# Extreme (X-) Testing with Binary Data and Applications to Reliability Demonstration

David Mease
Department of Marketing and Decision Sciences
San Jose State University
San Jose, CA 95192-0069
mease_d@cob.sjsu.edu

Vijayan N. Nair
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107
vnn@umich.edu

**Abstract**

Reliability demonstration techniques are used to formally verify that the reliability of a product meets a specified target with a certain degree of confidence. When the reliability target to be demonstrated is very high (close to one), traditional reliability demonstration plans require extremely large sample sizes or have low power. One solution to this problem is to inflate the failure probability by testing products under extreme conditions so that they are more likely to fail. It is sufficient then to demonstrate a lower reliability target which can be mapped back to the required reliability under standard conditions. This paper develops a general framework for this type of extreme testing, or "X-testing", with binary data. The effects of X-testing on sample size and power of reliability demonstration plans are discussed. Properties of various X-transforms are studied with respect to zero-failure plans, fixed sample size plans, and fixed power plans. Conditions under which X-transforms lead to inadmissible or uniformly efficient tests are obtained. X-testing is similar in spirit to accelerated testing as both methods are intended to induce failures, but there are some key differences. Several applications, in addition to reliability demonstration, are used to illustrate the general usefulness of the approach.

Key Words: accelerated testing; biased sampling; compressed testing; structural reliability.

# 1 Introduction

## 1.1 Reliability Demonstration Tests

Reliability demonstration is an integral part of the reliability program in manufacturing companies. Suppliers of parts and components are typically required to demonstrate that their products meet specified reliability targets. The requirement can be stated in terms of a component meeting, say, 0.98 reliability target for 36 months, 3 years in service, a million cycles of operation, and so on. Reliability demonstration is also a key part of internal reliability improvement programs within companies. For example, a large manufacturing company has a company-wide "high-mileage problem resolution process" in which reliability demonstration is one of the major phases.

The goal of a demonstration test is to formally verify that a given product meets the specified reliability requirement with a certain level of confidence. The test is typically based on binary (success/failure) data and consists of testing a sample of $n$ products and recording the number of failures (Martz and Waller, 1982). A general $k$-out-of-$n$ reliability demonstration plan can be described as follows. Throughout, let

$$BIN(k;\ n,p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}, \tag{1}$$

for given integers $k$ and $n$ with $k < n$ and failure probability $p$ (or reliability $1-p$) with $0 < p < 1$. Now, consider a demonstration test with $R_0$ as the reliability target and $n$ being the number of units put on test. Let $k$ be the largest integer such that

$$BIN(k;\ n, 1-R_0) \leq \alpha. \tag{2}$$

Then, we say that reliability of $R_0$ has been successfully demonstrated with $(1-\alpha)$-level confidence if the number of failures is less than or equal to $k$. That is, $k$ is the maximum number of acceptable failures out of the $n$ units tested in a $k$-out-of-$n$ plan. For example, suppose we need to demonstrate a reliability of $R_0 = 0.90$ with 95% confidence. If we put $n = 46$ units on test, we can have at most one failure since $k = 1$ is the largest value for which (2) holds. Note that successful demonstration of reliability $R_0$ implies successful demonstration of any reliability less than $R_0$ as a consequence of the stochastic ordering of the binomial distribution.

In developing reliability demonstration plans, test planners typically fix the number $k$ of acceptable failures up front and use this to determine $n$, the sample size or number of units that need to be put on test. In the above example, for a 1-out-of-$n$ failure plan with $R_0 = 0.90$ and $\alpha = 0.05$, the required sample size can be computed from (2) as $n = 46$. The most common plans are zero-failure plans ($k = 0$) since they require the smallest sample size. The sample size $n$ is then determined from the equation $R_0^n = \alpha$. The relation $C = (1-\alpha) = 1 - R_0^n$ has been called the "success-run" formula in the engineering literature (Lipson and Sheth, 1973).

The power of a demonstration plan is defined formally in the next section. Simply speaking, it is the probability that the demonstration plan will be successful when the true unknown reliability is some value $R_a$

which is greater than $R_0$. Among $k$-out-of-$n$ plans, plans with large $k$ have higher power. Thus, zero-failure plans, which are economical in terms of sample size, have the smallest power.

There is a great deal of emphasis in industry to continuously improve reliability to meet customer demands and reduce warranty costs due to safety-critical needs. However, when the reliability $R_0$ is very high (close to one), traditional demonstration plans become impractical. They require large sample sizes and considerable resources to achieve reasonable confidence levels with adequate power. For example, to demonstrate a reliability of $R_0 = .99$ with 95% confidence would require testing $n = 299$ units even for a zero-failure plan. Therefore, it is of interest to investigate different testing regimes based on inducing failure and using the data from this transformed environment to carry out reliability demonstration more efficiently.

We will refer to this type of testing as "extreme testing," or "X-testing" for short. At first glance, this sounds like the traditional accelerated testing that is used commonly in reliability, and indeed there are close connections. However, the X-testing framework is more general and includes other ways of increasing the failure probability. Several examples are given in Section 2.

## 1.2 General Formulation

Consider the following general formulation of X-testing with binary data. Let $Z$ be a $d$-dimensional performance characteristic, and let $\mathcal{A}$ be the region of interest. We observe binary data $Y = \mathrm{I}\{Z \in \mathcal{A}\}$, i.e., $Y = 1$ if $Z \in \mathcal{A}$ and zero otherwise. Let $R = \mathrm{E}(Y) = \mathrm{P}(Z \in \mathcal{A})$. The goal is to test

$$\mathrm{H}_0 : R \leq R_0 \text{ vs. } \mathrm{H}_a : R > R_0. \tag{3}$$

This formulation with the alternative hypothesis as $R > R_0$ might seem counter-intuitive at first. However, for reliability demonstration, we have to show that $R$ is at least as good as $R_0$ with a high degree of confidence. Thus, we take the null hypothesis as $R \leq R_0$ so that considerable evidence or a low $p$-value is required to reject the null and establish the alternative hypothesis or reliability target being met. (For further discussion on this relationship see Mann, Schafer and Singpurwalla, 1974.)

The usual testing is based on $n$ independent and identical experiments which yield *iid* Bernoulli observations $Y_i$, $i = 1, ..., n$. Thus, an $\alpha$-level test rejects the null hypothesis if (2) holds. For instance, if $n = 46$, $R_0 = 0.9$, and $\alpha = .05$, we will reject the null hypothesis if we observe one or zero failures.

There is a one-to-one relationship between this testing scheme and the demonstration plans discussed in the last section. In reliability demonstration, given the number of acceptable failures, say $k$, the goal is to choose the sample size $n$ to ensure that the level of the test is at least as small as $\alpha$. This leads to the $k$-out-of-$n$ failure plan with sample size given by the smallest integer $n$ solving (2).

Since the problem is now formulated in terms of a hypothesis test, we can define power in the usual way. If $R_a > R_0$ is the true reliability, then the power is given by

$$BIN(k;\ n, 1 - R_a). \tag{4}$$

This is nothing more than the probability of a successful demonstration when the true reliability is $R_a$.

As previously observed, when $R_0$ is close to one, very large sample sizes and high testing costs will be required to test at reasonable $\alpha$-levels or to achieve sufficient power. Thus, it is of interest to investigate different testing regimes based on inflating the failure probability and using the data from this transformed environment to test the hypothesis more efficiently.

Specifically, let $X$ be the binary outcome under the transformed environment with $E(X) = \tau(R)$. We refer to $\tau(\cdot)$ as the X-transform and assume throughout that it is a continuous strictly increasing function from $(0, 1)$ to $(0, 1)$. Let $\tau_0 = \tau(R_0)$. Then the new testing problem is

$$\text{H}_0 : \tau(R) \leq \tau_0 \text{ vs. } \text{H}_a : \tau(R) > \tau_0. \tag{5}$$

To implement the X-test, we need to know the value of $\tau_0$. (The validity of this assumption will be discussed in the context of the different applications we consider.) Under the new testing regime, the null hypothesis will be rejected if $BIN(k;\ n, 1-\tau_0) \leq \alpha$ where $k$ is the number of observed failures. Analogously, for a fixed number of allowable failures $k$ in a demonstration plan, the power is $BIN(k;\ n, 1 - \tau(R_a))$. Note that the power depends on $\tau(R_a)$, the transformed value of the true probability.

### 1.3 Overview

This paper provides a general treatment of X-testing with binary data. We study the properties of the X-transforms and characterize situations where the X-transform leads to higher efficiency in terms of sample size and power. The effect of inducing failure on the power of the test plan has not received as much attention in the reliability demonstration literature.

We study X-testing with respect to three problems of interest: (a) the zero-failure problem, (b) the fixed sample size problem, and (c) the fixed power problem. In case (a), a zero-failure X-test is used in place of a traditional zero-failure plan in order to reduce the sample size needed. For (b), the X-testing maintains the same sample size as the traditional demonstration plan but is applied for the purpose of increasing the power. In (c), X-testing is applied in order to achieve a required power using a smaller sample size than that needed to obtain the required power with a traditional demonstration plan.

The paper is organized as follows. Section 2 gives some examples of X-transforms that arise in various applications. Section 3 gives some theoretical results for X-transforms that apply to all three problems described in the previous paragraph. Sections 4, 5 and 6 give theoretical as well as numerical results specific to each of the three respective problems. Section 7 discusses practical implications of assuming an X-transform that differs from reality. Section 8 gives some additional examples of X-testing illustrating the applications of the results presented, and Section 9 provides some concluding remarks.

# 2 Applications and Examples of X-Transforms

## 2.1 Increasing Test Time

Increasing test time (over-testing) is a well-known technique in reliability demonstration (see, for example, Meeker and Escobar, 1998). Suppose the failure times follow a log location-scale distribution such as Weibull or lognormal, i.e., the time-to-failure distribution is $F((\log(t) - \mu)/\sigma)$ for some baseline CDF $F(\cdot)$. In traditional demonstration plans, one assumes that the scale parameter $\sigma$ is known. Then, for a given sample size and confidence level, one can work out the time required to test the units. This is often called bogey testing (Lipson and Sheth, 1973).

Let $t_1$ be the required time. Then, the reliability is $R = 1 - F((\log(t_1) - \mu)/\sigma)$. Suppose instead of testing for $t_1$ units, we inflate the testing time to $t_2$ units ($t_2 > t_1$). Then $\tau(R) = 1 - F((\log(t_2) - \mu)/\sigma)$. Letting $Q_F(u)$ be the quantile function of $F$, we can re-express this as

$$\tau(R) = 1 - F(Q_F(1 - R) + A) \tag{6}$$

where $A = (\log(t_2) - \log(t_1))/\sigma$. Note that $A$ does not depend on $\mu$ and so the function $\tau(\cdot)$ is completely known if $\sigma$ is known, which is the common assumption in demonstration testing.

Define $\tau(R_0) = \tau_0$ where $R_0$ is the probability under the null hypothesis (or reliability to be demonstrated). Then we get $A = Q_F(1 - \tau_0) - Q_F(1 - R_0)$. Using this, one can rewrite (6) as

$$Q_F(1 - \tau(R)) - Q_F(1 - R) = Q_F(1 - \tau_0) - Q_F(1 - R_0) = A. \tag{7}$$

Any X-transform that has the form in equation (7) is defined to be a *location X-transform* for distribution $F$. Location X-transforms are discussed in more detail in Section 3.

## 2.2 Increasing Stress Levels and Connections to Accelerated Testing

An alternative method of inducing failure is to modify the stress environment (Kececioglu, 1994). This is closely related to the well-known notion of accelerated stress testing. Consider again the log-location failure time model (also referred to as the accelerated failure time (AFT) model in this context, Meeker and Escobar, 1998). Suppose the location parameter $\mu$ decreases linearly with a stress factor $S$ (typically after a suitable transformation as in the Arrehenius model) so that we have $\mu(S) = \beta_0 + \beta_1 S$. In accelerated stress testing, the units are tested at two or more stress levels (much higher than the operating/design level $S_0$) and failure time data are collected from which the values of $\beta_0$, $\beta_1$ and $\sigma$ are estimated. These are then used to estimate quantiles at the design condition, which are of the form $\mu(S_0) + Q_F(p)\sigma$, where $Q_F(p)$ is the $p$-th quantile of the baseline distribution $F$. The notion of accelerated testing has been considered in the context of binary data (Joseph and Wu, 2004; Nair and Wang, 2004) although the scale parameter $\sigma$ cannot be estimated with binary data.

X-testing with binary data can be employed in this set up by testing at the higher stress level $S_1$, leading to the X-transform in (6) with $A = (\beta_1[S_0 - S_1])/\sigma$. Implementation of this scheme requires knowledge of $A$; hence of $\sigma$ and $\beta_1$. The latter will have to be obtained from prior knowledge, or one can use a two-stage procedure to get a preliminary estimate and then perform X-testing. (In this case, one has to take the uncertainty in the preliminary information into account.) Note that this X-transform is of the same form as (7) and hence is a location X-transform.

A similar situation arises when the effect of stress on time-to-failure is captured by a proportional hazards model (Cox, 1972). The reliability at time $t$ is then given by

$$R(t) = e^{-H_0(t)e^{\beta S}}$$

where $S$ is the stress level and $H_0(t)$ is the baseline cumulative hazard. Testing at a higher stress level $S_1$ than the design level $S_0$ as before gives $\tau(R) = R^{e^{\beta(S_1-S_0)}}$. This requires knowledge of only the coefficient $\beta$ and not of the underlying failure distribution.

If we parameterize these X-transforms in terms of $\tau_0$, we have $\tau(R) = R^{\frac{\log(\tau_0)}{\log(R_0)}}$. By applying the transformation $\log(-\log(u))$ to both sides, we get $\log(-\log(\tau(R))) = \log(-\log(R)) + A$. Note again that this is a location X-transform. In fact, regardless of the underlying failure distribution it can be written as $Q_{\text{sev}}(1-\tau(R)) = Q_{\text{sev}}(1-R)+A$ where $Q_{\text{sev}}$ is the quantile of the smallest extreme value (SEV) distribution.

## 2.3 Stress-Strength Models and Testing "Weaker" Units

Stress-strength models are commonly used in structural reliability. If $X$ denotes strength (or capacity) and $Y$ denotes stress (or load), reliability is defined as the probability that strength exceeds stress, or $R = \mathrm{P}(X > Y)$. There is no notion of time here (although there are some recent extensions to time-based stress-strength models in the literature).

Consider a specific example with $X$ and $Y$ independent and having lognormal distributions with parameters $(\mu_X, \sigma_X)$ and $(\mu_Y, \sigma_Y)$ respectively. Then, $R = 1 - \Phi((\mu_Y - \mu_X)/\sqrt{\sigma_Y^2 + \sigma_X^2})$ where $\Phi$ is the standard normal CDF. For simplicity of exposition, we take $\sigma_X = \sigma_Y = \sigma$, so $R = 1 - \Phi((\mu_Y - \mu_X)/(\sqrt{2}\sigma))$.

One possible approach to X-testing in this context is to test "weaker" units in order to increase the failure probability. Suppose we selected units for test according to a different strength distribution, say lognormal with parameter $\tilde{\mu}_X < \mu_X$ and same $\sigma$. Then, $1 - \tau(R) = \Phi\left(Q_{\text{nor}}(1-R) + \frac{\mu_X - \tilde{\mu}_X}{\sqrt{2}\sigma}\right)$ where $Q_{\text{nor}}$ denotes the standard normal quantile function. This is also of the form $Q_{\text{nor}}(1 - \tau(R)) = Q_{\text{nor}}(1 - R) + A$, a location X-transform.

It is more likely that the new strength distribution would have a smaller standard deviation $\tilde{\sigma}_X$. In this case, straightforward calculations show that the new X-transform is of the form $Q_{\text{nor}}(1 - \tau(R)) = A + BQ_{\text{nor}}(1 - R)$ where $A$ and $B$ are constants that depend on the $\mu$'s and $\sigma$'s. This is a special case of a location-scale transform.

To implement such a scheme, we need some way of identifying weaker units. This is possible if there is some characteristic that is related to strength. For instance, in the case of beams, strength can be related to the diameter of the beams. Then, X-testing can be done on beams with a smaller mean diameter $D$. This could be accomplished either by changing the production process to produce beams with smaller mean diameters (even fixed diameters) or by selecting from the production process using, for instance, a rejection sampling scheme.

One can also implement X-testing by varying the stress instead of the strength distribution. The details are similar to the case with testing weaker units. While this is also a form of accelerated testing, it is conceptually different from the traditional set up discussed in the last section.

## 2.4 Testing Component Reliability Using System Tests

Consider a situation where we are interested in the reliability $R$ of a certain type of component. Further, the components are relatively inexpensive but the cost of testing is high, so we want to reduce the number of tests. We can develop an X-test by constructing an artificial series system that places the components in series and tests the reliability of the series system. If the system has $m$ components in series, then $\tau(R) = R^m$. If we parameterize this X-transform in terms of $\tau_0$, we have $\tau(R) = R^{\frac{\log(\tau_0)}{\log(R_0)}}$. Just as in Section 2.2, this is a location X-transform based on the SEV distribution. For this X-testing formulation to be useful, however, it is essential that the reliability is not adversely affected by the new system configuration.

## 2.5 Attribute/Specification Testing

The X-testing formulation also arises in other applications such as inspection and process capability analysis in quality control (Sarkadi and Vincze, 1974). Consider the following application to attribute sampling and specification testing. Let $R$ be "yield" defined as the proportion of manufactured products for which the characteristic of interest (a continuous variable $Z$) is smaller than the specification limit of $L$. The goal in attribute sampling is to determine if $R$ is sufficiently large (at least as big as $R_0$). Since quality of tested products is often in parts per million, regular testing which involves observing *iid* copies of $Y = \mathrm{I}[Z > L]$ is not economical. Similar issues arise in yield testing in electronics and other semiconductor manufacturing applications.

Beja and Ladany (1974) examined the use of a "compressed limit" sampling plan for this problem (essentially the same as X-testing). They assumed that it is possible to increase (or decrease) the failure probability by observing $\mathrm{I}[Z \geq L_c]$ where $L_c$ is termed an "artificial attribute". They considered the problem of finding the optimal value of $L_c$ to give the smallest sample size meeting both a level and power restriction assuming that $Z$ was normal with a known variance but unknown mean. This is a special case of a problem we will examine in subsequent sections. In a later paper, Ladany (1975) developed a cost model for the compressed limit sampling plans. This model took into account the costs of destroying a unit (as a result of failing the test) as well as the cost of testing at the limit $L_c$ versus the usual limit $L$.

# 3 Some Theoretical Results

## 3.1 Inadmissible and Uniformly Efficient X-Transforms

Consider testing the null hypothesis $\mathrm{H}_0 : R = R_0$ vs. $\mathrm{H}_a : R = R_a$ with $R_0 < R_a$. We refer to an X-transform as *inadmissible* for the ordered pair $(R_0, R_a)$ if the corresponding X-test always results in an equal or lower power for fixed sample size and zero-failure plans and an equal or larger sample size for fixed power plans compared to the analogous traditional plans for any level $\alpha$. The corresponding X-tests are always inefficient and should not be used, whether the application is to zero-failure plans, fixed sample size plans, or fixed power plans.

The following proposition gives a sufficient condition to determine if an X-transform is inadmissible. For this proposition and hereafter we will only consider X-transforms with $\tau_0 \leq R_0$. The results can be applied to cases in which $\tau_0 > R_0$ by interchanging the X-test and the traditional test plan. All proofs are deferred to the Appendix.

**Proposition 1.** *An X-transform is inadmissible for* $(R_0, R_a)$ *if* $\frac{\tau(R_a)}{\tau(R_0)} \leq \frac{R_a}{R_0}$.

Conversely, we refer to an X-transform as *uniformly efficient* for $(R_0, R_a)$ if the corresponding X-test always results in an equal or larger power for fixed sample size and zero-failure plans and a smaller or equal sample size for fixed power plans compared to the analogous traditional plans for any level $\alpha$. The following proposition gives a sufficient condition to determine if an X-transform is uniformly efficient.

**Proposition 2.** *An X-transform is uniformly efficient for* $(R_0, R_a)$ *if* $\frac{1-\tau(R_a)}{1-\tau(R_0)} \leq \frac{1-R_a}{1-R_0}$.

In addition to comparing X-tests to traditional plans, these two results can be useful in deciding to what extent it is desirable to increase the failure probability in an X-test. For example, suppose we want to decide whether to increase the test time by one hour or two hours. We can then check if the X-transform that maps from one hour of additional testing to two hours of additional testing is inadmissible or uniformly efficient for the pair $(\tau_1(R_0), \tau_1(R_a))$, where $\tau_1$ is the X-transform mapping the traditional test to the test using only one additional hour.

Examples of inadmissible and uniformly efficient X-transforms are given in the following section, in Section 8.1, and in Mease (2003).

## 3.2 Location X-Transforms

Recall that location X-transforms are defined to be of the form $Q_F(1 - \tau(R)) = Q_F(1 - R) + A$ with $A = Q_F(1 - \tau_0) - Q_F(1 - R_0)$. We refer to these as the class of location X-transforms corresponding to the family $F$. For example, $F = \Phi$, the standard normal CDF, leads to normal location X-transforms. A specific value of $\tau_0$ specifies a single X-transform in the class. However, the same normal class can be generated from any normal distribution with mean $\mu$ and variance $\sigma^2$. To resolve this ambiguity, we will use throughout the standard CDF in a location-scale family with scale parameter one and location parameter zero.

As we saw in Section 2, location X-transforms can arise in the context of different X-tests. They arise naturally for common reliability distributions such as Weibull and lognormal which can be reduced to location-scale after a log-transformation of the data. We will also consider other models such as the double exponential and Cauchy in the sequel to illustrate some technical properties, although the log location-scale versions of these models are less common in reliability.

Location X-transforms can be inadmissible or uniformly efficient depending on the underlying location family.

**Proposition 3.** *Any X-transform from a location class with standardized CDF $F$ is inadmissible for $(R_0, R_a)$ if the corresponding hazard function $h(x)$ is non-increasing for $x \in (Q_F(1-R_a), Q_F(1-\tau(R_0)))$.*

Consider the double exponential (DE) location X-transforms. The hazard function $h(x)$ for this distribution is constant for $x > 0$, so any X-transform from the class will be inadmissible for $(A, B)$ if $B \le 1 - F_{\mathrm{DE}}(0) = .5$ where $F_{\mathrm{DE}}$ is the standard DE CDF. One can show that it is not useful to induce failure below $\tau(R_a) = .5$ for this class since using $\tau(R_a) = .5$ along with randomization would give the same power with the same sample size.

For the Cauchy family, the hazard function $h(x)$ is decreasing for $x > 1$. Consequently, any X-transform will be inadmissible for $(A, B)$ if $B \le 1 - F_{\mathrm{Cauchy}}(1) = .25$. Thus, if $\tau_1$ is the Cauchy location X-transform such that $\tau_1(R_a) = .25$ and $\tau_2$ is the Cauchy location X-transform that maps $R_a$ to a value $\tau_2(R_a) < .25$, we can conclude that $\tau_2$ cannot outperform $\tau_1$ in either the fixed power or fixed sample size problem. To see this, note that $\tau_2(\cdot)$ can be expressed as $\tau^*(\tau_1(\cdot))$ and $\tau^*$ is inadmissible for $(\tau_1(R_0), \tau_1(R_a))$. From this one can conclude that in searching for the optimal X-test from this class in either the fixed power or fixed sample size problem, one can limit the search to X-transforms $\tau$ such that $\tau(R_a) \ge .25$.

Since the condition for uniform efficiency in Proposition 2 is based on the complement of the reliability, we can obtain an analogous result to Proposition 3 by considering the hazard rate for the negative of a random variable.

**Proposition 4.** *Any X-transform from a location class with standardized CDF $F$ is uniformly efficient for $(R_0, R_a)$ if the hazard function $\tilde{h}(x)$ of the random variable $-X$ is non-increasing for $x \in (-Q_F(1-\tau(R_0)), -Q_F(1-R_a))$ where $X$ has CDF $F$.*

Consider again the Cauchy location class. Since the Cauchy distribution is symmetric and the hazard function $h(x)$ for the standardized Cauchy CDF is decreasing for $x > 1$, it follows any X-transform from the class will be uniformly efficient for $(R_0, R_a)$ if $\tau(R_0) \ge F_{\mathrm{Cauchy}}(1) = .75$. Thus, any X-test employing a Cauchy location X-transform with $\tau_0 \ge .75$ would be preferable to the traditional reliability demonstration plan for zero-failure plans as well as fixed sample size and fixed power plans.

# 4 Zero-Failure Plans: Power and Sample Size Behavior

Zero-failure X-tests are of special interest for several reasons. First, the results of zero-failure plans for binary data also apply to situations in which time-to-failure (TTF) data or other auxiliary information can be obtained from failures. The reason is that in a successful zero-failure plan there are no failures, so tests based on TTF data are the same as those based on binary data. Second, just as traditional zero-failure demonstration plans have the smallest sample size possible without inducing failure, zero-failure X-tests have the smallest sample size for a given degree of increased failure probability (i.e., for a given value of $\tau_0$). This is important since inducing failure implies increased uncertainty, which makes it desirable to keep $\tau_0$ close to $R_0$.

Consider again the example in Section 1 with the goal of demonstrating a reliability of $R_0 = .90$ with 95% confidence. Suppose the traditional zero-failure plan of testing $n = 29$ units is too costly, and one can afford to test only $n = 20$ units. Then, the failure probability must be inflated so that $\tau(R_0) = .86$. We get $.86^{20} \approx .05$, so a zero-failure plan satisfies the required confidence level.

The sample size for a zero-failure X-test is given by

$$n = \left\lceil \frac{\log(\alpha)}{\log(\tau_0)} \right\rceil \tag{8}$$

where $\lceil x \rceil$ is the ceiling function. Analogous sample size formulas to (8) are given by Meeker and Escobar (1998) for inducing failure by increasing test time and by Kececioglu (1994) for inducing failure by increasing stress in time-to-failure models. There is, however, not as much discussion in the literature on the effect on the power of the tests.

In the above example, suppose failures were induced by increasing the number of cycles tested and the number of cycles to failure was assumed to follow a Weibull distribution with known shape parameter $\beta$. The log of the cycles to failure would follow a SEV distribution with known scale parameter $\sigma = 1/\beta$. Then, the resulting X-transform is given by $Q_{\text{sev}}(1-\tau(R)) - Q_{\text{sev}}(1-R) = Q_{\text{sev}}(1-\tau(R_0)) - Q_{\text{sev}}(1-R_0)$ where $Q_{\text{sev}}$ is the quantile function of the standard SEV distribution. One can verify that the power of the zero-failure X-test is the same as the power for the traditional plan regardless of the true value of $R$ in this case. Thus, one is able to reduce the sample size without any loss in power.

For the lognormal distribution with known parameter $\sigma$, the X-transform is given by $Q_{\text{nor}}(1-\tau(R)) - Q_{\text{nor}}(1-R) = Q_{\text{nor}}(1-\tau(R_0)) - Q_{\text{nor}}(1-R_0)$. It can be shown that the power of the zero-failure X-test for this case is less than that of the traditional plan for all values of $R > R_0$. For example, if $R_a = .99$, the power of the traditional plan is $.99^{29} = .747$ while the power for the X-test is $[\tau(.99)]^{20} = .9832^{20} = .713$. This loss in power is small, so the X-test can still be desirable if the reduction in sample size is big enough to offset the loss in power.

The power behavior of zero-failure plans is simple to study since we have closed-form expressions for the sample sizes needed and consequently the power of the plans. The power of a zero-failure X-test is $\tau(R_a)^{\left\lceil \frac{\log(\alpha)}{\log(\tau_0)} \right\rceil}$ which will be increasing, constant, or decreasing as the sample size is decreased depending on whether

$$\tau(R_a)^{\frac{\log(\alpha)}{\log(\tau_0)}} \tag{9}$$

is increasing, constant, or decreasing as $\tau_0$ decreases.

We consider now some specific results for location X-tests. Let $h(x)$ be the hazard rate and $H(x)$ be the cumulative hazard function for the CDF $F$. The following can be established by using (9).

**Proposition 5.** *The power of a zero-failure X-test based on the location transform with standardized CDF $F$ is increasing, constant, or decreasing as $\tau_0$ decreases if the function $h(x)/H(x)$ is increasing, constant, or decreasing in $x$ for $x \in (Q_F(1-R_a), Q_F(1-\tau_0))$.*

Since $H(x)$ is always non-decreasing, it follows that $h(x)/H(x)$ will be decreasing if the hazard rate is decreasing. In fact, we saw earlier that decreasing hazard rate implies the stronger condition of inadmissibility.

For the standardized SEV distribution, $h(x)/H(x)$ is constant, so the power remains constant as the sample size is decreased. This can also be seen by expressing the X-transform as $\tau(R) = R^{\frac{\log(\tau_0)}{\log(R_0)}}$ and noting that the power does not depend on $\tau_0$.

For most common distributions, $h(x)/H(x)$ is decreasing for all values of $x$, so the power of the zero-failure X-test will decrease as sample size is decreased. This was shown to be true for the normal distribution by Mease (2003) and is illustrated in Figure 1. Figure 2 suggests that this may also be true for the logistic distribution. In such cases, the X-tests are preferable only if the benefits of sample size reduction outweigh the consequences of loss in power.

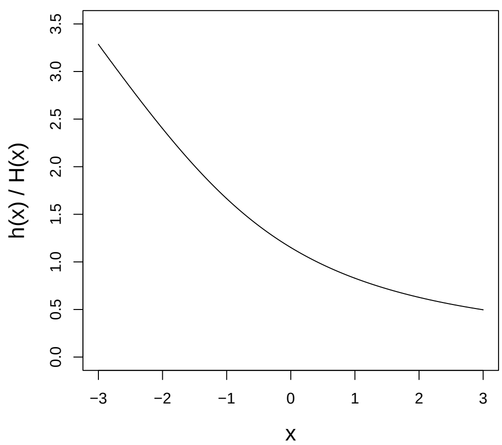


Figure 1: The Function $h(x)/H(x)$ for the (Standard) Normal CDF

As shown in Figure 3, $h(x)/H(x)$ is increasing for $x < 0$ for the standardized DE distribution. Thus, if $\tau_0 \geq .5$ the zero-failure X-test will actually have *greater* power than traditional plan (despite having an equal

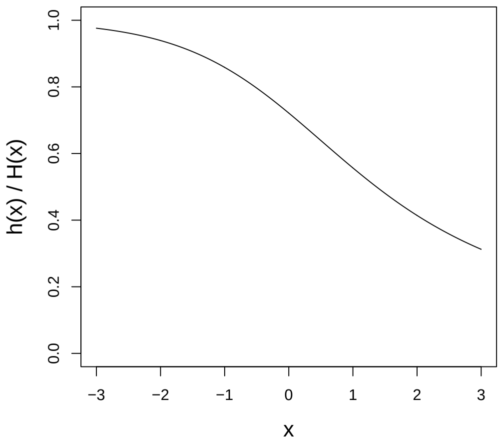


Figure 2: The Function $h(x)/H(x)$ for the (Standard) Logistic CDF

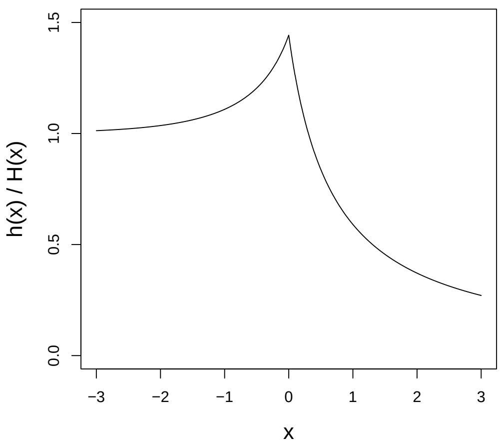


Figure 3: The Function $h(x)/H(x)$ for the (Standard) Double Exponential CDF

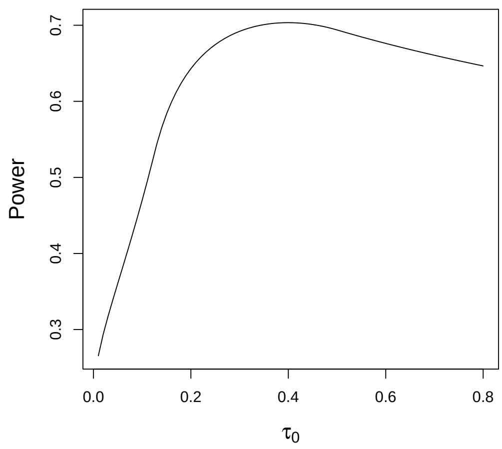


Figure 4: Power for Zero Failure X-tests from the DE Location Class with $\alpha = .15, R_0 = .8$ and $R_a = .95$

or smaller sample size). Moreover, any X-transform $\tau_1$ from the class with $\tau_1(R_0) \geq .5$ will give a larger power than a second X-transform $\tau_2$ from the class if $\tau_2(R_0) > \tau_1(R_0)$. The implication of this is that, in theory, if $R_a > .5$ one should always choose $\tau_0$ to be at most .5; however, costs associated with using extreme testing as well as concerns over model validity will generally prevent one from choosing $\tau_0$ too far from $R_0$ in practice.

Figure 3 also shows that $h(x)/H(x)$ is decreasing for the DE distribution for $x > 0$. Thus, if $R_a < .5$, the zero-failure X-test will necessarily have less power than the traditional zero-failure plan. However, values of $R_a < .5$ are rare in practice. More importantly, for $R_a > .5$, the power for DE location X-tests is decreasing as $\tau_0$ decreases for values of $\tau_0$ such that $\tau(R_a) < .5$. This can be seen by writing the DE location X-transform $\tau$ as the composition $\tau_1(\tau_2(\cdot))$ where $\tau_2(R_a) = .5$ and applying the result to $\tau_1$. For example Figure 4 shows the power of the DE location class with $\alpha = .15$, $R_0 = .8$ and $R_a = .95$. The power is in fact increasing as $\tau_0$ decreases for $\tau_0 > .5$ as proven by the argument in the previous paragraph and decreasing as $\tau_0$ decreases for $\tau_0 < F_{\text{DE}}(Q_{\text{DE}}(1 - R_a) - Q_{\text{DE}}(1 - R_0)) = .125$ as guaranteed by the argument in this paragraph. ($F_{\text{DE}}$ and $Q_{\text{DE}}$ denote the CDF and quantile functions of the standardized DE distribution.)

# 5 Power Behavior for Fixed Sample Size Plans

We now consider how power varies for $k$-out-of-$n$ X-tests when the sample size $n$ is fixed. As an example, suppose that the sample size of $n = 29$ for the zero-failure plan in the previous section was in fact an affordable sample size, but that the resulting power of .747 for a true reliability of $R_a = .99$ was unacceptably low. X-testing can be used to potentially increase the power with same fixed sample size by increasing the number of allowable failures $k$. For example, if failure probability was inflated so that the reliability under the null hypothesis was reduced to $\tau_0 = .847$, then a one-failure plan would meet the required $\alpha$ level for the same sample size of $n = 29$ since $BIN(1;\ 29, 1 - .847) = 0.05$.

Whether the power of this one-failure X-test is larger or smaller than the traditional zero-failure plan would depend on the X-transform. For instance, suppose the X-transforms belong to the normal location class so that

$$\tau(R) = 1 - \Phi(Q_{\text{nor}}(1 - R) + Q_{\text{nor}}(1 - \tau_0) - Q_{\text{nor}}(1 - R_0)). \tag{10}$$

If $R = 0.99$, $\tau(.99) = .9807$ and the power of the X-test is $BIN(1;\ 29, 1 - .9807) = 0.893$. This is a significant increase over the power of the traditional zero-failure plan. One could also consider using a two-failure X-test. This is done by increasing the failure probability even further such that $\tau_0 = .798$ (since $BIN(2;\ 29, 1 - .798) = .05$). For the normal location X-transforms, this results in a further increase of power to .945.

The extent to which the power will increase (or decrease) as failure probability is increased depends on the specific class of X-transforms. It is difficult to develop analytical results in general. The choice of $k$ and $\tau_0$ that will yield the largest power must be determined by a numerical search over all possible values of $k$ and $\tau_0$ that solve $BIN(k;\ n, 1 - \tau_0) = \alpha$.

Inducing failure as a means of increasing power for a fixed sample size was considered briefly by Sibuya and Suzuki (2001) who examined inducing failure through increasing test time. They carried out some numerical calculations for $n = 3$ for a number of different time-to-failure distributions including normal, lognormal, logistic, negative Gumbel, and Weibull. They concluded that "it looks difficult to develop a general theory" (p. 199) and acknowledged "the choice of $k$ is still an open problem" (p. 191).

We now consider optimal plans for some specific classes of location X-transforms. Numerical results for the normal, SEV, logistic, and Cauchy location X-transform classes for various values of $n$, $R_0$ and $R_a$ are displayed in the plots in Figure 5 and given in tables in Mease (2003). For each of these distributions, as the failure probability is increased (i.e., as $\tau_0$ is decreased), the power generally initially increases, reaches a maximum, and then decreases. (The last three plots in Figure 5 only show a limited range for $\tau_0$.) In the case of the SEV location class, the power usually reaches this maximum when $k$ is close to $.8 \times n$, while for the other three location classes considered (which are all based on symmetric distributions), this peak generally occurs when $k$ is close to $n/2$.

It is important to note that when searching over the values for $k = 0, 1, ..., n - 1$, the value of $k$ and corresponding value of $\tau_0$ that yield the largest power may not be the global optimum since randomized plans are being ignored. For example, if $R_0 = .6$, $n = 2$, and $\alpha = .36$, the two possible non-randomized plans are $k = 0$ with $\tau_0 = R_0 = .6$ or $k = 1$ with $\tau_0 = .2$. However, if failure is induced instead so that $\tau_0 = .3$, one would employ the randomized plan which rejects the null hypothesis always when there are zero failures and 64.29% of the time when exactly one failure occurs since $.3^2 + .6429[2(.3)(.7)] = .36 = \alpha$. Sibuya and Suzuki (2001) did not consider randomized plans in their analysis.

For the location X-transforms that we have studied, such randomized plans generally do not exceed the power of neighboring non-randomized plans. This can be verified by plotting the power of all the test plans (including the randomized plans) as a continuous function of $\tau_0$ and noting that the power for the non-randomized plans are local maxima. This behavior can be observed in the plots in Figure 5 which show the randomized plans as lines connecting the symbols for the non-randomized plans. In such cases, ignoring the randomized plans does not miss the optimum. However, examples can be constructed in which this does not hold true. One such case is obtained by using the example in the above paragraph with the class of Cauchy location X-transforms and $R_a = .8$. The power for the zero-failure (traditional) plan is then $.8^2 = .64$, as is the power for the one-failure X-test since $.4^2 + 2(.4)(.6) = .64$. However, the power for the randomized plan described is .669 which is slightly larger than the power for both of the possible non-randomized plans.

# 6 Sample Size Behavior for Fixed Power Plans

In the fixed power problem, the goal is to reduce the sample size while maintaining a power at least as large as a specified level. Suppose as in the example before that the goal is to demonstrate a reliability of $R_0 = .90$ with 95% confidence, but now it is required that the power is at least .95 when the true reliability is $R_a = .99$. The traditional zero-failure plan at $\alpha = .05$ and $n = 29$ test units results in a power of just .747. Thus a traditional plan will require more than $n = 29$ test units to achieve the required power. However, by using an X-test it may be possible to meet both the power and confidence level requirements with the sample size of $n = 29$. For an X-test with the normal location X-transform with $\tau_0 = .754$, a $k = 3$-failure plan will achieve this power requirement with $n = 29$ test units. Recall this plan is one of the plans displayed in the first plot in Figure 5. Thus, the X-test is preferred over the traditional plan as it has a smaller sample size while still meeting the power requirements.

It is difficult to characterize in general the values for $\tau_0$ that will lead to the test plan with the smallest sample size $n$ that meets both the power and level requirements. In most cases it is necessary to carry out a numerical search over the possible values of $k = 0, .., n-1$ for several values of $n$. However, when the required power is equal to $1 - \alpha$ (where $\alpha$ is the required level of the test), we can often limit the values of $\tau_0$ that must be considered. Specifically, suppose that we have a $k$-failure X-test that maps $R_0$ to some value $\tau_0^1 > .5$ and has level equal to $\alpha$ (exactly) but does meet the required power of $1 - \alpha$ when the true reliability is $R_a$. Let $\tau_a^1$ denote the value to which $R_a$ maps under this X-transform. Next consider a second X-transform from the same class which increases failure probability even more so that $R_0$ maps to $1 - \tau_a^1$. From the symmetry of the binomial distribution, we know that for this second X-transform there is an $(n - k)$-out-of-$n$ failure plan which has a level equal to 1 minus the power of the first plan and thus greater than $\alpha$. It follows that for this second plan if $R_a$ maps to a value less than or equal to $1 - \tau_0^1$, the power will be less than or equal to the required $1 - \alpha$. Consequently, we see that the new value of $\tau_0$ corresponding to the X-transform for the second plan cannot give a test with level $\alpha$ that will meet the required power of $1 - \alpha$. We do not need

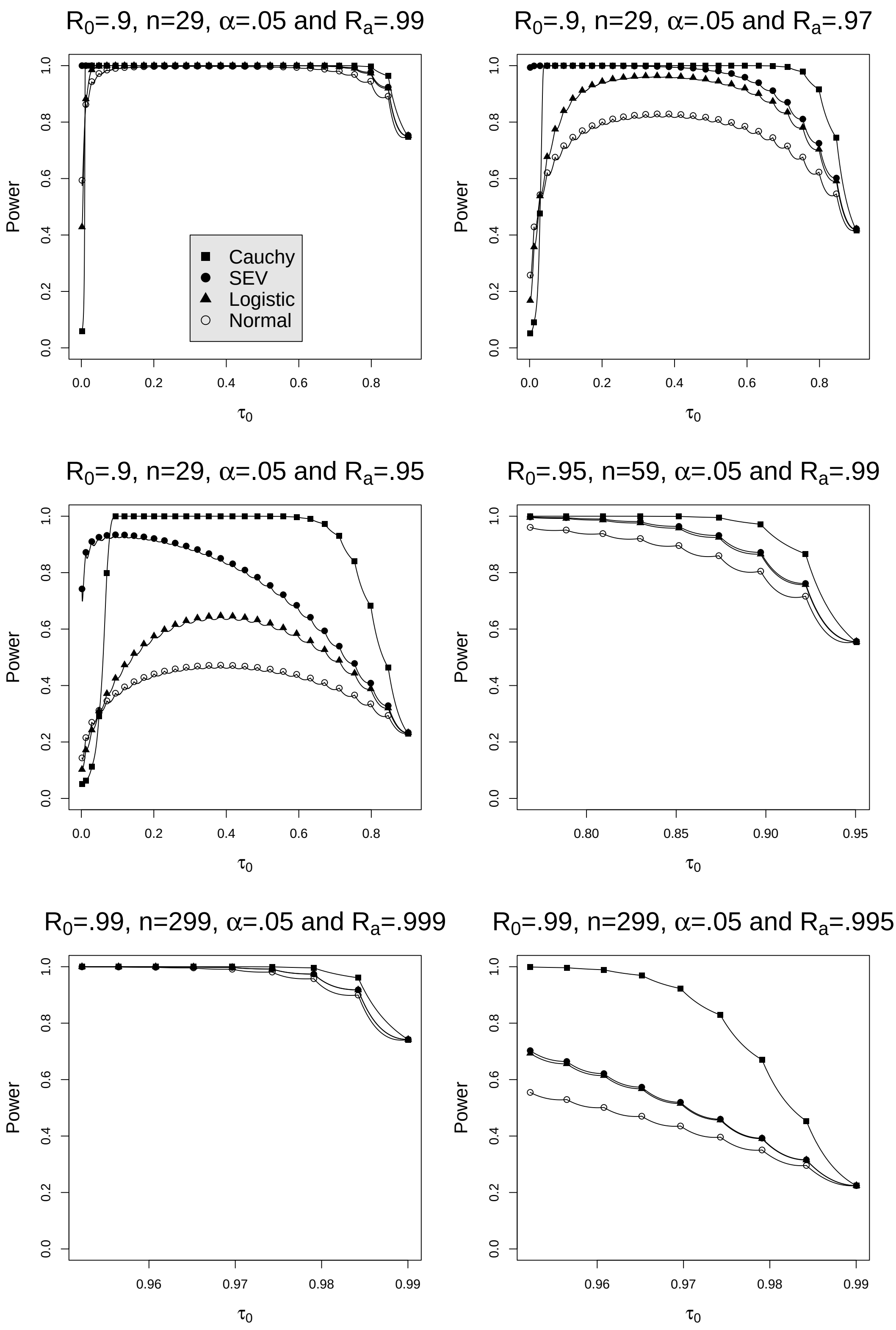


Figure 5: Power for Four Classes of Location X-Transforms Including Randomized Plans

to consider this second value of $\tau_0$ in our search since it will not allow us to meet the requirements, even though it corresponds to a further increase in failure probability over the first plan. This situation arises for any class of location X-transforms for which the distribution is symmetric. This result was also noted by Sibuya and Suzuki (2001). Conversely, if this second X-transform maps $R_a$ to a value larger than $1-\tau_0$, when the second X-transform cannot be used to construct a plan meeting the requirements then we know the first cannot either, and we do not need to consider the value of $\tau_0^1$ in our search. It can be shown that this latter result holds for the class of SEV location X-transforms.

Beja and Ladany (1974) also considered the problem of finding the optimal plan to minimize the sample size for a fixed level and power. Their method reduces to the normal location X-transforms. They noted that no closed form solution was available, and as a practical guideline they suggested using a value for $L_c$ that we recognize as being equivalent to using $\tau_0 = 1 - \tau(R_a)$. In a later paper Ladany (1976) gave two different suggestions for $L_c$. One was based on using a normal approximation to the binomial distribution and the other involved a graphical method.

# 7 Practical Implications of Misspecified X-Transforms

The results so far are based on the assumption that the given X-transform in the X-test is correct. However, in practice the X-transform is typically estimated from field data, often from related products, and may only be an approximation to the true function. This section considers the practical consequences of using a X-transform that differs from the true function.

One way an X-transform can differ from the true function is by overestimating or underestimating the amount by which the failure-inducing process would decrease the reliability from the target value of $R_0$. This would lead to a value of $\tau(R_0)$ that is not correct. Since choosing the smallest value of $n$ for which (2) holds depends on the assumed value of $\tau(R_0)$, this in turn leads to the level of the X-test being different from $\alpha$. Specifically, if the assumed value of $\tau(R_0)$ is less than the true value, the level will be greater than $\alpha$, while if the assumed value of $\tau(R_0)$ is greater than the true value, the level will be less than $\alpha$. For this reason, it is suggested that one errs on the side of smaller values of $\tau(R_0)$ in order to be conservative (Meeker and Escobar, 1998).

As an example, suppose that a manufacturer of an expensive electrical device is required to demonstrate with 95% confidence that at least 99.9% of the devices produced by the factory can operate continuously for at least 10 hours. Use of a traditional zero-failure demonstration plan would require testing $n = 2995$ of the devices for 10 hours each since $.999^{2995} \approx .05$. This is a very large number, so the manufacturer may consider using an X-test in which the units are tested instead for 30 hours to reduce the sample size. However, in order to know the X-transform, the manufacturer must know the distribution of the failure times up to one parameter. When there is uncertainty in this knowledge, one should err on the side of smaller values of $\tau(R_0)$. For instance, suppose that based on data from similar products, the manufacturer believes that the failure times of these devices follow a Weibull distribution with a shape parameter between 1 and 2. If the shape parameter is equal to 1 (which implies an exponential distribution), the X-transform resulting from testing for 30 hours instead of 10 hours is $\tau_1(R) = R^3$, while if the shape parameter is equal to 2, the corresponding X-transform is $\tau_2(R) = R^9$. Thus, in order to be conservative, the manufacturer should use $\tau_1$. The resulting zero-failure X-test would require testing $n = 999$ units for 30 hours each. Now if in fact the failure times were Weibull with a shape parameter of 2, this would be a .0001 level test.

In addition to affecting the level, a misspecified X-transform can also affect the power. In the case of using a conservative value of $\tau(R_0)$ as described above, this will lead to a lower power than if the X-transform assumed were correct. In the example described above, if the failure time distribution is in fact Weibull with a shape parameter of 1 so that the assumed X-transform $\tau_1$ is correct, the power of the X-test would be the same as the power of the traditional test (a property of SEV location X-transforms), so that if the true reliability is $R_a = .99999$ the power would be equal to .97. However, if the true distribution is instead Weibull with a shape parameter of 2, the resulting power would be .91. This lower power is the consequence of using a conservative value for $\tau(R_0)$.

Even in cases in which the value of $\tau(R_0)$ assumed is correct, if the true X-transform is different from the one assumed, the actual power may differ substantially from the predicted power. To illustrate this, we return to the above example in which $\tau_1$ was the assumed X-transform, but suppose that instead of having the assumed Weibull distribution with a shape parameter of 1, the failure time distribution is actually lognormal with a shape parameter of 3.21 so that the true X-transform is given by $\tau_3(R) = 1 - \Phi[Q_{\text{nor}}(1 - R) + \frac{\log(30)-\log(10)}{3.21}] = 1 - \Phi[Q_{\text{nor}}(1-R) + Q_{\text{nor}}(1 - .999^3) - Q_{\text{nor}}(1 - .999)]$. Since $\tau_3(R_0) = \tau_1(R_0)$ the X-test will still have exactly level $\alpha = .05$; however, a property of normal location X-transforms is that the power will be less than that of the traditional test. For the value of $R_a = .99999$ in the example, the power would be roughly .96 versus the value of .97 that would result from the traditional test as well as from the X-test had $\tau_1$ been the correct X-transform.

X-transforms can also be misspecified if the method of increasing the failure probability induces a new failure mode different from the mode of failure under standard conditions. For instance, in the example above suppose that the time to failure distribution up to 10 hours is in fact Weibull with a shape parameter of 1, but that a second independent failure mode causes 0.1% of the products to fail between 10 and 30 hours of use. Further, suppose that in the testing process the failures resulting from this second failure mode cannot be distinguished from those resulting from the first failure mode. Since the reliability of interest is the reliability at 10 hours, this second failure mode is not of interest and may not be known to the manufacturer. However, this second failure mode can cause products to fail during the 30-hour X-test and as such would affect both the level and the power of the X-test. Specifically, the level of the X-test will be equal to .018 instead of the assumed .05, while the power would be only .357 assuming the true reliability is $R_a = .99999$. Note that this reduction in power is much more severe than in the case of the conservative X-test described before which had an even lower level.

We conclude that when there is a substantial amount of uncertainty in the failure inducing process and thus the X-transform, one should be conservative in the use of X-testing. This is necessary to ensure the level does not exceed $\alpha$, but can substantially reduce the power. Moreover, one should always consider not only the consequences of the power if the assumed X-transform is correct (as has been studied in the previous sections), but also the consequences if the assumed X-transform is misspecified and instead various other X-transforms more accurately represent the true failure inducing process.

# 8 Additional Examples

This section considers two additional examples of X-testing with binary data. These examples serve to illustrate how the methodology and results developed in the previous sections can be applied, and also motivate the extension of the results beyond the context of the physical reliability experiments considered up to this point.

## 8.1 A Structural Reliability Computer Experiment

This example is an application dealing with a computer experiment. A more complex version of the stress-strength model is a structural reliability model in which reliability is given by $\mathrm{P}(g(W) > 0)$, where the random vector $W$ is a collection of what are called basic variables and can include loading, strength and geometric variables and $g$ is any function separating the "safe region" from the "failure region" (Madsen, Krenk and Lind, 1986). The distribution of $W$ is usually known and the function $g$ is also known. However, $g$ is often not available in closed form, and it may take hours or even days to evaluate for a single input vector using a high-speed computer. Thus, the problem of computing the reliability is a problem of numerical integration of an indicator function with respect to a density.

In this example, we will consider a computer experiment to demonstrate that the reliability $R$ is at least .99999 with 90% confidence. By generating $n = 230258$ pseudo-random vectors from the distribution of $W$ and evaluating $g$ for each vector, a zero-failure plan can be used to demonstrate the reliability. In other words, if all $n = 230258$ function evaluations of $g$ are returned by the computer as positive values, the reliability can be said to be demonstrated with the required confidence, since $R_0^n = .99999^{23058} \approx .10$.

While the above demonstration plan would be fine for function evaluations that can be done quickly, for function evaluations that take a substantial amount of time this would be impractical. In such cases, further information about the function $g$ can be used to reduce the number of function evaluations necessary in an X-test. In our example, we will suppose that it is known that the closest point to the origin for which $g(x) \leq 0$ is the point $x^*$. (In structural reliability (Madsen, Krenk and Lind, 1986) this point is called the MPP - Most Probable Point - and various algorithms are available to determine this point.) Therefore it can be assumed that $g(x) > 0$ within the sphere with its center at the origin and radius $||x^*||$. Thus an X-test can be employed by sampling $W$ from a new distribution which is the same as the original distribution of

$W$ but restricted to the set that is outside of this sphere.

The degree to which failure is induced and thus sample size is reduced in this example will depend on the value of $||x^*||$. For instance, suppose that $W$ is multivariate standard normal. For $||x^*|| = 4$ we have $\tau(R_0) = 1 - \frac{10^{-5}}{1-\Phi(4)} = .6843$ requiring a sample size of only $n = 7$ evaluations of $g$ to carry out the reliability demonstration. Furthermore, it can also be verified that while the sample size is reduced, the power will slightly increase. For example, with a value of $R_a = 1 - 10^{-6}$ the power of the traditional plan would be .794 while for the X-test described the power would be .799. The reason for this is that the X-transform is uniformly efficient for any ordered pair $(R_0, R_a)$ with $0 < R_0 < R_a < 1$ which is easily verified by noting its parameterization in terms of $\tau_0$ is $\tau(R) = 1 - \frac{1-\tau_0}{1-R_0}(1 - R)$ regardless of the distribution of $W$.

### 8.2 Optimal Binary Measurement for A Hypothesis Test

This example illustrates how the results can be applied in a much more general context than reliability testing. Suppose $W_1$,...,$W_n$ are $n$ independent observations of the random variable $W$ which is normal with unknown mean $\theta$ and a known standard deviation equal to one. Further suppose that we wish to test the null hypothesis that $\theta \leq 0$ versus the alternative that $\theta > 0$ in an $\alpha = .05$ level test with a sample size of $n = 29$. We can collect binary data $I_1, ..., I_n$ where $I_i$ is the indicator that $W_i$ is greater than a fixed constant $w$. The problem is to determine a good choice for the value of $w$.

We will begin by relating this problem to reliability demonstration plans. Let $R = \mathrm{P}(W > Q_{\text{nor}}(.1))$ so that the null hypothesis becomes $R \leq .90$. Thus taking $w = Q_{\text{nor}}(.1)$ corresponds to a traditional reliability demonstration plan, while using larger values for $w$ can be thought of as X-tests. For the traditional plan, a zero-failure plan can be used since $.9^{29} \approx .05$. For an X-test, numeric searches over the values $k = 0, ..., 28$ show that for $R_a > R_0 = .90$ the largest power usually occurs when $k$ is close 13. This can be seen, for example, in the second and third plots in Figure 5. Since an $\alpha = .05$ level X-test with $k = 13$ corresponds to $\tau_0 = .384$, a good choice for $w$ is the value of $w$ which corresponds to $\tau_0 = .384$ which is $w = Q_{\text{nor}}(1 - .384) = 0.295$.

## 9 Discussion

We have shown that the use of X-testing can lead to substantial gains in efficiency, in terms of both power and sample size reduction in many cases. We have introduced a methodology for quantifying the effects of X-testing by way of studying functions which are named X-transforms. These X-transforms exclusively determine the efficiency of the X-tests and are the same across many different applications, allowing us to give results that apply very generally. Specifically we have explored the performance of the X-transforms from the normal, SEV, logistic, and Cauchy location classes with respect to zero-failure plans, fixed sample size plans, and fixed power plans. Furthermore, we have seen that X-transforms can be uniformly efficient or inadmissible and have given conditions under which these situations occur.

Table 1: Power for Normal Scale X-Transforms with $R_0 = .85$, $n = 150$, and $\alpha = .05$ when $R_a = .95$

| Number of Failures Allowed ($k$) | $\tau(R_0)$ | Power |
|---|---|---|
| 0 | 0.980 | 0.922 |
| 5 | 0.931 | 0.997 |
| 10 | 0.890 | 0.998 |
| 11 | 0.882 | 0.998 |
| 12 | 0.874 | 0.998 |
| 13 | 0.866 | 0.997 |
| 14 | 0.858 | 0.997 |
| 15 | 0.850 | 0.997 |
| 16 | 0.843 | 0.996 |
| 17 | 0.835 | 0.995 |
| 18 | 0.827 | 0.994 |
| 19 | 0.820 | 0.993 |
| 20 | 0.812 | 0.992 |
| 30 | 0.739 | 0.945 |
| 40 | 0.667 | 0.756 |
| 50 | 0.598 | 0.400 |
| 60 | 0.530 | 0.112 |

While we have focused mostly on classes of location X-transforms, in some cases it is useful to define analogously *scale X-transforms* (Mease, 2003). Also, we have only considered examples for which efficiency is gained by increasing the failure probability, while in fact there are cases where decreasing the failure probability might be beneficial. We can illustrate this point as well as provide an illustration of a scale X-transform through the following example. Suppose that a manufacturer of a certain device must demonstrate a reliability of $R_0 = .85$ at 1000 hours and believes that the failure times follow a lognormal distribution $LN(\mu, \sigma)$ for which $\mu$ is known but $\sigma$ is unknown (which is not the usual case). If he or she wishes to have 95% confidence and has $n = 150$ devices available for testing, then the traditional plan would use $k = 15$ since $BIN(15;\ 150, 1 - .85) \approx .05$. If the manufacturer believes that the true reliability of the products is near $R_a = .95$ then he or she should expect a power near .997 since $BIN(15;\ 150, 1 - .95) \approx .997$. Now using the knowledge of the failure time distribution and $\mu$ the manufacturer could employ an X-test by testing longer than 1000 hours with the hope of increasing the power. However, consider Table 1 which gives the power for different values of $\tau_0$. It is apparent from the table that, for $R_a = .95$, increasing the failure probability will not increase the power since all values of $\tau(R_0)$ less than .85 give a smaller power than the .997 value for the traditional plan. In this case, the knowledge of the failure time distribution can instead be used to actually decrease the failure probability by testing the products for a period of time that is less than 1000 hours. For instance if the units were tested for a period of time corresponding to $\tau(R_0) = .874$, the power of the test would increase slightly to .998 as shown in Table 1.

Finally, we note that while we have approached the problem from an optimization/minimization angle with respect to power and sample size, there are also other issues that must be considered when deciding to what extent X-testing should be used, if at all. For instance, if there are costs that increase as failure probability is increased, selection of the X-test should take into account these cost considerations in addition to power. Second, in many cases the uncertainty in the underlying model and concerns over extrapolation may prohibit inducing failure beyond a certain point. In practice benefits predicted from assumed X-transforms must be weighed against uncertainty associated with extrapolation.

## Acknowledgments

This research was supported in part by NSF grant DMS-0204247 and a Ford URP grant. Mease's research was also supported by a University of Michigan Rackham Predoctoral Fellowship and an NSF-DMS Postdoctoral Fellowship. The authors are grateful to Agus Sudjianto and Steve Wang for helpful discussions in formulating the problem, to Jeff Robinson for useful comments, and to the Associate Editor and two referees for a careful reading of the paper and a number of constructive comments.

# References


Beja, A. and Ladany, S. P. (1974), "Efficient Sampling by Artificial Attributes," *Technometrics*, 16, 601-611.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society Series B*, 34, 187-220.

Joseph, R. V. and Wu, C. F. J. (2004), "Failure Amplification Method: An Information Maximization Approach to Categorical Response Optimization," *Technometrics*, 46, 1-12.

Kececioglu, D. (1994), *Reliability and Life Testing Handbook*, Englewood Cliffs, NJ: Prentice Hall.

Ladany, S. P. (1975), "Least Cost Acceptance Sampling For Destructive Testing," *Journal of Quality Technology*, 7, 123-126.

Ladany, S. P. (1976), "Determination of Optimal Compressed Limit Gaging Sampling Plans," *Journal of Quality Technology*, 8, 225-231.

Lipson, C. and Sheth, N. J. (1973), *Statistical Design and Analysis of Engineering Experiments*, New York: McGraw-Hill.

Madsen, H.O., Krenk, S., and Lind, N.C. (1986), *Methods of Structural Safety*, Englewood Cliffs, NJ: Prentice-Hall.

Mann, N. R., Schafer, R. E. and Singpurwalla, N. D. (1974), *Methods for Statistical Analysis of Reliability and Life Data*, New York: Wiley.

Martz, H. F. and Waller, R. A. (1982), *Bayesian Reliability Analysis*, New York: Wiley.

Mease, D. (2003), "Contributions to Engineering Statistics: X-Testing for Reliability and Optimal Partitioning Strategies," Unpublished Ph. D. Dissertation, University of Michigan.

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Nair, V. and Wang, X. (2004), Discussion of "Failure Amplification Method: An Information Maximization Approach to Categorical Response Optimization" by Joseph and Wu, *Technometrics*, 46, 19-23.

Nelson, W. (1990), *Accelerated Testing : Statistical Models, Test Plans and Data Analysis*, New York: Wiley.

Sarkadi, K. and Vincze, I. (1974), *Mathematical Methods of Statistical Quality Control*, New York: Academic Press.

Sibuya, M. and Suzuki, K. (2001), "Optimal Threshold for the $K$-out-of-$N$ Monitor with Dual Failure Modes," *Annals of the Institute of Statistical Mathematics*, 53, 189-202.

# 10 Appendix

*Proof of Proposition 1:* We begin with the fixed sample size problem. Recall that the power of the traditional demonstration plan (TDP) is the same as that for testing the null hypothesis $\mathrm{H}_0 : \theta = R_0$ versus $\mathrm{H}_a : \theta = R_a$ where the data are independent Bernoulli trials $X_1, ..., X_n$ with $\mathrm{P}(X_i = 1) = \theta$. Now suppose $\frac{\tau(R_a)}{R_a} = \frac{\tau(R_0)}{R_0}$ and there exists an X-test with level $\alpha$ and sample size $n$ using the X-transform $\tau$ that has greater power than the traditional plan with the same level $\alpha$ and sample size. The power of this X-test is the same as that for the test of the hypothesis $\mathrm{H}_0 : \theta = R_0$ versus $\mathrm{H}_a : \theta = R_a$ where the data are $X_1Y_1, ..., Y_nX_n$ where $Y_1, ...Y_n$ are independent Bernoulli random variables with $\mathrm{P}(Y_i = 1) = \frac{\tau(R_0)}{R_0} = \frac{\tau(R_a)}{R_a}$. This is due to the fact that the data in the X-test have the same distributions under both hypotheses as $X_1Y_1, ..., X_nY_n$. Now by the Neyman Pearson lemma, we know the most powerful test for $\mathrm{H}_0 : \theta = R_0$ versus $\mathrm{H}_a : \theta = R_a$ involves only $X_1, ..., X_n$, so the power using the X-test cannot be greater than the power using the traditional test. Further if $\frac{\tau(R_a)}{R_a} < \frac{\tau(R_0)}{R_0}$, it is easy to see that the power of the X-test will be even smaller.

For the fixed power problem, any X-transform for which the condition in Proposition 1 holds will lead to an X-test that requires a sample size at least as large as that of the traditional test plan. This is due to the fact that for the same sample size the power of the X-test cannot be larger than the power of the TDP, using the previous argument. Also, for the zero-failure plan, any X-transform for which the condition in Proposition 1 holds cannot result in a larger power than the TDP, since the sample size for the zero-failure X-test is smaller than the sample size for the traditional zero-failure plan. □

*Proof of Proposition 2:* Again we begin with the fixed sample size problem. The power of the X-test is the same as the power for testing the hypothesis $\mathrm{H}_0 : \theta = \tau(R_0)$ versus $\mathrm{H}_a : \theta = \tau(R_a)$ where the data are independent Bernoulli trials $X_1, ..., X_n$ with $\mathrm{P}(X_i = 1) = \theta$. Now suppose $\frac{1-\tau(R_a)}{1-R_a} = \frac{1-\tau(R_0)}{1-R_0}$ and there exists a TDP with level $\alpha$ and sample size $n$ using the X-transform $\tau$ that has greater power than the X-test with the same level $\alpha$ and sample size. In the TDP the power of the test is the same as the power for testing the hypothesis $\mathrm{H}_0 : \theta = \tau(R_0)$ versus $\mathrm{H}_a : \theta = \tau(R_a)$ where the data are $1 - Y_1(1 - X_1), ..., 1 - Y_n(1 - X_n)$ where $Y_1, ..., Y_n$ are independent Bernoulli trials with $\mathrm{P}(Y_i = 1) = \frac{1-R_0}{1-\tau(R_0)} = \frac{1-R_a}{1-\tau(R_a)}$. This is due to the fact that the data in the TDP have the same distributions under both hypotheses as $1 - Y_1(1 - X_1), ..., 1 - Y_n(1 - X_n)$. Now by the Neyman Pearson lemma we know the most powerful test for $\mathrm{H}_0 : \theta = \tau(R_0)$ versus $\mathrm{H}_a : \theta = \tau(R_a)$ involves only $X_1, ..., X_n$ so the power using the X-test cannot be less than the power of the traditional test. Further, if $\frac{1-\tau(R_a)}{1-R_a} < \frac{1-\tau(R_0)}{1-R_0}$ the power of the X-test will be even larger. From this it also follows that for the fixed power problem any X-transform for which the condition in Proposition 2 holds will never require a larger sample size than the traditional test plan since for the same sample size the power of the X-test cannot be smaller than the power of the TDP. For zero-failure plans, any X-transform for which the condition in Proposition 2 holds will not result in a smaller power than the TDP even though the sample size is decreased.

To see this suppose $\frac{1-\tau(R_a)}{1-R_a} \leq \frac{1-\tau(R_0)}{1-R_0}$ in which case the power of the X-test will be at least as large as

$$\left[1 - \frac{1-\tau_0}{1-R_0}(1-R_a)\right]^{\frac{\log(\alpha)}{\log(\tau_0)}}$$

which is at least as large as the power of the traditional plan since the above expression increases as $\tau_0$ decreases. □

*Proof of Proposition 3:* As a result of Proposition 1 it is sufficient to check that the condition given in Proposition 1 holds. Taking $x_1 = Q_F(1-R_a)$, $x_2 = Q_F(1-R_0)$ and $c = Q_F(1-\tau(R_0)) - Q_F(1-R_0)$ we can rewrite the condition in Proposition 1 as $\frac{1-F(x_1+c)}{1-F(x_1)} \leq \frac{1-F(x_2+c)}{1-F(x_2)}$. Now since $\int_{-\infty}^{t} h(x)dx = -\log(1-F(t))$, by taking the log, the condition becomes $\int_{x_1}^{x_1+c} h(x)dx \geq \int_{x_2}^{x_2+c} h(x)dx$ which is true provided $h(x)$ is non-increasing for $x \in (x_1, x_2+c) = (Q_F(1-R_a), Q_F(1-\tau(R_0)))$. □

*Proof of Proposition 4:* As a result of Proposition 2 it is sufficient to check that the condition given in Proposition 2 holds. Let $x_1 = Q_F(1-R_a)$, $x_2 = Q_F(1-R_0)$ and $c = Q_F(1-\tau(R_0)) - Q_F(1-R_0)$ and let $\tilde{F}$ denote the CDF of $-X$. The condition in Proposition 2 can be written as $\frac{F(x_1+c)}{F(x_1)} \leq \frac{F(x_2+c)}{F(x_2)}$ which is equivalent to

$$\frac{1-\tilde{F}(-x_1-c)}{1-\tilde{F}(-x_1)} \leq \frac{1-\tilde{F}(-x_2-c)}{1-\tilde{F}(-x_2)}.$$

Taking log, this inequality becomes

$$\int_{-x_1-c}^{-x_1} \tilde{h}(x)dx \leq \int_{-x_2-c}^{-x_2} \tilde{h}(x)dx$$

which is true provided $\tilde{h}(x)$ is non-increasing for $x \in (-x_2-c, -x_1) = (-Q_F(1-\tau(R_0)), -Q_F(1-R_a))$. □

*Proof of Proposition 5:* For a location X-transform with CDF $F$, (9) can be written as

$$\{1 - F[Q_F(1-R_a) + Q_F(1-\tau_0) - Q_F(1-R_0)]\}^{\frac{\log(\alpha)}{\log(\tau_0)}}$$

or

$$(1-F(x_1+c))^{\frac{\log(\alpha)}{\log(1-F(x_2+c))}}$$

where $x_1 = Q_F(1-R_a)$, $x_2 = Q_F(1-R_0)$ and $c = Q_F(1-\tau_0) - Q_F(1-R_0)$. Taking log, this becomes

$$\log(\alpha)\frac{\log[1-F(x_1+c)]}{\log[1-F(x_2+c)]}.$$

This will be increasing, constant, or decreasing in $\tau_0$ depending on its derivative with respect to $c$ which can be written as the product of a positive function and

$$\log(\alpha)\left[\frac{-f(x_1+c)}{1-F(x_1+c)}\log(1-F(x_2+c)) - \frac{-f(x_2+c)}{1-F(x_2+c)}\log(1-F(x_1+c))\right]$$

where $f$ is the density function for the CDF $F$. Thus (9) will be increasing, constant, or decreasing as $\tau_0$ decreases depending on whether

$$\frac{f(x_1+c)}{[1-F(x_1+c)]\log(1-F(x_1+c))} - \frac{f(x_2+c)}{[1-F(x_2+c)]\log(1-F(x_2+c))}$$

is positive, zero, or negative respectively. To check this, it is sufficient to check that the function

$$\frac{f(x)}{[1 - F(x)]\, |\log(1 - F(x))|} = \frac{h(x)}{H(x)}$$

is increasing, constant, or decreasing in $x$ for $x \in (x_1, x_2 + c) = (Q_F(1 - R_a), Q_F(1 - \tau_0))$. $\square$