two-sample situation, we have data from two populations characterized by the two *cdf*s $F(x)$ and $G(x)$. We will consider testing the simple hypothesis

$$F(x) = G(x) \quad \text{for all} \quad x \ , \tag{4}$$

as well as methods for assessing the common shape of the two *cdf*s.

Two types of techniques, graphical methods and formal goodness-of-fit tests, will be discussed. Situations with complete data as well as those with censored data will be considered. Extensions to more complex problems, such as regression problems and data with selection-bias, will be briefly mentioned.

Graphical methods are attractive because they allow one to visually examine the data and look for systematic patterns as well as aberrant behavior. If the specified model does not fit the data, plots will often suggest what alternative models to consider. However, they are informal in nature and hence the interpretations are often subjective. Goodness-of-fit tests, on the other hand, provide formal criteria for accepting or rejecting a hypothesis, but if a hypothesis is rejected, they provide little information on what alternative models might be suitable. We will show how the advantages of both methods can be combined by using graphical methods with confidence bands.

1.3 Organization

In Section 2, we describe several real data sets and use them to motivate different types of data and goodness-of-fit problems. These data sets will be used in subsequent sections to illustrate the various techniques. The well-known chi-squared tests are discussed in Section 3. These tests are, however, best suited for discrete distributions or grouped data whereas the emphasis of the chapter is on modeling data from continuous distributions. Sections 4–6 deal with the one-sample problem. The quantile-quantile (Q-Q) plotting technique is discussed in Section 4. Simultaneous confidence bands for use with Q-Q plots are developed in the next section. Formal test procedures for testing the simple hypothesis (1) and the composite hypothesis (2) are discussed in Section 6. The two-sample problem is considered in Section 7. Section 8 deals with some miscellaneous issues.

2 Data

2.1 Types of Data

The paper deals with methods for *complete* data as well as for different types of *censored* data. By complete data, we refer to situations where we observe a sample $X_1, \ldots, X_n$ independently and identically distributed (iid) according to $F(x)$. Of course, the X_i's are never observed "exactly" in practice and are always subject to some form of rounding. By use of the term "complete" data, we implicitly assume that the level of coarseness can be ignored for practical purposes. Often, data are also subject to various forms of censoring. We will consider several types of censored data.

Right censoring refers to situations where we do not know the actual value of an observation but know only that it exceeded some threshold. To define it formally, let $X_1^0, \ldots, X_n^0$ be iid from $F(x)$. Associated with each X_i^0 is a censoring variable C_i. We observe only $X_i = \min(X_i^0, C_i)$ and $\delta_i = I[X_i \leq C_i]$, $i = 1, \ldots, n$ where $I[A]$ is the indicator function of the event A. Thus, $\delta_i = 0$ means that the observation has been censored to the right, i.e., the true observation is larger than C_i but its actual value is unknown. The case where $C_i \equiv C$, a pre-specified, fixed value, is called *Type I censoring.* In some reliability and medical studies, the investigation is terminated as soon as $r < n$ "failures" have occurred. In this case, $C_i \equiv X_{(r)}$, the rth order-statistic, and is random. This is called *Type II censoring.* More generally, the C_i's vary from observation to observation, and the observations are said to be *multiply right censored.* One important case of this occurs when the C_i's are themselves random variables. The data are then said to be *randomly right censored.*

Left censoring occurs when we know that an observation is smaller than a known threshold but its actual value is unknown. It is common in areas that involve measurements, such as astronomy, metrology and environmental studies. If a measuring instrument cannot record observations below a certain threshold, the observations will be left-censored. The problem also arises in reliability and medical studies where some units have already failed before they enter the study. Left censoring can be defined formally as in the previous paragraph but with $X_i = \max(X_i^0, C_i)$ and $\delta_i = I[X_i^0 \geq C_i]$. The special cases of Type I, II and random censoring are defined similarly.

Sometimes data are grouped into intervals and we only know the number of observations that fall in each interval. Specifically, let $T_0 < T_1 < \cdots < T_J$ denote the end points of the J intervals. We observe only n_j, the number of X_i's that fall in the interval $(T_{j-1}, T_j]$, for $j = 1, \ldots, J$. Such data are called *grouped* or *interval-censored* data. Interval-censoring arises in reliability and medical studies where units are inspected periodically over time, and at the inspection times one records whether a unit is "alive" or "dead." Another example is particle size studies where the particles are "sieved" according to particle diameters, and the sieved particles are then weighed.

2.2 Data Sets

We describe some real data sets and use them to illustrate the different types of data we have discussed thus far. These data sets will also be used in later sections to illustrate the methods that we develop.

Data on Emission of α-Particles: This is from a well-known experiment by Rutherford and Geiger (1910) on the emission of α-particles from a radioactive source. The data in Table 1 are n_j = the number of intervals (of a given duration) in which j α-particles were observed, $j = 0, 1, \ldots, 11$. If emissions of α-particles follow a homogeneous Poisson process, the data in Table 1 come a Poisson distribution. We will describe methods for assessing this assumption empirically. This is the only example where we consider a discrete distribution. Most of our attention will be focused on modeling data from continuous distributions.

Table 1. Data on Emission of α-Particles
(number of intervals in which j α-particles are observed)

j	0	1	2	3	4	5	6	7	8	9	10	11
O_j	57	203	383	525	532	408	273	139	49	27	10	6

Data on Failure Times of a Mechanical Device: This data set was obtained from a reliability study of a mechanical device, a type of switch. The data in Table 2 are times to failure (in millions of operations) of 37 devices. There were two possible modes of failure – essentially two springs – for these devices. When a device failed, its mode of failure was noted. These are indicated by the codes A and B in Table 2. The two springs were identical in construction but were subjected to different stress levels, so the life distributions of the two failure modes are likely to be different. For illustrative purposes, we consider here only part of the data from the study. See Nair (1984) for the complete data set.

This data set is a typical example of a competing-risks problem where a system fails due to one or more competing causes, and one observes only the time to failure of the system and the corresponding failure mode (compo-

Table 2. Data on Failure Times of a Mechanical Device
(in millions of operations)

Time	Mode	Time	Mode
1.151	B	2.116	B
1.170	B	2.119	B
1.248	B	2.135	A
1.331	B	2.197	A
1.381	B	2.199	B
1.499	A	2.227	A
1.508	B	2.250	B
1.534	B	2.254	A
1.577	B	2.261	B
1.584	B	2.349	B
1.667	A	2.369	A
1.695	A	2.547	A
1.710	A	2.548	A
1.955	B	2.738	B
1.965	A	2.794	A
2.012	B	2.910	A
2.051	B	3.015	A
2.076	B	3.017	A
2.109	A		

nent). In Table 2, when a device failed due to mode A (B), the failure time of mode B (A) was *randomly right-censored.* Based on these censored data, we have to make inferences about the distributions of the device and of the two failure modes. In particular, we want a parsimonious parametric model for the reliability of the components and the system. For example, if the two failure modes can be modeled by a Weibull distribution with the same shape parameter, then the failure time of the device, which is the minimum of the failure times of the two modes, also follows a Weibull model with the same shape parameter. We will discuss methods for assessing the adequacy of this model.

Data from an Accelerated Life Test of Printed Circuit Boards: This data set is from an accelerated life test (ALT) of printed circuit boards conducted at AT&T. The goal of the original experiment was to determine the effects of temperature, relative humidity (RH), and electric field on the reliability of printed circuit boards. The problem studied in this experiment was the formation of conductive anodic filaments between copper plated holes in the printed circuit board. In the ALT, failure was defined as the time taken for the resistance between positively and negatively charged columns (in the board) to drop low enough to cause a short circuit. Table 3 gives the data for two of the accelerated test conditions used in this experiment, corresponding to the levels of $RH1 = 62.8\%$ and $RH2 = 85\%$. See Meeker and Luvalle (1992) for additional data and details about this experiment.

The data in this example were subject to *interval-censoring.* The circuit boards were inspected periodically, so it was possible to only record whether a board had failed or not at the time of inspection. The inspection times are indicated as the lower (L) and upper (U) values in the table and N, the number of boards that failed in each interval, are also given.

The experiment at the stress condition $RH1$ was terminated at 3067 hours, and provides an example of *Type I right censoring.* About 15% of the circuit boards were censored at this point. In fact, for the data at a lower stress condition (see Meeker and Luvalle, 1992), almost 70% of the circuit boards were censored. Hence, even the median of this distribution cannot be estimated nonparametrically and parametric models must be used.

The goal in ALT's is to use the data at accelerated stress conditions to make inferences about reliability at normal stress conditions. To do this, one has to extrapolate across stress conditions using a model that relates the failure times as a function of stress. Most of the ALT regression models in the literature assume that the distributions of the log-failure times at different stress conditions follow a location model. That is, if $F_k(x)$ denotes the *cdf* of the log-failure time at stress level k, then

$$F_k(x) = F_0(x - \mu_k) \ , \tag{5}$$

for some (location) parameter μ_k and some *cdf* $F_0(x)$ for $k = 1, \ldots, K$. An ALT regression model is used to describe how the parameters μ_k vary as a function of the stress conditions (see Nelson, 1990). A first step in analyzing data from ALT experiments is, therefore, checking that the location model (5) holds. We will consider methods for assessing this assumption.

Table 3. Data on an Accelerated Life Test of Printed Circuit Boards (interval times to failure in hours)

$RH1: 62.8\%$						$RH2: 85\%$		
L	U	N	L	U	N	L	U	N
112	116	1	401	405	1	14	18	1
128	132	1	461	465	1	18	22	1
136	140	1	481	485	1	30	34	8
148	152	1	533	537	1	42	46	3
173	177	1	537	541	2	46	50	3
185	189	1	561	565	1	50	54	3
189	193	1	565	569	1	58	62	4
201	205	1	569	573	1	62	66	1
237	241	1	637	641	1	66	70	5
261	265	1	688	692	2	70	74	2
269	273	1	760	764	2	74	78	2
281	285	1	796	800	1	78	82	1
285	289	2	916	920	1	82	86	4
289	293	2	940	944	1	86	90	8
297	301	1	964	968	1	90	94	4
301	305	1	1014	1026	1	94	98	5
317	321	1	1086	1098	1	98	102	5
321	325	1	1122	1134	1	106	110	1
337	341	3	1206	1218	1	110	114	2
341	345	1	1446	1458	1	114	118	1
353	357	1	1650	1662	1	122	126	1
361	365	1	1974	1986	1	130	134	1
365	369	1	2046	2058	1	134	138	1
385	389	1	2142	2154	1	138	142	1
397	401	1	3054	3066	1	142	146	2
			>3067		11			

3 Chi-Squared Tests

The simplest, and perhaps most versatile, goodness-of-fit procedure is the chi-squared test introduced by Pearson in 1900. The general form of the test statistic is

$$X^2 = \sum_{\{\text{all cells}\}} \frac{(O-E)^2}{E} \tag{6}$$

where the O is the actual number of observations and E is the "expected" number of observations in a given cell. To see what these are in concrete terms, let us consider specific testing problems.

3.1 Simple Hypothesis

We discuss first the testing problem (1). Suppose the actual observations $X_1, \ldots, X_n$, which are independent and identically distributed (iid) from $F(x)$, have been grouped into J cells $(T_{j-1}, T_j]$, $j = 1, \ldots, J$. Then O_j is the number of X_i's that fall in the jth cell, $j = 1, \ldots, J$. The "expected" number of observations in the jth cell, E_j, is given by $E_j = np_{0j}$ where p_{0j} is the probability content of the jth cell under the null hypothesis, $F_0(T_j) - F_0(T_{j-1})$.

The test is carried out by computing the value of X^2 and comparing it to the quantiles of the distribution of X^2 under the null hypothesis. The hypothesis (1) is rejected at level α if X^2 exceeds the $(1-\alpha)$th quantile; otherwise it is accepted. The exact distribution of X^2 under the null hypothesis is complicated. In large samples, however, this distribution can be well approximated by a χ^2 distribution with $(J-1)$ degrees of freedom. Since quantiles of χ^2 distributions are widely available, it is easy to implement this test.

3.2 Composite Hypothesis

Next consider the composite hypothesis (2) that $F(x) = F_0(x; \boldsymbol{\theta})$ where $F_0(\cdot)$ is a specified parametric family and $\boldsymbol{\theta}$ is some unknown p-dimensional parameter. The O_j's are again defined as the number of observations falling in the jth cell. To compute the E_j's, however, we now have to estimate the unknown $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}}$ be an estimate, and $\widehat{p}_{0j}$ be the probability content of the jth cell under $F_0(x; \widehat{\boldsymbol{\theta}})$. Then $E_j = n\widehat{p}_{0j}$. By substituting these values of O_j and E_j in (6), we get the chi-squared statistic for testing the composite hypothesis (2). The limiting distribution of this test statistic depends on the method of estimation used to obtain $\widehat{\boldsymbol{\theta}}$. Fisher (1924) showed that if maximum likelihood estimation based on the grouped data is used, the large-sample distribution is again χ^2 but now with $(J-p-1)$ degrees of freedom, to account for the p unknown parameters that have been estimated. The result holds true if other asymptotically equivalent methods of estimation are used. Use of maximum likelihood estimators based on the original X_i's leads to a more complicated distribution for the test statistic, however.

Example: We use the data in Table 1 on emission of α-particles to illustrate the method. Table 1 gives O_j = the number of times j emissions occurred in an interval of fixed duration for $j = 0, \ldots, 11$ ($J = 12$). If the data have a Poisson distribution with mean λ, the probability content of the jth cell is $p_{0j}(\lambda) = \lambda^j e^{-\lambda}/j!$ We have to estimate the unknown parameter λ in order to get the E_j's. For this data set, the maximum likelihood estimate of λ based on the grouped data is very close to the sample mean $\overline{X} = 3.88$. Using this value as $\widehat{\lambda}$, the chi-squared statistic is computed to be 11.39. The P-value from a χ^2distribution with $J - p - 1 = 10$ degrees of freedom is $P\{\chi^2_{10} > 11.39\} = 0.33$, so there is no evidence to reject the null hypothesis of a Poisson model.

Before proceeding with chi-squared tests for other situations, we digress slightly to describe a graphical technique for assessing the fit of a Poisson model. Under the Poisson assumption, the O_j's should be close to $n\lambda^j e^{-\lambda}/j!$ If we let $Y_j = \log(O_j) + \log(j!) - \log(n)$, then Y_j should be close to $-\lambda + log(\lambda)j$. Thus, a plot of Y_j vs. j should be approximately linear with slope $\log(\lambda)$ and intercept $-\lambda$. A nonlinear configuration would suggest that the assumption of a Poisson model does not hold. This plot was suggested by Hoaglin (1980) who called it the "Poissonness" plot. See Hoaglin and Tukey (1985) for graphical techniques for checking the shapes of other discrete distributions.

Example: Figure 1 shows the Poissonness plot for the emission data. The plot is remarkably linear indicating that the Poisson model is reasonable for this data set. Note that we did not have to estimate the unknown value of λ to assess the fit of the Poisson model. In fact, we can estimate λ from Figure 1 by fitting a line to the linear configuration of points.

3.3 Two-Sample Tests

The chi-squared test can also be used for two and K-sample problems. We will describe it for the two-sample hypothesis (4). Extension to $K > 2$ is straightforward.

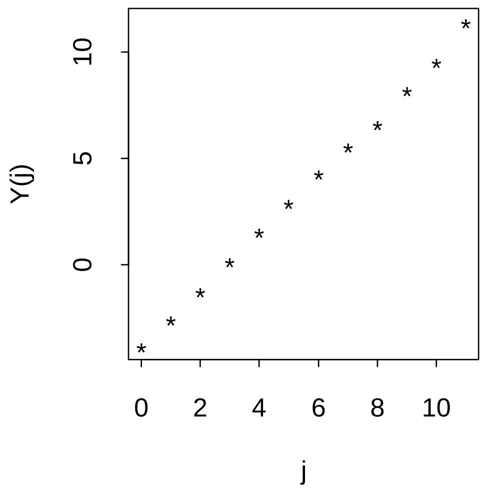

Figure 1: Poissonness plot for data on emission of α-particles.

Let X_i, $i = 1, \ldots, n$ and Y_k, $k = 1, \ldots, m$ be iid from $F(\cdot)$ and $G(\cdot)$ respectively. We assume that the two data sets are mutually independent. Then O_{1j} is the the number of X_i's in cell j and O_{2j} is the number of Y_k's in cell j, $j = 1, \ldots, J$. The corresponding "expected" number of observations under the null hypothesis (4) are $E_{1j} = n\widehat{p}_j$ and $E_{2j} = m\widehat{p}_j$ where $\widehat{p}_j = (O_{1j}+O_{2j})/(n+m)$. The chi-squared statistic for the two-sample problem is then obtained from (6) by summing over all the cells involving both samples. If n and m are both large, the null distribution of the X^2 statistic for the two-sample problem can also be well approximated by a χ^2 distribution with $(J-1)$ degrees of freedom. For the general K-sample case, the test statistic has a limiting χ^2 distribution with $(K-1)(J-1)$ degrees of freedom.

3.4 Discussion

The χ^2 approximations to the distributions of the test statistics work well provided the expected number of observations in each cell is not too small. A conservative guideline is that the expected number in each cell is at least five. There are also guidelines available for selecting the number of cells and cell sizes. It is generally recommended that the cells are chosen to have equal probability under the null hypothesis. Moore (1986) suggests a rough rule of thumb for the number of cells as $J \approx 2n^{2/5}$.

The chi-squared tests are based on grouped data, and so they are ideally suited for situations where data are only available in grouped form or for testing hypotheses about discrete distributions. For other situations, there is a loss of information from grouping the data into cells, and the procedure will not perform as well as the others that we will discuss in later sections. The chi-squared tests also do not take into account the natural ordering of the J cells or categories. Variations of the chi-squared test have been proposed to address this shortcoming. See Moore (1986) and references therein for other, chi-squared type tests.

4 Quantile-Quantile (Q-Q) Plots

The goodness-of-fit methods in the rest of this chapter rely on nonparametric estimators of the underlying distribution, i.e., estimators that are not based on on any assumptions about the shape of the distribution. So we begin with a discussion of nonparametric estimation of $F(x)$. Throughout, we use the notation $\widehat{F}(x)$ generically to denote any nonparametric estimator of $F(x)$. Whenever it is necessary to specify the particular method of estimation, more specific notation will be used.

4.1 Nonparametric Estimation of the CDF

Consider first the complete data situation with $X_1, \ldots, X_n$ iid from $F(x)$. To estimate $F(x)$, we can use the empirical cumulative distribution function (*ecdf*)

$$\widehat{F}_e(x) = n^{-1} \sum_{i=1}^{n} I[X_i \leq x] \ , \tag{7}$$

where $I[X_i \leq x] = 1$ if $X_i \leq x$ and equals zero otherwise. Let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the order statistics of the sample data, i.e., the values of the X_i's arranged in increasing order. Then, it can be seen that the *ecdf* is a step-function that jumps a height of $1/n$ at the value of each order statistic $X_{(i)}$.

Example: Figure 2 is the plot of the *ecdf* for the mechanical device failure time data in Table 2 ignoring the failure mode information. There were 37 observations, so the *ecdf* jumps by 1/37 at each of the order statistics and is constant in between the values of the order statistics. The estimated cumulative probabilities and quantiles can be read off from this plot. In addition, it is possible to identify features peculiar to the data set, such as the striking gap immediately preceding 2 million operations, indicating a relatively long time between successive failures at that point.

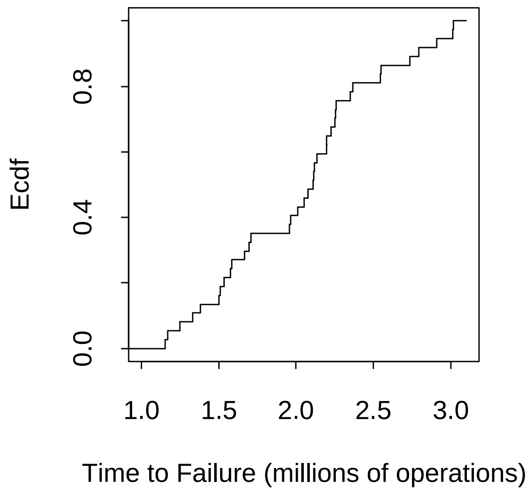

Figure 2: *Plot of the ecdf for the mechanical device failure time data.*

As we have already seen, there are situations where the data are subject to some form of censoring. For Type I, Type II and interval censored data, the *ecdf* $\widehat{F}_e(x)$ in (7) can be used to estimate $F(x)$ for a restricted range of x-values. If the data are right censored at C, then $\widehat{F}_e(x)$ can be used for $x \leq C$. No nonparametric estimates are available for $x > C$ without additional information about the censored observations. Left-censored data are handled similarly. For grouped or interval-censored data, $\widehat{F}_e(x)$ can again be used to estimate $F(x)$ at $x = T_j$, $j = 1, \ldots, J$, the upper end-points of the intervals. At other points, the *cdf* cannot be estimated precisely in general.

Example: The RH1 data in Table 3 are right-censored at $C = 3067$ hours. Since 11 out of 71 units are censored at this value, the *ecdf* goes up only to $\widehat{F}(C) = 60/71$. The ecdf is not defined for $x > 3067$ hours.

Nonparametric estimation is more involved for multiply-censored data. Under the assumption that the observations and censoring variables are independent (actually a slightly weaker assumption is sufficient), the Kaplan-Meier estimator can be used to estimate $F(x)$.

We define the Kaplan-Meier estimator for right-censored data. (The estimator for left-censored data can be obtained from this by observing that if X_i's are left-censored, $-X_i$'s are right-censored.) Let R_i denote the rank of X_i in the (lexicographic) ordering of $(X_i, 1-\delta_i)$, $i = 1, \ldots, n$, and let $X_{(n)} = \max_i X_i$. Then, the Kaplan-Meier estimator is given by

$$\widehat{F}_{KM}(x) = 1 - \prod_{\{i: X_i \leq x\}} [(n - R_i)/(n - R_i + 1)]^{\delta_i} \ , \quad x \leq X_{(n)} \ . \qquad (8)$$

This is the nonparametric maximum likelihood estimator of $F(x)$ with multiply right censored data. Note that it is a step function that jumps only at the uncensored observations.

Example: Figure 3 is a plot of the Kaplan-Meier estimates of the *cdf*s of the two failure modes for the mechanical device data. Recall that when a device fails, we can observe only the mode that failed and that the failure time of the competing mode is randomly right censored. It is actually quite interesting that one can estimate the individual distributions from such data. The Kaplan-Meier estimate of the *cdf* for mode B goes only up to about 0.7. It lies above the estimate for mode A for up to about 2.5 million operations indicating that the system is more likely due to mode B in the early stages.

Estimation of the distribution is more difficult when we have complicated censoring patterns involving combinations of right, left, and interval censoring and truncation. Turnbull (1976) discusses methods of estimation for such complicated situations.

4.2 Assessing the Fit of a Completely Specified Distribution

Consider the simple hypothesis (1) where we want to determine if $F(x)$ can be reasonably approximated by a given distribution $F_0(x)$. Since we have a nonparametric estimate, it is natural to plot both the nonparametric estimate $\widehat{F}(x)$ and the hypothesized *cdf* $F_0(x)$ on the same figure and visually compare the fit. Under the null hypothesis, $\widehat{F}(x)$ is an estimate of $F_0(x)$, so

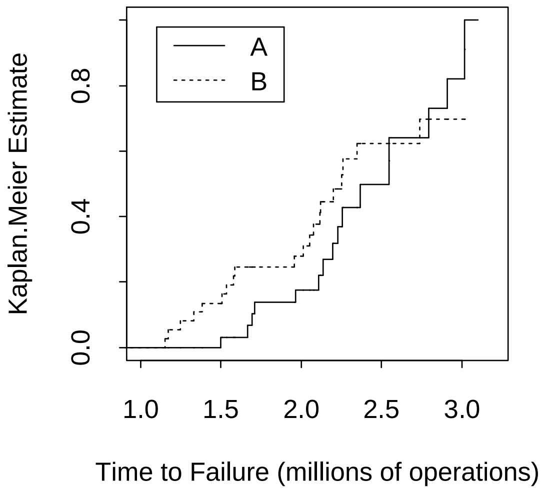

Figure 3: *Plot of the Kaplan-Meier estimates of the cdf's for the two failure modes for the mechanical device data.*

we can expect them to be close to each other. Large, systematic differences would indicate that the null hypothesis is not true.

This "comparative" plot has several drawbacks. First, it is hard to visually assess differences between two curves. Second, if the hypothesis is not true, it is not easy to determine from this plot what other distributions might be reasonable. It is much easier to assess and interpret deviations from a straight line. There are two ways of transforming the scales so that the reference/hypothesized *cdf* is a straight line: transforming the x-axis or the y-axis. The corresponding transformations applied to the *ecdf* result in two well-known *probability plotting* techniques called *percentage-percentage* (P-P) and *quantile-quantile* (Q-Q) plots respectively. We will discuss only the Q-Q plots here. Readers interested in P-P plots should refer to Wilk and Gnanadesikan (1968), and Gan, Koehler and Thompson (1991).

The Q-Q plot, as the name suggests, is a plot of quantiles from the hypothesized *cdf* against the "nonparametric" quantiles from the *ecdf*,

$$F_0^{-1}(p_i) \quad \text{vs.} \quad \widehat{F}^{-1}(p_i) \; , \tag{9}$$

for some suitably chosen plotting positions p_i's. We define the nonparametric quantiles by the right-inverse of $\widehat{F}(x)$, i.e., $\widehat{F}^{-1}(p) = \inf\{x : \widehat{F}(x) \geq p\}$, $p \in (0,1)$. There has been considerable discussion in the literature about the choice of plotting positions. For complete data, we will use $p_i = (i - .5)/n$. With this choice, the Q-Q plot for complete data is simply a plot of

$$F_0^{-1}\left(\frac{i-.5}{n}\right) \quad \text{vs.} \quad X_{(i)} \ , \tag{10}$$

for $i = 1, \ldots, n$, where the $X_{(i)}$'s denote the order statistics.

Example: Consider the data set on failure times of a mechanical device in Table 2. Suppose we want to assess the simple null hypothesis that the distribution of failure times of the devices is a standard exponential distribution with parameter $\lambda = 1$, i.e., $F_0(x) = 1 - \exp(-x)$. To obtain the exponential Q-Q plot, we just substitute in (10) the quantiles of the exponential distribution with parameter $\lambda = 1$, $F_0^{-1}(p) = -\log(1-p)$. Figure 4(a) shows this plot with the order statistics plotted on the x-axis and the standard exponential quantiles on the y-axis. The probabilities are also shown on the right-hand side of the y-axis. If the data are from the standard exponential distribution, the plot should follow the straight line with slope 1 and intercept zero. It is quite clear from this plot, however, that the null hypothesis does not hold. In fact, as we shall see later, none of the exponential distributions fit this data set well.

The Q-Q plots for Type I and Type II censored data are also obtained from (9) using the *ecdf*, but the plotting positions are restricted to the range where $\widehat{F}_e(x)$ is defined. For multiply censored data, the Kaplan-Meier estimate is used to obtain the Q-Q plot. Let $r_1 \leq \ldots \leq r_K$ denote the ranks of those X_i's for which $\delta_i = 1$. Then, the Q-Q plot for multiply censored data is a plot of $F_0^{-1}(p_i)$ vs. $X_{(r_i)}$ where $p_i = [\widehat{F}_{KM}(X_{(r_i)}) + \widehat{F}_{KM}(X_{(r_{(i-1)})})]/2$. In the absence of censoring, this reduces to the plotting positions $p_i = (i - .5)/n$.

The Q-Q plot is essentially a plot of $F_0^{-1}(\widehat{F}(x))$, i.e., a plot of the nonparametric estimate $\widehat{F}(x)$ with the y-axis transformed. It still retains all of the original information in the plot of $\widehat{F}(x)$. Since the y-axis is also labeled in the probability scale, the *cdf* can still be visually estimated from this plot. Unlike the *ecdf* plot, however, it is now easier to interpret the deviations from linearity and identify possible alternative distributions. To develop guidelines on how to interpret the deviations, let us consider the possible shapes of the true (or theoretical) Q-Q plot

$$\tau(x) = F_0^{-1}(F(x)) \ . \tag{11}$$

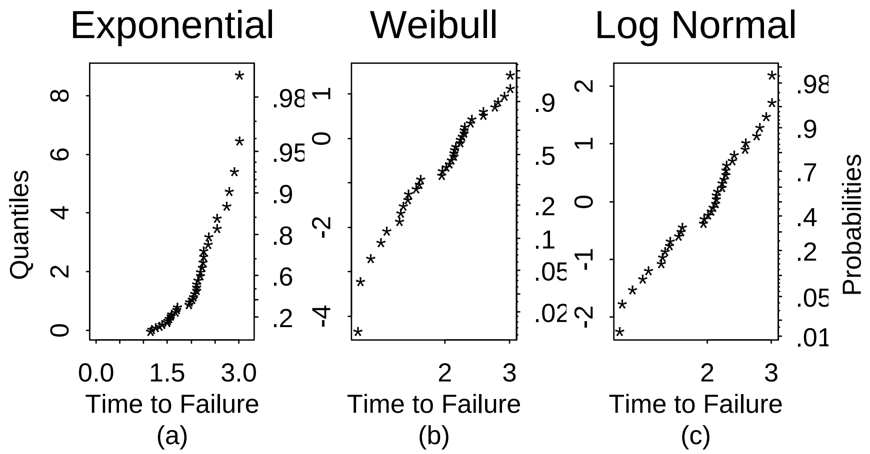

Figure 4: *Q-Q plots for the mechanical device failure time data. The quantile axis is plotted on the left of the plot, the probability axis on the right of the plot for the distribution given in the plot heading. A logarithmic scale is used on the x-axis for the Weibull and log-normal plots.*

- If $F(x) = F_0([x-\mu]/\sigma)$, i.e., the location-scale model (3) holds, it is easy to see that $\tau(x) = [x-\mu]/\sigma$. Thus, if the Q-Q plot does not follow the straight line with slope one and intercept 0 but still appears linear, this would suggest that a location-scale model is reasonable. This will be discussed in more detail in the next section.

- If $\tau(x)$ has an S- or inverted S-shape, then $F(x)$ is *symmetric with respect to* $F_0(x)$. Thus, if $F_0(x)$ itself is symmetric, an S- or inverted S-shaped Q-Q plot would suggest that $F(x)$ is also symmetric. An S-shape implies that $F(x)$ is heavier-tailed than $F_0(x)$ while an inverted $S-$shape implies that it has lighter tails.

- If $\tau(x)$ is piecewise linear, that suggests that the data could be from two different populations so a mixture model might be appropriate.

Fowlkes (1987) provides an extensive catalog of the shapes of the theoretical Q-Q plot, $\tau(x)$, for different distributions. This catalog is very useful in interpreting the shapes and identifying specific parametric alternatives.

4.3 Assessing the Fit of Parametric Families

We discuss first the special case of location-scale families. Examples include the normal, logistic, exponential, and smallest extreme-value (SEV) families. The methods here can also be applied to check the fit of models that can be transformed to location-scale families. For example, the lognormal and Weibull models can be transformed to the normal and SEV families, respectively, by applying a logarithmic transformation to the data.

As we have already noted, under the location-scale model $F_0([x-\mu]/\sigma)$, the Q-Q plot is approximately linear with slope $1/\sigma$ and intercept $-\mu/\sigma$. So we can assess the location-scale hypothesis by checking if the Q-Q plot follows some linear configuration without specifying the slope and intercept of the line. Thus, we can use the same Q-Q plot, without any modification, to check the fit of an entire location-scale family of *cdf*s. Note also that we do not have to estimate the unknown parameters μ and σ. These properties make the use of the Q-Q plot particularly attractive. Let us consider some important cases of location-scale families.

The one-parameter exponential model with *cdf* $1-\exp(-\lambda x)$ is a scale family. It arises naturally in reliability and other applications where interest centers on the failure rate (or rate of decay). The exponential distribution has a constant failure rate (or rate of decay). Constructing an exponential Q-Q plot was already discussed in the last section. Under exponentiality, the plot should be approximately linear with slope λ and intercept zero. If the exponential model does not hold, one is typically interested in knowing whether the underlying distribution has an increasing failure/hazard rate (IFR) or decreasing failure rate (DFR). To obtain such information, observe that the exponential Q-Q plot is just an empirical plot of $H(x)$ where $H(x) = \log(1-F(x))$ is the cumulative hazard function. $H(x)$ is linear under the exponential model, is convex when the distribution has IFR (i.e., dH/dx is increasing) and concave when it has DFR. Thus, the shape of the exponential Q-Q plot can be interpreted to identify the behavior of the hazard rate and possible alternative distributions.

If the exponential model does not fit the data, a natural next step is to consider the more general Weibull model with *cdf* $1-\exp(-[x/\alpha]^{\beta})$ for some $\alpha > 0$ and $\beta > 0$. This family of distributions includes the exponential model as a special case when $\beta = 1$. For $\beta < 1$ it has DFR and for $\beta > 1$ it has IFR. It is thus quite flexible and can be used to model a variety of problems. The Weibull family is not a location-scale family but can be made into one by applying the log-transformation. Specifically, if Y has a

Weibull distribution, then $X = \log(Y)$ has the smallest extreme-value (SEV) distribution $F(x) = 1 - \exp(-e^{[x-\mu]/\sigma})$ which is a location-scale family with location parameter $\mu = \log(\alpha)$ and scale parameter $\sigma = \beta^{-1}$. Thus, the Q-Q plot with $X_{(i)} = \log(Y_{(i)})$ and $F_0^{-1}(p_i) = \log(-\log(1-p_i))$ can be used to assess the fit of Weibull models. (Evidently, it can also be used to assess the fit of the SEV location-scale family.) Under a Weibull model, the plot will be linear with slope $1/\sigma = \beta$ and intercept $-\mu/\sigma = \beta \log(\alpha)$. In particular, if the exponential model holds, the line will have slope 1. The guidelines discussed in the previous section can be used to interpret deviations from linearity.

The normal distribution plays a central role in statistics and assessing whether a given distribution belongs to a normal family is a problem that occurs frequently. In addition, the lognormal distribution has been found useful in modeling data with heavy tails, such as oil and gas exploration data and environmental data on air quality. The normal Q-Q plot is obtained by taking $F_0(x)$ in (9) to be the standard normal *cdf* $\Phi(x)$. Under a normal model, the plot will be approximately linear with slope $1/\sigma$ and intercept $-\mu/\sigma$. The Q-Q plot for assessing lognormality is obtained by first taking logarithms of the data and then using the normal Q-Q plot.

Example: As we have already seen, Figure 4(a) is the exponential Q-Q plot for the mechanical device failure data. Since no line through the origin fits the data well, we can conclude that the exponential model is not reasonable for this data set. The plot appears to be nonlinear with a convex shape, suggesting a distribution with IFR as a possible alternative. The Weibull Q-Q plot in Figure 4(b) is approximately linear, suggesting that the Weibull model is plausible. The lognormal Q-Q plot is shown in Figure 4(c), and this suggests that that the lognormal model might be a slightly better fit.

Example: Consider now the randomly right-censored data corresponding to the two failure modes A and B. Figures 5(a)–(c) show the exponential, Weibull and lognormal Q-Q plots based on the Kaplan-Meier estimates of the two *cdf*s. Again, the exponential model does not fit the data well. The Weibull model seems reasonable for both failure modes, but the slopes of the two plots appear to be different, however. This suggests that a simple Weibull model with the same shape parameter for both distributions may not be appropriate. Figure 5(c) indicates that the lognormal model is also plausible.

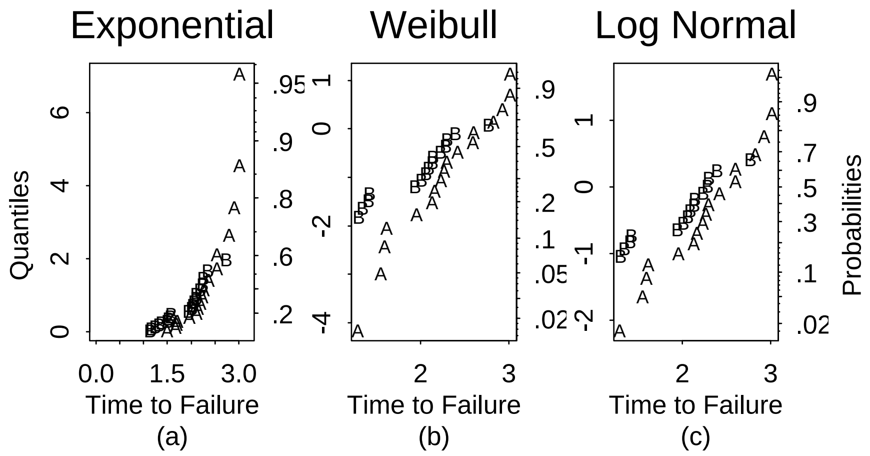

Figure 5: *Q-Q plots for the distributions of the two failure modes for the mechanical device data.*

To extend the Q-Q plots to more general parametric families such as beta or gamma, we have to first estimate the unknown parameters $\boldsymbol{\theta}$. The maximum likelihood method can be used to estimate these parameters. Once an estimate of $\boldsymbol{\theta}$ is available, the Q-Q plot for assessing the composite hypothesis (2) is obtained by replacing $F_0^{-1}(p_i)$ in (9) with $F_0^{-1}(p_i; \widehat{\boldsymbol{\theta}})$. Under the null hypothesis, the plot should be approximately linear with slope one and intercept zero. Deviations from linearity can be interpreted using the guidelines discussed in the last section.

5 Plots with Simultaneous Confidence Bands

One of the advantages of the Q-Q plot, over the P-P plot, is that it emphasizes the tails of distributions where there is greater information for discriminating among different parametric models. However, there is also greater variability in the tails. In addition, the high correlation among neighboring order statistics can give the appearance of systematic deviations even under the null hypothesis. Hence, interpreting the plots is often subjective and can be difficult. Therefore, it is useful to have confidence bands that allow us to account for the uncertainty in the data and provide more objective assessment. We first develop confidence bands for the true *cdf* $F(x)$ and use these to get bands associated with the Q-Q plots.

Throughout this paper, a confidence band of the form $[a-b, a+b]$ will be denoted by the notation $a \pm b$. Further, since $F(x)$ is a probability, we can always restrict our confidence bands to lie in $[0,1]$. Although we will not state it explicitly, we assume throughout that this is always done. Finally, we consider in this paper only two-sided bands and tests since they are more useful in goodness-of-fit problems. Corresponding one-sided procedures can be constructed in a similar fashion.

5.1 Confidence Bands for $F(x)$

The confidence bands are based on the nonparametric estimators discussed in Section 4.1. Consider first the situation with complete data. Exact confidence intervals for $F(x)$ that are valid pointwise (i.e., for a given x-value) can be constructed using the fact that $[n\widehat{F}_e(x)]$ has a binomial distribution. If n is moderately large, the quantity

$$\frac{n^{1/2}[\widehat{F}_e(x) - F(x)]}{[\widehat{F}_e(x)(1-\widehat{F}_e(x))]^{1/2}} \tag{12}$$

is approximately normally distributed. From this, an approximate $(1-\alpha)$ level, *pointwise* confidence interval for $F(x)$ is obtained as:

$$\widehat{F}_e(x) \pm n^{-1/2} z_{\alpha/2}[\widehat{F}_e(x)(1-\widehat{F}_e(x))]^{1/2} \ , \tag{13}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal *cdf.*

For goodness-of-fit problems, we need *simultaneous* (instead of pointwise) confidence bands for $F(x)$, i.e., confidence bands that are valid for all, or at least for a large region of, x-values. A class of such bands can be based on the maximum weighted distance

$$M = \sup_x n^{1/2} |\widehat{F}_e(x) - F(x)| \psi(\widehat{F}_e(x)) \ , \tag{14}$$

where $\psi(u), u\epsilon[0,1]$ is a nonnegative function. We can choose different weight functions $\psi(\cdot)$ to emphasize different regions of the *cdf.* We will consider below two special cases of $\psi(u)$.

Let a_α be the $(1-\alpha)$th quantile of the distribution of M, i.e.,

$$P\{M \leq a_\alpha\} = 1 - \alpha \ . \tag{15}$$

We will also refer to a_α as the $(1-\alpha)$ level critical value. It can be shown that, if the underlying *cdf* $F(x)$ is continuous and strictly increasing, the distribution of M does not depend on $F(x)$, i.e., it is distribution-free. So the critical value depends only on α and the weight function $\psi(u)$. It can be computed assuming, for example, that $F(x)$ is Uniform $(0,1)$. The distribution of M has been studied extensively for special cases of $\psi(u)$ [see Stephens (1986) and references therein].

From (14) and (15), we get a $(1-\alpha)$ level confidence band for $F(x)$, valid simultaneously for all $\{x : \psi(\widehat{F}_e(x)) > 0\}$, as

$$\widehat{F}_e(x) \pm n^{-1/2} a_\alpha / \psi(\widehat{F}_e(x)) \ . \tag{16}$$

The most well-known special case of (16) is obtained by taking $\psi(u) \equiv 1$ which yields the Kolmogorov (1933) band (or K-band)

$$\widehat{F}_e(x) \pm n^{-1/2} c_\alpha \ . \tag{17}$$

We use the notation c_α to denote the critical value for this choice of $\psi(u)$. This band is valid simultaneously for all x. However, the variance of $\widehat{F}_e(x)$ is small in the tails since it is proportional to $[F(x)(1-F(x))]$. Hence a band with constant width is not particularly useful in the tails.

It is, therefore, desirable to have a simultaneous band that, analogous to the pointwise band, has width proportional to the (estimated) standard deviation of $\widehat{F}_e(x)$. Such a band can be obtained by taking $\psi(u) = [u(1-u)]^{-1/2}$ for $a \le u \le b$ and $\psi(u) = 0$ otherwise, where $0 < a < b < 1$. The resulting simultaneous band, valid for $\{x : a \le \widehat{F}_e(x) \le b\}$, is

$$\widehat{F}_e(x) \pm n^{-1/2} d_\alpha [\widehat{F}_e(x)(1-\widehat{F}_e(x))]^{1/2} \ , \tag{18}$$

where d_α is the critical value for this choice of $\psi(u)$. We refer to this as the S-band (S refers to the fact that the weight function corresponds to standardizing the distance by the estimated standard deviation). The restriction to the interval $[a \le \widehat{F}_e(x) \le b]$ is necessary because the critical value $d_\alpha \to \infty$ as either $a \to 0$ or $b \to 1$. In our experience, the choice $a = 1 - b = .05$ provides a reasonable compromise between the size of the critical value (and hence the width of the band) and coverage. Note the similarity between the S-band and the pointwise confidence interval (13); only the critical values are different. The value for the S-band is larger to account for the simultaneous nature of the band.

Table 4 in the next section provides large-sample critical values for the K- and S-bands for some commonly used α-levels. For small to moderate sample sizes, the asymptotic critical values for the K-band can be used with the correction $c_\alpha/(n^{1/2} + .12 + .11/n^{1/2})$ instead of $c_\alpha/n^{1/2}$ [see (17)]. A similar correction is not available for the S-band, and readers are referred to Niederhausen (1981) for exact critical values.

Example: Figure 6 shows the *ecdf* together with the 95% K- and S-bands for the mechanical device failure data ignoring the failure mode information. For the S-band, $a = 1-b = .05$ so the band is valid for all $\{x : .05 \leq \widehat{F}_e(x) \leq .95\}$. Notice that the K-band is narrower in the middle of the distribution while the S-band is narrower in the tails. This property can be shown to hold in general. For detecting alternative distributions, most of the information is typically in the tails, and thus the S-band is to be preferred in this regard.

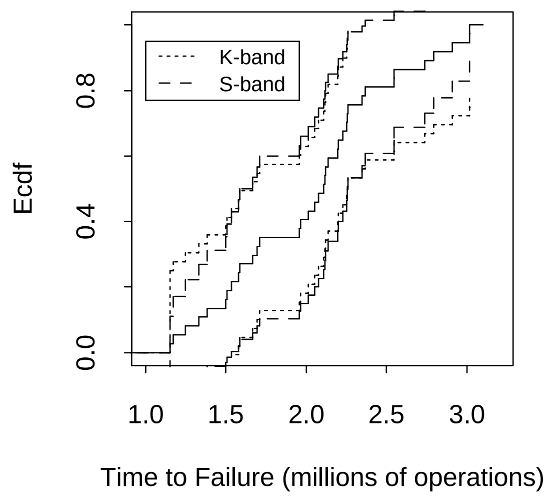

Figure 6: *Comparison of 95% K- and S-bands for the mechanical device failure time data.*

We turn now to censored data. The pointwise band in (13) can be used for Type I, Type II and interval censored data for all x's at which the *ecdf* is defined. For simultaneous bands, the range of the x-values is now restricted, so different critical values must be used in (16)–(18). Barr and Davidson (1973) and Dufour and Maag (1978) provide tables of critical values for the K-band for Type I and Type II censored data. The critical values for complete data can also be used as a conservative approximation.

To obtain interval estimates for multiply censored data, we need to estimate the variance of the Kaplan-Meier estimator $\widehat{F}_{KM}(x)$. Let

$$\widehat{v}_n^2(x) = n \sum_{\{i: X_i \leq x\}} \delta_i / [(n - R_i)(n - R_i + 1)] \ , \tag{19}$$

where X_i, δ_i, and R_i are defined as in Section 4.1. Then, the variance of $\widehat{F}_{KM}(x)$ can be estimated by $n^{-1}[1 - \widehat{F}_{KM}(x)]^2 \ \widehat{v}_n^2(x)$. Using the fact that $[\widehat{F}_{KM}(x) - F(x)]/[\widehat{Var}(\widehat{F}_{KM}(x))]^{1/2}$ is approximately normally distributed in large samples, we get the following approximate $(1 - \alpha)$ level pointwise band for $F(x)$:

$$\widehat{F}_{KM}(x) \pm z_{\alpha/2} n^{-1/2} [1 - \widehat{F}_{KM}(x)] \widehat{v}_n(x) \ . \tag{20}$$

To obtain the censored analog of simultaneous bands, let

$$\widehat{K}(x) = \widehat{v}_n^2(x)/[1 + \widehat{v}_n^2(x)] \ . \tag{21}$$

$\widehat{K}(x)$ behaves like an *ecdf* and in fact reduces to the usual *ecdf* $\widehat{F}(x)$ when there is no censoring. The censored analog of (14) that we need is (see Nair, 1984)

$$M_C = \sup_x n^{1/2} \frac{|\widehat{F}_{KM}(x) - F(x)| \ \psi(\widehat{K}(x))}{[1 - \widehat{F}_{KM}(x)]/[1 - \widehat{K}(x)]} \ , \tag{22}$$

where $\psi(u)$ is defined as before. This pivotal quantity has been carefully chosen so that it is asymptotically distribution-free. It is not possible to obtain bands that are distribution-free in finite samples. It turns out that M_C has the same limiting distribution as M in (14). So if b_α denotes the asymptotic critical value of M, we get a large-sample $(1 - \alpha)$ level band for $F(x)$, valid simultaneously for all $\{x : \psi(\widehat{K}(x)) > 0\}$, as

$$\widehat{F}_{KM}(x) \pm n^{-1/2} b_\alpha [1 - \widehat{F}_{KM}(x)]/\{[1 - \widehat{K}(x)]\psi(\widehat{K}(x))\} \ . \tag{23}$$

By specializing to $\psi(u) \equiv 1$, we get the censored analog of the K-band (due to Hall and Wellner, 1980)

$$\widehat{F}_{KM}(x) \pm h_\alpha n^{-1/2} [1 - \widehat{F}_{KM}(x)]/[1 - \widehat{K}(x)] \ , \quad \text{for all} \quad x < X_{(n)} \ , \tag{24}$$

where h_α is the corresponding critical value. The large-sample value for the uncensored case in Table 4 can be used for h_α unless there is a high degree of censoring. See Hall and Wellner (1980) for improved critical values.

The censored analog of the S-band (due to Nair, 1984) can be obtained by taking $\psi(u) = [u(1-u)]^{-1/2}$ for $a \leq u \leq b$ and $\psi(u) = 0$ otherwise. This yields the band

$$\widehat{F}_{KM}(x) \pm e_\alpha n^{-1/2}[1 - \widehat{F}_{KM}(x)]\widehat{v}_n(x) \ , \tag{25}$$

and it is valid for all $x \in \{a \leq \widehat{K}(x) \leq b\}$. The critical value e_α is the same as the asymptotic critical values of the S-bands in Table 4.

The LT-band, which is related to the S-band but has better performance in some censored data situations, is discussed by Weston and Meeker (1991).

Example: Figure 7 shows the Kaplan-Meier estimate and the 95% S-band for the *cdf* of failure mode A for the mechanical device data. Again, we have taken $a = 1 - b = .05$ here. Notice that, due to censoring, the width of the band is now larger and that the width at the upper tail does not shrink to zero as rapidly as in the complete data case. We have not shown the K-band here, but the relative performances of the K-and S-bands are similar to that with complete data (see Nair, 1984).

5.2 Q-Q Plots with Simultaneous Bands for Assessing the Simple Hypothesis

The simultaneous confidence bands for $F(x)$ can be used to formally test the simple null hypothesis (1). The null hypothesis is accepted at level α if the corresponding $(1-\alpha)$ level band contains the *cdf* $F_0(x)$; otherwise it is rejected. This test can be implemented visually by plotting the simultaneous band and the *cdf* $F_0(x)$ on the same plot and checking whether the band contains $F_0(x)$. However, as we have already seen, visual comparisons are much easier when the plots are displayed on the quantile scale. Thus, we want bands associated with the Q-Q plots.

Recall that the Q-Q plot is just a plot of $\widehat{F}(x)$ with the y-axis transformed by $F_0^{-1}(\cdot)$. By applying the same transformation to the simultaneous bands we get bands associated with Q-Q plots. Specifically, let $L(x)$ and $U(x)$ denote the lower and upper envelopes of a particular simultaneous confidence band for $F(x)$. (As noted before, we implicitly assume that they are restricted $[0,1]$.) Then, the confidence band associated with the Q-Q plot is just $[F_0^{-1}(L(x)), F_0^{-1}(U(x))]$.

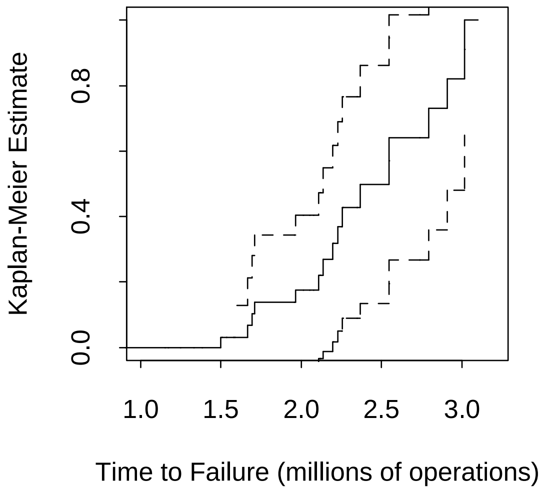

Figure 7: *95% S-band for the cdf of failure mode A for the mechanical device failure time data.*

For complete data, this band can be written explicitly for a general weight function ψ as

$$F_0^{-1}\left(\widehat{F}_e(x) \pm n^{-1/2} a_\alpha / \psi(\widehat{F}_e(x))\right) \ , \tag{26}$$

where a_α is the appropriate critical value for the weight function ψ and the band is valid for $\{x : \psi(\widehat{F}(x) > 0\}$.

This band is a simultaneous $(1-\alpha)$ level confidence band for the true (theoretical) Q-Q plot $\tau(x)$ in (11). The simple null hypothesis that $F(x) = F_0(x)$ can be tested at level α by checking if the band contains the straight line with slope 1 and intercept zero.

Example: Figure 8 shows the exponential, Weibull and lognormal Q-Q plots (same as Figure 4) with a 95% S-band. Let us first consider Figure 8(a). Since the straight line with slope 1 and intercept 0 does not fit into the band, we can reject, at level .05, the simple null hypothesis that $F(x)$ is given by the standard exponential distribution with $\lambda = 1$. Figures 8(b) and 8(c) can be used similarly to assess the fit of Weibull and lognormal models respectively. Suppose we want to test the hypothesis that $F(x)$

is a lognormal distribution with $\mu = 1.0$ and $\sigma = 1.0$. Under this null hypothesis, the lognormal Q-Q plot will be linear with slope $1/\sigma = 1$ and intercept $-\mu/\sigma = -1$. This line does not fit into the 95% S-band, so the hypothesis must be rejected.

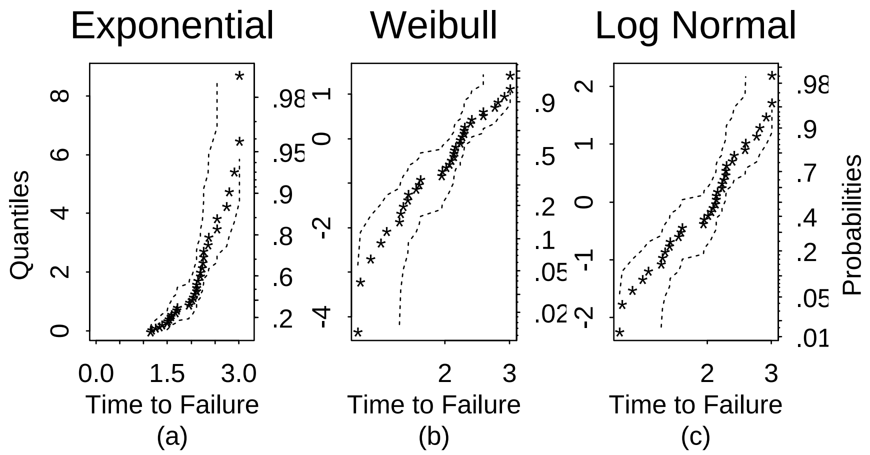

Figure 8: *95% S-band associated with the Q-Q plots for the mechanical device failure time data.*

In addition to being a formal hypothesis test, the band provides us with an indication of the amount of variability in the plot at different x-values and thus allows us to account for the uncertainty while making visual judgments. It retains all the advantages of the original graphical method. In particular, the right-hand side of the y-axis is labeled in probability scale, so one can still use it to obtain simultaneous interval estimates for $F(x)$ and its quantiles.

5.3 Q-Q Plots with Simultaneous Bands for Assessing the Location-Scale Hypothesis

We now consider the use of simultaneous bands for assessing the complex hypotheses in equations (2) and (3). Recall that the Q-Q plot for assessing the general composite hypothesis (2) is obtained by replacing ${F_0}^{-1}(p_i)$ in (9) with ${F_0}^{-1}(p_i; \widehat{\boldsymbol{\theta}})$. It is possible to get a simultaneous band associated with this Q-Q plot. It has the same structure as the band in (26) but the

the critical values α have to be replaced by appropriate values that account for the fact that the parameter $\boldsymbol{\theta}$ has been estimated. See the next section for discussion of the difficulties that arise from this and for appropriate references to critical values. The band is used in the same way as before: The null hypothesis is accepted if the straight line with slope 1 and intercept zero fits into the band and is rejected otherwise.

The location-scale case deserves some special attention. Recall that in this case we can just use a single Q-Q plot, based on the standard member of the location-scale family, to graphically assess the fit of the entire family of distributions. However, to get simultaneous bands to formally assess the location-scale hypothesis in (3), we need to estimate the location-scale parameters. Let $\widehat{\mu}$ and $\widehat{\sigma}$ be the estimators of the location and scale parameters respectively, and let $\tilde{a}_\alpha$ be the critical value of the corresponding test statistic (see next section for a discussion of some properties required of the estimators and of the corresponding test statistics). Then, the simultaneous band for checking the location-scale model is given by:

$$F_0^{-1}\left(\widehat{F}_e(x) \pm n^{-1/2}\tilde{a}_\alpha/\psi(\widehat{F}_e(x))\right). \tag{27}$$

The band is valid for $\{x : \psi(\widehat{F}(x) > 0\}$. The location-scale hypothesis is assessed by checking visually if the band contains the fitted line $[(x-\widehat{\mu})/\widehat{\sigma})$.

Example: Suppose we want to check the location-scale hypothesis for the printed wiring board data at stress condition RH1. Figures 9(a)–(c) show the Q-Q plots for assessing the exponential, Weibull, and lognormal models respectively. The figures also show 95% S-bands for testing the location and/or scale hypothesis. The lines in the figures correspond to the "best-fitting" member of the parametric family. The null hypothesis is accepted if the band contains this line; otherwise it is rejected. We conclude from the figures that the exponential model is not appropriate while the more general Weibull model is reasonable. The lognormal model, however, does not fit the data.

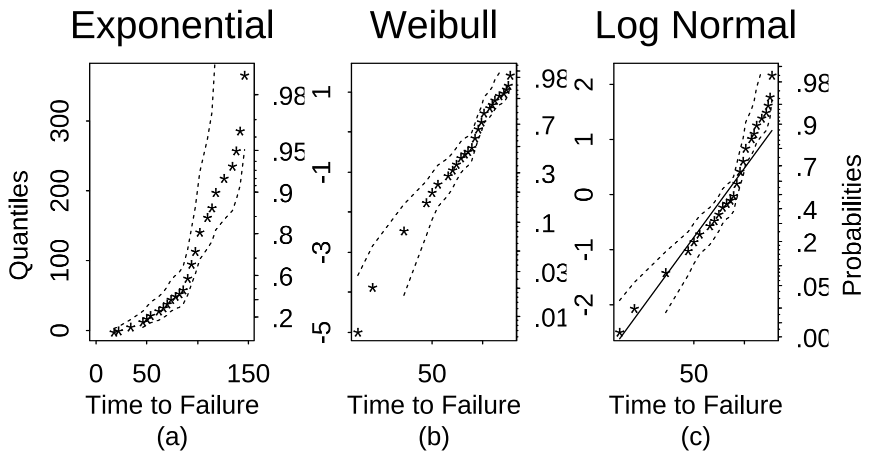

Figure 9: *Q-Q plots and 95% S-bands for testing the location-scale hypothesis for the printed wiring board data at stress condition RH1. The line drawn in Figure 9c represents the line with intercept $\overline{X}$ and slope S, where $\overline{X}$ and S are the sample mean and standard deviation of the log-failure times.*

6 Formal Test Procedures

6.1 Assessing the Fit of a Completely Specified CDF

We have already seen that the simultaneous confidence bands in Section 5 yield formal tests of the simple hypothesis (1). In fact, there is a one-to-one correspondence between these bands and a class of goodness-of-fit tests. For complete data, this class of test statistics is obtained by substituting the hypothesized *cdf* $F_0(x)$ for the true *cdf* $F(x)$ in (14). This results in the class of *maximal distance* tests

$$M_0 = \sup_x n^{1/2} |\widehat{F}_e(x) - F_0(x)| \psi(\widehat{F}_e(x)) \ , \tag{28}$$

where $\psi(u)$ is defined as before. The special case of $\psi(u) \equiv 1$ corresponds to the well-known Kolmogorov (1933) test, which we call the K-test. The choice of $\psi(u)$ that corresponds to the S-band was discussed by Doksum and Sievers (1976) in the context of two-sample comparisons. We refer to this as the S-test. The properties and critical values of these tests are the same as the corresponding bands.

Another class of well-known tests are the *average distance* tests, based on the average weighted squared distance

$$A_0 = n \int [\widehat{F}_e(x) - F_0(x)]^2 \phi(F_0(x)) dF_0(x) \ , \tag{29}$$

where $\phi(u)$ is a non-negative function defined on $[0,1]$. The choice of the weight function $\phi(u) \equiv 1$ yields the well-known Cramer-von Mises statistic

$$CM = n \int [\widehat{F}_e(x) - F_0(x)]^2 dF_0(x) \ . \tag{30}$$

If we want to standardize the squared deviation by weighting it inversely proportional to its variance (analogous to the S-band), we should take $\phi(u) = [u(1-u)]^{-1}$. This yields the Anderson-Darling statistic

$$AD = n \int \frac{[\widehat{F}_e(x) - F_0(x)]^2}{[F_0(x)(1 - F_0(x))]} dF_0(x) \ . \tag{31}$$

Like the maximal distance tests, the average distance tests are also distribution-free. Table 4 provides asymptotic critical values for several commonly used α-levels. See Stephens (1986) and references therein for additional tables of critical values.

Table 4. Asymptotic Critical Values for K, S, CM, and AD Tests

Band	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
K	1.63	1.36	1.22
$S : a = 1 - b = .01$	3.81	3.31	3.08
$S : a = 1 - b = .05$	3.68	3.16	2.91
$S : a = 1 - b = .10$	3.59	3.06	2.79
CM	0.74	0.46	0.35
AD	3.86	2.49	1.93

Example: Suppose we want to formally test the hypothesis that the distribution $F(x)$ of the mechanical device failure data is lognormal with $\mu = 1.0$ and $\sigma = 1.0$. The computed values of the test statistics are: $K = 2.78, S =$

$11.27, CM = 1.98$, and $AD = 9.46$. All of these test statistics are significant at the 0.01 level, and so this null hypothesis must be rejected. This is consistent with our conclusion in the last section based on the 95% S-band.

We turn now to a comparison of the performances of the various tests. As noted already, most of the information for discriminating between parametric models is in the tails, so procedures that give more weight to the tails will perform better. We have already seen this in the context of confidence bands where the S-band did better than the K-band. Similarly, the AD test, which gives greater weight to the tails, is better than the CM test in this regard. Overall, the K-test is dominated by the other tests discussed in this section and is not recommended. The others are all competitive. The advantage of the S-test is that it yields confidence bands which can be displayed graphically. However, it has to be restricted to a range $0 < a \leq \widehat{F}(x) \leq b < 1$. The AD test, on the other hand, does not suffer from this problem. There are many other tests in the literature for testing the simple hypothesis (1). Interested readers should refer to D'Agostino and Stephens (1986, Chapters 4 and 8).

Extensions of the maximal distance tests to censored data have already been discussed in Section 5 when we discussed confidence bands. The average weighted statistics extend to Type I and II censored data in a straightforward manner. The only modification is that the range of integration in (29) has to be restricted to the range over which the data are observable. This would affect the distribution of the test statistic. See Pettit and Stephens (1976) for tables of critical values.

The average distance tests have also been extended to multiply censored data (see, for example, Nair 1981). The analog of (29) for multiply censored data is

$$A_C = n \int \left(\frac{[\widehat{F}_{KM}(x) - F_0(x)]}{[1 - \widehat{F}_{KM}(x)]/[1 - \widehat{K}(x)]} \right)^2 \phi(\widehat{K}(x)) d\widehat{K}(x) \ , \qquad (32)$$

where $\widehat{K}(x)$ is defined in Section 5. Since $\widehat{K}(x)$ is a step-function which jumps only at the uncensored observations, (32) can be expressed explicitly as a sum and computed easily. By choosing special cases of $\phi(u)$ as before, we get the censored analogs of the Cramer-von Mises and Anderson-Darling tests. The class of test statistics in (32) is asymptotically distribution-free under the null hypothesis, and the asymptotic critical values are the same as those for the uncensored case. So the values in Table 4 can be used to implement the tests. The performances of the various tests have not been as extensively studied, however, in this case.

6.2 Assessing the Fit of Parametric Location-Scale Families

We consider first the class of maximal distance tests for testing the location-scale hypothesis (3) for complete data. Let $F_0([x-\mu]/\sigma)$ denote the hypothesized location-scale family. In Section 6.1 where we considered the simple null hypothesis, the class of maximal distance tests were obtained by replacing $F(x)$ in (14) with the hypothesized *cdf* $F_0(x)$. But now we have a composite hypothesis with unknown μ and σ. A natural step is to estimate the unknown parameters and replace $F(x)$ in (14) with $F_0([x-\widehat{\mu}]/\widehat{\sigma})$. This yields the following class of maximal distance tests for testing the location-scale hypothesis:

$$\widehat{M}_0 = \sup_x n^{1/2} |\widehat{F}_e(x) - F_0([x-\widehat{\mu}]/\widehat{\sigma})| \psi(\widehat{F}(x)) \ , \tag{33}$$

where, as before, $\psi(u)$ is a non-negative function on $[0,1]$.

The class of statistics (33) is no longer distribution-free. In general, the distribution depends on the underlying family F_0, the method of estimation and the true values of μ and σ. However, if the estimations $\widehat{\mu}$ and $\widehat{\sigma}$ satisfy certain equivariance properties, the distribution will not depend on the unknown parameters. Most of the usual estimators of location and scale, such as the MLE's, satisfy these properties.

Let $\tilde{a}_\alpha$ denote the critical value of the statistic in (33). Then, the location-scale hypothesis is formally rejected at level α if the test statistic exceeds the critical value $\tilde{a}_\alpha$. Table 5 below provides asymptotic critical values for several parametric families for the two choices of $\psi(u)$ corresponding to the K and S tests. To use these critical values, the unknown parameters must be estimated using the method of maximum likelihood, except for the scale parameter for the normal model which is estimated using the usual sample standard deviation. For additional tables of critical values for K, CM, and AD, see Stephens (1986) and references therein. Note that the values in Table 5 are smaller than the corresponding ones in Table 4. This is because we are now considering the distance between the *ecdf* and the "best-fitted" member of the hypothesized parametric family rather than a fixed member of the family.

The class of average distance tests can be extended similarly by replacing F_0 in (29) by $F_0([x-\widehat{\mu}]/\widehat{\sigma})$. By specializing to the two cases $\phi(u) \equiv 1$

Table 5. Asymptotic Critical Values for Testing the Location-Scale Hypothesis for Selected Parametric Families

Test	Normal			Exponential			Weibull		
α	.01	.05	.10	.01	.05	.10	.01	.05	.10
K	1.04	.90	.82	1.30	1.10	1.00	1.01	.87	.80
$S : a = 1 - b = .01$	2.87	2.48	2.30	3.39	2.95	2.71	2.91	2.51	2.33
$S : a = 1 - b = .05$	2.40	2.11	1.97	3.20	2.70	2.45	2.50	2.17	2.00
$S : a = 1 - b = .10$	2.92	2.00	1.86	2.99	2.54	2.32	2.32	2.01	1.86
CM	.18	.13	.10	.34	.22	.18	.18	.12	.10
AD	1.04	.75	.63	1.96	1.32	1.06	1.04	.76	.64

and $\phi(u) = [u(1-u)]^{-1}$, we get, respectively, the Cramer-von Mises and Anderson-Darling statistics for testing the location-scale hypothesis (3). These statistics are also no longer distribution-free. Table 5 also provides asymptotic critical values for the Cramer-von Mises and Anderson-Darling tests for several parametric families. See Stephens (1986) and references therein for additional tables.

So far we have discussed the two-parameter location-scale problem. In some cases, we might be interested in one-parameter models such as the exponential scale family or the normal location family (with known standard deviation). The above test statistics can also be used in these situations but with only the unknown parameter estimated. The critical values of the test statistics will now be different, however. Table 5 provides asymptotic critical values for the exponential model. See Stephens (1986) for other important cases. The comparisons of the performances of the various statistics are qualitatively similar to the case for the simple null hypothesis. Stephens (1986) suggests, however, that for the composite hypothesis where unknown parameters have to be estimated from the same sample, the differences in the power of the various tests tend to get smaller.

Example: Suppose we now want to test the hypothesis of lognormality for the mechanical device failure data. To do this, we first estimate the parameters of the lognormal distribution using $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$, where $\overline{X}$ and S are the sample mean and standard deviation of the logarithm of the failure times. The computed values of the test statistics are: $K = 0.83$, $S = 1.75$, $CM = 0.11$, and $AD = 0.56$. None of these test statistics is significant at any of the usual α-levels, so we do not reject the hypothesis of lognormality for this data set.

Tests for censored data can be obtained in the same way by substituting $F_0([x-\widehat{\mu}]/\widehat{\sigma})$ for $F_0(x)$ in the tests discussed in Section 6.1. The distributions of the test statistics have not been studied extensively in the literature, and critical values are not widely available.

6.3 Special Tests for Normality

There is a large variety of tests in the literature for assessing the assumption of normality. We have already considered distance based tests for testing normality. We consider briefly some other important tests here. See D'Agostino (1986) for additional details and references to other procedures.

Perhaps the test with best overall performance for assessing normality is the Shapiro-Wilk (1965) test. It is simply a ratio of two estimators of the variance σ^2,

$$SW = \widehat{\sigma}_1^2/\widehat{\sigma}_2^2 \ . \tag{34}$$

The term $\widehat{\sigma}_1$ in the numerator is the best linear unbiased estimator of σ under the assumption of normality while the term $\widehat{\sigma}_2$ in the denominator is the usual sample standard deviation S. If the data are normal, both will estimate σ, and hence the ratio will be close to one. If normality does not hold, $\widehat{\sigma}_1$ does not estimate σ, and hence the ratio will be quite different from one. The exact computation of $\widehat{\sigma}_1$ involves calculating the expected values, variances, and covariances of the order statistics of a sample of size n from a standard normal distribution. This is quite involved, although tables of these values are available in the literature. Many approximations to the Shapiro-Wilk test have been suggested to ease the computational problem. In particular, Weisberg and Bingham (1975) show that the numerator, $\widehat{\sigma}_1^2$, can be approximated well by

$$\tilde{\sigma}_1^2 = \frac{1}{n-1} \frac{[\ \sum_{i=1}^n \Phi^{-1}(p_i) X_{(i)}\ ]^2}{\sum_{i=1}^n [\ \Phi^{-1}(p_i)\ ]^2} \ , \tag{35}$$

where $p_i = [i-3/8]/[n+1/4])$. Note that the p_i's are close to the plotting positions used in the normal Q-Q plot. In fact, the Shapiro-Wilk test is directly related to the measure of linearity in the normal Q-Q plot. Table 6 gives some critical values for the statistic $n(1 - SW^2)$ with SW computed using Weisberg and Bingham's approximation in (35). See Stephens (1986) for a more extensive table of critical values. The hypothesis of normality is rejected at level α if the computed value of $n(1 - SW^2)$ exceeds the critical value.

Table 6. Critical Values of $n(1 - SW^2)$ for Different Sample Sizes n

n	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
10	2.27	1.58	1.26
20	2.81	1.91	1.56
40	3.30	2.23	1.83
60	3.43	2.39	1.96
80	3.54	2.49	2.05
100	3.61	3.01	2.11

The procedures we have discussed thus far are omnibus tests in the sense that they have not been designed with any particular alternatives in mind. Among the various omnibus tests for normality, the Shapiro-Wilk and its modifications have the best performance over a wide variety of alternatives. With the exception of the K-test, the distance based tests also perform reasonably well. The S-test yields a simultaneous band that can be used with Q-Q plots and hence is attractive.

There are also directional tests for detecting particular types of deviations. For detecting asymmetry, moment tests based on the sample skewness have been proposed. Similarly, for detecting heavier or lighter tailed distributions, moment tests based on the sample kurtosis can be used. For details about these tests, readers should consult Bowman and Shenton (1986). When certain types of deviation from normality is suspected a priori, then directional tests specifically geared for such an alternative, such as the moment tests, should be used.

6.4 Tests for General Parametric Families

So far we have restricted attention to parametric location-scale families. More generally, we are interested in testing the composite hypothesis (3). For example, we may want to assess the fit of a gamma model with unspecified shape and scale parameters. The distance-based test statistics in Section 6.2 can be extended in the same manner to this more general situation. However, the distribution of the test-statistics are now more complicated and will depend on the unknown parameter $\boldsymbol{\theta}$. Stephens (1986) gives approximate critical values for some selected parametric families. One can also use simulation to obtain the necessary critical values.

7 Two-Sample Comparisons

Let $F(x)$ and $G(x)$ be two *cdf*'s corresponding to two populations of interest. We have to compare the two *cdf*'s based on two mutually independent sets of data from the two populations. We consider methods for testing the simple hypothesis (4) as well as methods for assessing the common shape of the two *cdf*'s.

7.1 Two-Sample Q-Q Plots

Let $\widehat{F}(x)$ and $\widehat{G}(x)$ be the nonparametric estimates of $F(x)$ and $G(x)$ respectively, based on either complete or censored data. The Q-Q plotting technique has been extended to two-sample problems (see Wilk and Gnanadesikan, 1968). The two-sample Q-Q plot is just a plot of

$$\widehat{G}^{-1}(p_i) \ \text{ vs. } \ \widehat{F}^{-1}(p_i) \tag{36}$$

for some suitable plotting positions p_i's. We discuss some special cases.

Consider first the complete data case with X_i, $i = 1, \ldots, n$ iid from $F(\cdot)$ and Y_j, $j = 1, \ldots, m$ iid from $G(\cdot)$. If the two sample sizes n and m are equal, we can take the p_i's to be i/n, $i = 1, \ldots, n$. Then, (36) is simply a plot of

$$Y_{(i)} \ \text{ vs. } \ X_{(i)} \ , \quad i = 1, \ldots, n \ . \tag{37}$$

When the sample sizes n and m are unequal, let $K = \max(n, m)$, and $p_i = i/K$, $i = 1, \ldots, K$. Then, the two-sample Q-Q plot is given explicitly as

$$Y_{[in/K]} \ \text{ vs. } \ X_{[im/K]} \ , \quad i = 1, \ldots, K \ , \tag{38}$$

where $[x]$ denotes the smallest integer $j : \ j \geq x$. If the samples are subject to censoring, the p_i's in (36) have to be modified suitably. For Type I and II censoring, the p_i's have to be restricted to the range of values for which the quantile functions of both $\widehat{F}$ and $\widehat{G}$ are well-defined. If the data are multiply censored, the plotting positions have to be obtained from the Kaplan-Meier estimates. They can be taken as the set of distinct values of the Kaplan-Meier estimates $\widehat{F}_{KM}$ and $\widehat{G}_{KM}$.

The two-sample Q-Q plot is a plot of $\widehat{G}^{-1}(\widehat{F}(x))$ *vs* x; it estimates $\tau(x) = G^{-1}(F(x))$. If the two distributions are the same, $\tau(x) \equiv x$ so that the Q-Q plot will be linear with slope one and intercept zero. If the distributions differ only in location and scale, i.e., $F(x) = G([x-\mu]/\sigma)$, $\tau(x) = [x-\mu]/\sigma$ and so the plot will be linear with slope $1/\sigma$ and intercept $-\mu/\sigma$. Thus, the same Q-Q plot can be used to check the composite hypothesis that the two distributions belong to the same location-scale family. Unlike the one-sample situation, however, we are not specifying here the shape of the underlying location-scale family. If the Q-Q plot appears to be nonlinear, the guidelines suggested in Section 4.1 can be used to interpret the deviations, such as whether F is symmetric or heavier/lighter-tailed compared to G.

Example: Figure 10(a) shows the two-sample Q-Q plot for comparing the distributions of the ALT data at the two stress conditions. Since the data at RH1 are right-censored and the *ecdf* goes up to only 60/71, the plotting positions p_i's are restricted to this range. The plot exhibits a significant nonlinearity, suggesting that the two distributions are not related through a location-scale model. Thus, the assumption of an accelerated failure time model (5) will not be satisfied. Meeker and Luvalle (1992) consider this data set and develop more complex parametric models based on the underlying reliability kinetics.

7.2 Formal Tests and Confidence Bands

Two-sample versions of the test procedures and confidence bands in Sections 5 and 6 can be obtained. We discuss here only the simple Kolmogorov-Smirnov test and the related confidence band. We consider only the complete data case.

Let $N = mn/(m+n)$. To test the null hypothesis (4), we can use the Kolmogorov-Smirnov (KS) test statistic

$$\sup_x N^{1/2} |\widehat{F}_e(x) - \widehat{G}_e(x)| \,, \tag{39}$$

the maximum distance between the two *ecdf*s. Under the null hypothesis, the distribution of this test statistic does not depend on the (unknown)

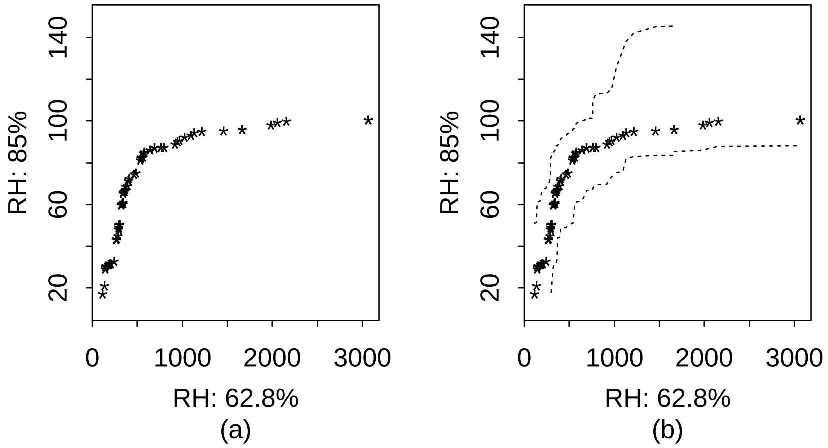

Figure 10: *Two-sample comparison of the ALT data. (a) Two sample Q-Q plot. For illustrative purposes, data grouped in an interval were distributed uniformly within the interval. (b) Two-sample Q-Q plot with 95% KS-band.*

common distribution. Its finite-sample distributions have been tabulated extensively and critical values are readily available. The limiting distribution is the same as that of the one sample Kolmogorov test statistic, and hence the values in Table 4 can be used in large samples. Let c_α denote the $(1-\alpha)$ level critical value of the KS-statistic, i.e.,

$$P\left\{\sup_x |\widehat{G}_e(x) - \widehat{F}_e(x)| \leq N^{-1/2} c_\alpha\right\} = 1 - \alpha \ . \qquad (40)$$

Then, the null hypothesis is rejected at level α if the computed value of the test statistic exceeds the critical value; otherwise it is accepted.

We now show how to use this test statistic to construct simultaneous confidence bands associated with two-sample Q-Q plot, i.e., bands for the function $\tau(x) = G^{-1}(F(x))$. It can be shown that, if $\widehat{G}_e(x)$ is the *ecdf* based on an iid sample from $G(x)$, $\widehat{G}_e(\tau(x))$ has the same distribution as an *ecdf* from $F(x)$. Hence, the null hypothesis (4) is satisfied for the two *ecdf*'s $\widehat{G}_e(\tau(x))$ and $\widehat{F}_e(x)$. From (40) we get

$$P\{\widehat{F}(x) - N^{-1/2} c_\alpha \leq \widehat{G}_e(\tau(x)) \leq \widehat{F}_e(x) + N^{-1/2} c_\alpha \text{ for all } x\} = 1 - \alpha \ . \quad (41)$$

This yields a $(1-\alpha)$ level simultaneous confidence band for $\tau(x)$ as

$$\widehat{G}_e^{-1}(\widehat{F}_e(x) - N^{-1/2} c_\alpha) \leq \tau(x) \leq \widehat{G}_e^{-I}(\widehat{F}_e(x) + N^{-1/2} c_\alpha) \ , \qquad (42)$$

where $\widehat{G}^{-I}(p) = \sup\{x : \widehat{G}(x) \leq p\}$, the left-inverse of $\widehat{G}(x)$. This band can be written explicitly in terms of the order statistics as

$$[Y_{[m\ a(x)]}\,,\ Y_{[m\ b(x)+1]}] \tag{43}$$

where $a(x) = (\widehat{F}_e(x) - N^{-1/2}c_\alpha)$, $b(x) = (\widehat{F}_e(x) + N^{-1/2}c_\alpha)$ and $[z]$ is the largest integer smaller than or equal to z,

Example: Figure 10(b) shows the two-sample Q-Q plot and 95% KS-band for comparing the two distributions for the ALT data. The line with slope one and intercept zero does not fit into the band, so we can reject the null hypothesis that the two distributions are the same. This band can also be used as a conservative test of the composite hypothesis of location-scale model. This is done by checking to see if there is some line that will fit into the band. For a more precise test of the location-scale hypothesis, the procedures in Section 6.2 have to be extended to the two-sample situation.

8 Concluding Remarks

We have focused attention in this paper on one and two sample comparisons. The problem of assessing distributional assumptions is also important in many other, more complex, situations such as regression problems, data with selection-biases, and multivariate data. Many of the graphical techniques discussed here can be extended to handle these situations as well. For example, the Q-Q plotting technique can be used in one-sample problems with selection bias as long as there is a nonparametric estimate of the underlying *cdf.* The formal goodness-of-fit techniques, however, do not extend as easily to the more complex situations. The major problem is that the distribution theory for the test statistics is now much more involved. One alternative approach for obtaining the critical values in some of these cases is the use of simulation/bootstrap techniques.

References

[1] Baar, D. R. and Davidson, T. (1973) "A Kolmogorov-Smirnov Test for Censored Samples," *Technometrics*, 15, 739–757.

[2] Bowman, K. O. and Shenton, L. R. (1986), "Moment ($\sqrt{b_1}$, b_2) Techniques," Chapter 7 in *Goodness-of-Fit Techniques*, edited by D'Agostino, R. B. and Stephens, M. A., Marcel Dekker, Inc., NY.

[3] D'Agostino R. B. (1986) "Tests for the Normal Distribution," Chapter 9 in *Goodness-of-Fit Techniques*, edited by D'Agostino, R. B. and Stephens, M. A., Marcel Dekker, Inc., NY.

[4] D'Agostino, R. B. and Stephens, M. A. (1986) *Goodness-of-Fit Techniques*, Marcel Dekker, Inc., NY.

[5] Doksum, K. A. and Sievers, G. L. (1976), "Plotting with Confidence: Graphical Comparisons of Two Populations," *Biometrika*, 63, 421–434.

[6] Dufour, R. and Maag, U. R. (1978) "Distribution Results for Modified Kolmogorov-Smirnov Statistics for Truncated or Censored Data," *Technometrics* 20, 29–32.

[7] Fisher, R. A. (1924), "The Conditions Under Which χ^2 Measures the Discrepancy Between Observation and Hypothesis," *Journal of the Royal Statistical Society*, 87, 442–450.

[8] Fowlkes, E. B. (1987), *A Folio of Distributions: A Collection of Theoretical Quantile-Quantile Plots*, Marcel Dekker.

[9] Gan, F. F. Koehler, K. J. and Thompson J. C. (1991), "Probability Plots and Distribution Curves for Assessing the Fit of Probability Models," *The American Statistician*, 45, 14–21.

[10] Hall, W. J. and Wellner J. A. (1980), "Confidence Bands for a Survival Curve," *Biometrika*, 67, 133–143.

[11] Hoaglin, D. C. (1980), "A Poissonness Plot" *The American Statistician*, 34, 146–149.

[12] Hoaglin, D. C. and Tukey, J. W. (1985), "Checking the Shape of Discrete Distributions," Chapter 9 in *Exploring Data Tables, Trends, and Shapes* edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley

[13] Kolmogorov, A. N. (1933) "Sulla Determinazione Empirica di una Legge di Distribuziane," *Giorna. Inst. Attuaari.*, 4, 83–91.

[14] Meeker, W. Q. and Luvalle, M. J. (1992), "An Accelerated Life Test Model Based on Reliability Kinetics," submitted for publication.

[15] Moore, D. S. (1986) "Tests of Chi-Squared Type," Chapter 3 in *Goodness-of-Fit Techniques*, edited by D'Agostino, R. B. and Stephens, M. A., Marcel Dekker, Inc., NY.

[16] Nair, V. N. (1981), "Plots and Tests for Goodness-of-Fit with Randomly Censored Data," *Biometrika*, 68, 99–103.

[17] Nair V. N. (1984), "Confidence Bands for Survival Functions with Censored Data: A Comparative Study," *Technometrics*, 26, 265–275.

[18] Niederhausen, H. (1981), "Tables and Significant Points for the Variance Weighted Kolmogorov-Smirnov Statistic," Technical Report 298, Stanford University, Dept. of Statistics.

[19] Pettit, A. N. and Stephens, M. A. (1976), "Modified Cramer-von Mises Statistics for Censored Data," *Biometrika*, 63, 291–298.

[20] Rutherford, E. and Geiger, H. (1910), "The Probability Variations in the Distribution of α-Particles," *Phil. Mag. S6*, 20, 698.

[21] Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611.

[22] Stephens, M. A. (1986) "Tests Based on EDF Statistics," Chapter 4 in *Goodness-of-Fit Techniques*, edited by D'Agostino, R. B. and Stephens, M. A., Marcel Dekker, Inc., NY.

[23] Turnbull, B. W. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society* 38, 290–295.

[24] Weisberg, S. and Bingham, C. (1975), "An Approximate Analysis of Variance Test for Normality Suitable for Machine Calculations," *Technometrics*, 17, 133–134.

[25] Weston, S. A. and Meeker, W. Q. (1991), "Coverage Probabilities of Nonparametric Simultaneous Confidence Bands for a Survival Function," *Journal of Statistical Computation and Simulation*, 32, 83–97.

[26] Wilk, M. B. and Gnanadesikan, R. (1968) "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–19.