Goodness-of-Fit for Sequentially Censored Lifetests

Lt. Colonel Andrew G. Glen, Ph.D.
Department of Mathematical Sciences
United States Military Academy
West Point, New York 10996
aa1275@usma.edu

Bobbie Leon Foote, Ph.D.
Department of Systems Engineering
United States Military Academy
West Point, New York 10996
fb9690@usma.edu

June 16, 2003

ABSTRACT

We propose a methodology for gaining statistical inference on censored samples, especially during the actual conduct of the lifetest experiment, in order to reduce cost and time on test while preserving reasonable levels of statistical power, and in at least one case, the methodology has increasing statistical power of a censored sample over that of a full sample. The outcome of the methodology will produce design efficiencies in lifetime testing. The method is distribution free for any fully specified continuous distribution under the null hypothesis, and produces p-values that are exact. Transforming ordered lifetest data into iid uniformly distributed data on (0,1), we use the T_n statistic, discussed in a companion paper (Glen and Foote 2003), to gain inference on mean life of systems with resulting power increases of up to 30% higher than that of the Anderson–Darling statistic. We investigate, with simulation, the power of the method as r (the number of failures currently observed) increases to n. We look specifically at null hypotheses from the exponential, normal, and gamma families of random variables. We introduce an automated tool that allows for immediate implementation of the new method using the probabilistic software package "A Probability Programming

Language" running in the Maple environment. We provide conclusions that will give insight on how to gain statistical inference with less time and materiel on test. We also show a counter-intuitive result where in certain cases, censored samples produce higher power than full samples. We investigate this counter-intuitive result more fully.

Keywords: Computational Algebra Systems, Exact Distributions, Conditional Order Statistics, Censored Lifetesting.

1 INTRODUCTION AND LITERATURE REVIEW

In lifetesting applications, tests are designed to gain an understanding of the probabilistic properties of a component or a system of components. Often, the costs of lifetests, in both time and money, constrain the design of the experiment, limiting the number of items placed on test and the length of the test. Many times, like in pharmaceutical drug testing, the length of the experiment cannot be estimated accurately in advance, and often one is faced with un-analysed, censored data in an ongoing experiment. For such cases we propose a methodology that gives exact statistical inference on censored samples. Consider an existing component, process, or drug with an all-parameters known lifetime reliability distribution $F(x)$. Should an improved component, process, or drug come along, both producers and consumers would like to verify that the new item is better than the existing item, most often by determining if its mean lifetime has improved (whether a decrease or an increase). In the lifetesting of the new component, it would be highly desirable to stop the test when enough evidence exists to support either claim. Such censoring, commonly called Type I (stop after time t) or Type II (stop after r items fail), can produce statistical inference, however, existing methods are not widely known, nor do they have remarkable statistical power. We propose a methodology that will specifically rely on Type II censoring in the design and conduct of the lifetest. If for example, one could afford a lifetest with $n = 5$ items to fail, a certain level of statistical power could be achieved if the test continued until completion of n failures. Consider, however, an example where $n = 25$ items are placed on test with $r = 5$ as the designated censoring value. Obviously the second test would conclude more quickly, as the expected time on test would be the mean failure time of $X_{(25:5)}$, the

fifth order statistic from a sample of 25 items, under the null hypothesis. Now consider a slightly different example, where $n = 25$ items are placed on test. Experimenters notice that after $r = 3$ failures, lifetimes seem to be substantially better than the original system. After $r = 6$ failures, they are convinced, at least anecdotally, that the new system is better. We propose a new methodology and a new test statistic that will allow for instantaneous assessment at every failure, with exact p-values, from an exact distribution of the test statistic. We rely on properties of conditional order statistic distributions to provide statistical inference for censored data that has acceptable statistical power. We also show that for the case of the Gamma distribution, given certain conditions, it is possible to achieve higher power with a censored sample than it is for a full sample, a counter-intuitive result that has warranted in depth investigation on our part. We use the test statistic T_n, presented in a companion paper (Glen and Foote 2003), which has significantly more power than the Anderson–Darling statistic, given changes in mean lifetime. The method we propose transforms censored data, via two probability integral transforms (PITs) and conditional order statistics, into an un-ordered, iid sample of uniformly distributed data on the open interval (0,1), which we abbreviate $U(0, 1)$. Furthermore, the test statistic T_n, designed as a test of uniformity, enjoys significantly higher power than the A–D statistic when finding differences in the mean of the distribution of the item in question, thus higher power is generally possible by combining the censored methodology with the use of the T_n statistic. The net effect of combining the new statistic with the new methodology is an very strong advantage in assessing censored data, to include the possibility of purposefully designing lifetests with higher values of n so that the test can be censored early at a reasonable value of r, saving time, money, and items that were destroyed during the test.

Rosenblatt (1952) presents theory that transforms joint conditional statistics to ordered, uniformly distributed statistics for the censored case (we will instead transform censored data to a complete un-ordered set of uniform data). David (1981) discusses the Markov nature of conditional order statistics, and equates the conditional order statistic with the truncated order statistic, a result that we will use as part of our method. O'Reilly and Stephens (1988) use a Rosenblatt transform, then invert that transformed data to test ordered uniform data (we will test un-ordered uniform data). Hegazy and Green (1975) present work on goodness-

of-fit using expected values of order statistics with approximations used for critical values. Balakrishnan, Ng, and Kannan (2002) present a test for exponentiality that is based on progressively censored data, which uses a T statistic, however this statistic and this method is unrelated to the T_n statistic and the sequentially censored data analysis that we use. Michael and Schucany (1979) also present a transformation that takes censored data and transforms it into ordered uniform data. Since Michael as well as Stephens (1974) also point out that the A–D statistic is generally more powerful than the other well-known goodness-of-fit statistics in the case when the mean has shifted, we will rely on T_n, which has even higher power in detecting shifts in the mean than A–D, as shown our earlier, companion paper (Glen and Foote, 2003).

2 TRANSFORMING THE CENSORED DATA INTO IID $U(0,1)$

Let the lifetime of an existing system (also that of the null hypothesis) be distributed by the all-parameters known continuous rv X with CDF $F(x)$. Let n items be on lifetest and let the Type II censoring value be r. Recall that in a lifetest, failure data arrives in ordered fashion. The ordered lifetime data $x_{(i)}$ have CDFs from their appropriate order statistic $F_{X_{(n:i)}}(x_{(n:i)}),\ i = 1, 2, \ldots, r$, (note $X_{(n:i)}$ is abbreviated $X_{(i)}$). Now consider the conditional order statistics of the lifetest, $X_{(1)}, X_{(2)}|X_{(1)}, \ldots, X_{(r)}|X_{(r-1)}$. Theorem 2.7 from David (1981, pg. 20) explains the Markov nature of these conditional order statistics. Thus for our purposes the CDF of the ith order statistic, given the $(i-1)^{\text{th}}$ data point, $F(x_{(i)}|x_{(i-1)})$, is that of the rv $X_{(n-i+1:1)}$ with support $x_{(i-1)} < x_{(i)} < 1$. David shows this is the first order statistic from a sample size $n-(i-1)$ from the parent distribution of X truncated on the left at $x_{(i-1)}$. In other words, the distribution is independent of $x_{(1)}, x_{(2)}, \ldots, x_{(i-3)}$, and $x_{(i-2)}$, and is therefore memoryless in this regard. Since each of the conditional distributions can be computed, conducting separate PITs on each data value, $F_{X_{(i)}|X_{(i-1)}}(x_{(i)}),\ i = 2, 3, \ldots, r$ will give a sample of r iid $U(0,1)$ random variables (see Rosenblatt 1952, pg. 470) to which a uniformity test can be applied. As mentioned earlier, we use T_n, as it is better at finding changes in μ_X than A–D in many cases (Glen and Foote, 2003). The statistic T_n has the

distribution of the convolution of n iid $U(0,1)$ random variables. Therefore, the test statistic we will use is as follows:

$$T_r = \sum_{i=1}^{r} F_{X_{(i)}|X_{(i-1)}}(x_{(i)}),$$

where $F_{X_{(1)}|X_{(0)}}$ is defined to be $F_{X_{(1)}}$, and r is the size of the censored sample.

3 IMPLEMENTATION USING APPL

The theory of the statistic is straightforward, however the implementation is made practicable only with automated probabilistic software. We implement the new method and new statistic in APPL (Glen, et. al. 2001) for a number of reasons. The software allows us to use exact distributions of the original data, the distributions of the conditional order statistics, and the distribution of the T_n statistic so that exact p-values are attainable. Additionally, the author has already calculated the PDFs of the sum of n $U(0,1)$ random variables from $n = 1$ to $n = 50$, the last PDF requiring 91 pages of ASCII text to enumerate. APPL reads these PDFs exactly and can thus compute the exact p-values. APPL allows for the use of any continuous distribution (well-known distributions as well as ad hoc) to specify the null hypothesis and conducts the necessary PITs for these distributions. We will demonstrate power of the censored and full samples using T_n and A–D statistics with data from the Normal, Exponential, and Gamma prior distributions, however we are not limited to just these distributions.

The methodology can be confusing to those not used to using conditional order statistics, thus we present more clearly the algorithm for computing the test statistic.

- Specify the null distribution of the existing (old) system, $F(x)$.
- During the lifetest experiment, note n and create the vector of r observed occurrences.
- Calculate $z_{(i)} = F(x_{(i)}),\ i = 1, 2, \ldots, r$, which is ordered uniform (not iid).
- Calculate the unordered, iid $U(0,1)$ (under the null hypothesis) $u_i = F_{Z_{(i)}|Z_{(i-1)}}(z_{(i)}),\ i = 1, 2, \ldots, r$. Note: we perform the PIT with F(x) and then conduct the conditional order statistics PIT using the uniform conditional order statistic distributions. These two

methods have been shown to be equivalent (Glen, et. al., 2001), but this method is preferred as the conditional order statistics of the uniform distributions are much more tractable than conditional order statistics using the parent distribution F. Also note, we find the conditional order statistic using the truncation of the parent distribution method outlined by David (1981).

- Sum the u_i values to get the T_r statistic.
- Calculate the p-value with the appropriate tail of the T_r distribution.

The APPL code that enacts this algorithm to calculate the statistic is as follows:

```
# take the r censored values in 'data' and PIT them into the list 'Zdata'
for i from 1 to r do
  Zdata := [op(Zdata), CDF(Nulldist, data[i])];
od;
# sum the independent uniforms to for the statistic 't_stat' starting with the first failure ...
t_stat:=CDF(OrderStat(U(0, 1), n, 1), Zdata[1]);
# ... then adding up the subsequent failures until r is reached.
if (r > 1) then
  for i from 2 to r do
    t_stat := t_stat + CDF(OrderStat(Truncate(U(0,1), evalf(Zdata[i-1]), 1),
      n - (i - 1), 1), Zdata[i]);
  od;
fi;
Tr_distn := cat('T',r);
# now return the statistic, the lower tail pvalue and the upper tail pvalue
# using the APPL command 'CDF'
RETURN(t_stat, CDF(Tr_distn, t_stat), 1 - CDF(Tr_distn, t_stat));
```

This APPL code is implemented in a new APPL procedure called `CensoredT` and its use is illustrated in the example that follows. Assume there exists a medical treatment that has an established time-to-healing record that is modeled by the Gamma(2.1, 4.41) distribution,

where time is measured in years. A new treatment is developed and experimenters hope to show an improvement (decrease) in healing time. The new treatment is administered to $n = 25$ patients, and it is noted that the first five healing times are 0.40, 0.54, 0.66, 0.75, 0.84 years. Completion of the full experiment, under the null hypothesis, has an expected time of $E(X_{(25)}) \approx 4.52$ years, the expected healing time of the slowest patient to heal. However, the fifth patient's expected healing time, under the null hypothesis, is $E(X_{(5)}) \approx 1.21$ years. Since the observed time of the fifth patient's healing was only 0.84 years, it would useful to know if there is enough statistical evidence to stop the experiment, concluding that the new treatment is better. The following APPL code will analyse this Type-II censored experiment:

```
> Old_Treatment := GammaRV(2.1, 4.41);
> n := 25;
> data := [0.40, 0.54, 0.66, 0.75, 0.84];
> CensoredT(Old_Treatment, data, n);
```

The procedure output is the test statistic, the lower tail p-value and the upper tail p-value. In this case those values are 1.309743, 0.031999, 0.968001. Since we are interested in the lower tail, we have a p-value of 0.031999, significant evidence that the new treatment is better and we can consider terminating the experiment.

4 POWER SIMULATION RESULTS

In this section, we discuss the results of various power simulations to see the effect of increasing r on the power of the test. We will use the T_n statistic and benchmark it against the A–D statistic. In the case of the Exponential and Normal prior distributions, power tests confirmed what was expected: as r increased, the power of the test increased, but never exceeded the power of letting all n components fail. In the case of the Gamma prior distribution, however, a non-intuitive result was observed. Power initially increased as r increased, but then started to decrease after reaching a 'maximum' power. Even more unexpectedly, in certain cases of parameter values, the maximum power of the mid-values of r was actually higher that the power of the full sample. We have investigated some of this unexpected phenomena and report on it below, as the implications of more power with lower

r is very significant.

To implement this simulation, we wanted to set up the experiment so that, where possible, the underlying data had changing μ, but constant σ^2. We fixed σ^2 so that we could see if we could spot a change in μ by itself. This ability to detect a change is μ is helpful to lifetesters who have a new component that they would like to show superior to an existing component with a well defined distribution and well established μ. In the case of Exponentially distributed data, we could not fix σ^2 as $\sigma^2 = \mu^2$. We are able to fix σ^2 for the Normal and Gamma distributions. Table 1 shows the parameter values, as well as μ, and σ^2 for the Exponential, Normal, and Gamma distributions that were used in the power experiment.

Table 1: Distribution families, parameters, mean and variances for Monte Carlo Simulation

Normal Distribution, $H_0 : \mu = 1$, fixed $\sigma = 1$										
μ_a	-1	-0.8	-0.6	-0.4	-0.2	0.2	0.4	0.6	0.8	1
Exponential Distribution, $H_0 : \lambda = \mu = 1$										
$\lambda_a = \frac{1}{\mu_a}$	0.4	.6	0.7	0.8	0.9	1.25	1.5	1.9	2.3	2.7
Gamma Distribution, $H_0 : \alpha = \mu = 2.1$, $\beta = 4.41$ fixed $\sigma = 1$										
$\alpha_a = \mu_a$	1.1	1.3	1.5	1.7	1.9	2.3	2.5	2.7	2.9	3.1
β_a	1.21	1.69	2.25	2.89	3.61	5.29	6.25	7.29	8.41	9.61

As we see in the Normal and Exponential cases (Figures 1 and 2), higher r values produced higher power. Also, as can be seen in Figures 2 and 3, the T_r statistic produced higher power than the A–D statistic except for the extremely high values of μ (though not shown in Figure 1, the same result was observed for the Normal distribution). This switch is interesting since in the full sample experiments from the companion paper, The T_r appeared to always be higher in power than the A–D statistic.

A very counter-intuitive phenomena occurs with the Gamma distribution. Highest power for censored samples appears to come approximately $r = 10$ and then decrease as r approaches n. This result happened for the T_r and the A–D statistics. An enlargement of Figure 3 is shown in Figure 4 that further shows that the power increases then decreases. Figure 4 clearly shows that power starts out moderately at $r = 5$, then seems to achieve a maximum at $r = 10$ (for both statistics) then clearly decreases by the time $r = 20$ and

$r = 25$. (Note the conditional order statistic approach at $r = n$ appears to be a different, less powerful statistic than the full sample for Gamma prior.) Most striking was that, for some values of μ_a lower than μ_0 we have achieved higher power for the censored, $r = 10$, case than we did for the full sample. As this is very counter-intuitive, we experimented in detail the case where the Gamma parameter $\alpha = 1.7$ and calculated the power for each value of $r = 1, 2, \ldots, 25$. The results of this in depth simulation are shown in Figure 5. Here we clearly see both phenomena occur: 1) power increases until approximate $r = 9$, then it decreases, and 2) for values for $r = 6, 7, 8, 9,$ and 10 power for the censored sample is at least has high or higher than power for the full sample. A note on the simulations: as these Gamma prior results were so counter-intuitive that our colleagues have had difficulty believing that a censored sample could possibly produce higher power than a full sample, we have re-designed and re-run this experiment a number of times over the last year, achieving similar results each time. For a copy of the simulation code, readers may contact the first author.

5 APPLICATIONS AND IMPLICATIONS

This methodology has potential for significant advances in reliability engineering lifetesting, pharmaceutical drug tests, or any sort of experiment where data comes naturally in ordered form. The sequential testing ability allows for a test to be terminated early, hence ending a dangerous experiment or giving early vindication allowing an effective therapy to go to market earlier. In particular, if a new therapy or component is more effective than the old, early failures may be remarkably small or large. This will result in acceptance and termination without running until all cases have failed. The test can then be used to accept the new component or medical treatment. Similarly, a few early failures can render a judgment and the remaining patients can be switched to potentially better therapies. Other implications of this research is as follows:

- Good statistical power for censored samples is possible for a wide ranges of experiments.
- Experiments can be designed for high n values, knowing that they will stop at a predetermined, relatively small r value.

- Experiments can be tracked real-time to see a pattern of p-values that indicates enough inference has been gained.
- With a Gamma prior distribution, higher power is achieved in censored samples than with full samples in some cases.

6 CONCLUSIONS

A new goodness of fit methodology has been developed and tested. Significant increases in power on the order of 30% have been found compared to the standard Anderson - Darling statistic. Also, relatively high power is achieved using the T_n statistic on censored samples, allowing for lifetests to be terminated early. Finally, in at lease one special case, that of a Gamma prior, a phenomena has been found, that at approximately $r = 0.4\,n$, power is greater that with a full sample.

7 TOPICS OF FURTHER RESEARCH

The cause of the phenomena revealed by validation testing of the slightly higher power in one special case needs to be further investigated. A possible basis for the explanation lies in the variance of successive, truncated order statistics, when data that originates from the alternate hypothesis is passed through the PIT of the Gamma distribution based on the null hypothesis. Somehow, the transformation of the data has a different characteristic than when is it passed thru a PIT based on a null hypothesis with, say, an Exponential prior. Also, further research is needed to investigate how high to set n and r in experimental design, in order to gain possible advantages in lower time on test, lower cost, and fewer failed items as a result of the experiment. For example, if a budget can afford 25 items failing, perhaps it would be more effective to put 50 items on test, knowing ahead of time that the desired increase on μ should be evident by about the $r = 10^{\text{th}}$ failure. Clearly a time savings and component savings is evident here. Finally, one of our goals was to find the exact power functions instead of using simulation of power. Due to the complexity of sending data from one distribution thru the PIT of another, the resulting transformations were so complicated

that we could only find the exact power function for the Exponential prior with $r = 2$.

REFERENCES

Balakrishnan, N., Ng, H. K. T., and Kannan, N. (2002), "A Test for Exponentiality," published in *Goodness-of-fit Tests and Model Validity*, editors C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, and M. Mesbauh, Birkhaeuser.

David, H. A. (1981), *Order Statistics*, Second edition, John Wiley and Sons.

Glen, A., Leemis, L., and Barr, D. (2001) "Order Statistics in Goodness of Fit Testing," *IEEE Transactions on Reliability*, **50**, Number 2, pp. 209–213.

Glen, A., Leemis, L., and Evans, D. (2001), "APPL: A Probability Programming Language," *The American Statistician*, **55**, Number 2, pp. 156–166.

Glen, A., Foote, B. (2003), "Test for Uniformity based on Convolutions of the Uniform Distribution," *Technical Report, USMA* , West Point, NY.

Hegazy, Y., Green, J. (1975), "Some New Gooness-of-Fit Tests Using Order Statistics," *Applied Statistics*, **24**, Issue 3, pp. 299–308.

Maple Version 7 (2001), Waterloo Maple Inc., Waterloo, Canada. *Order Statistics*, Second edition, John Wiley and Sons.

Michael, J. R. and Schucany, W. R. (1979), "A New Approach to Testing Goodness of Fit for Censored Samples," *Technometrics*, **21**, Number 4.

O'Reilly, F. J. and Stephens, M. A. (1988), "Transforming Censored Samples for Testing Fit," *Technometrics*, **30**, Number 1.

Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, **23**, Issue 3.

Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, **69**, Issue 347.

Figures

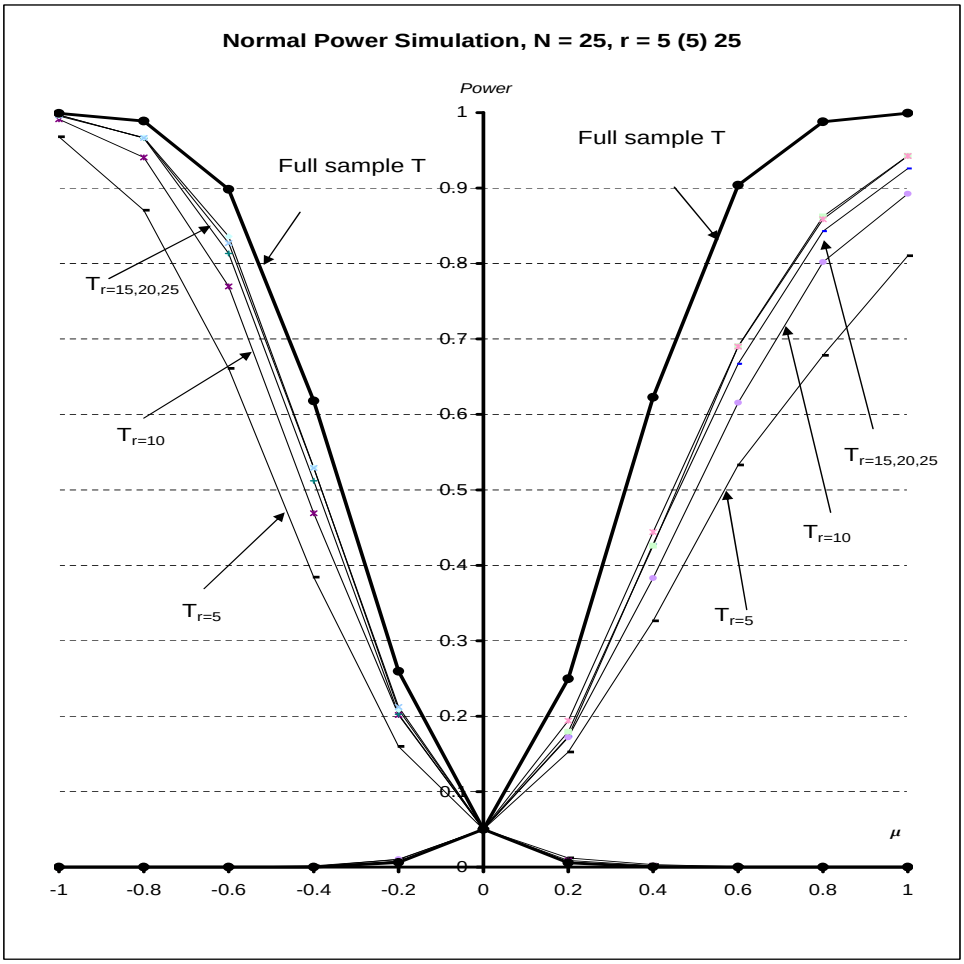

Figure 1: Results of Monte Carlo power simulation with underlying normally distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, $\mu = 0$. Only the T_n results are shown. Notice how well behaved the power functions are, in that higher r produced higher power.

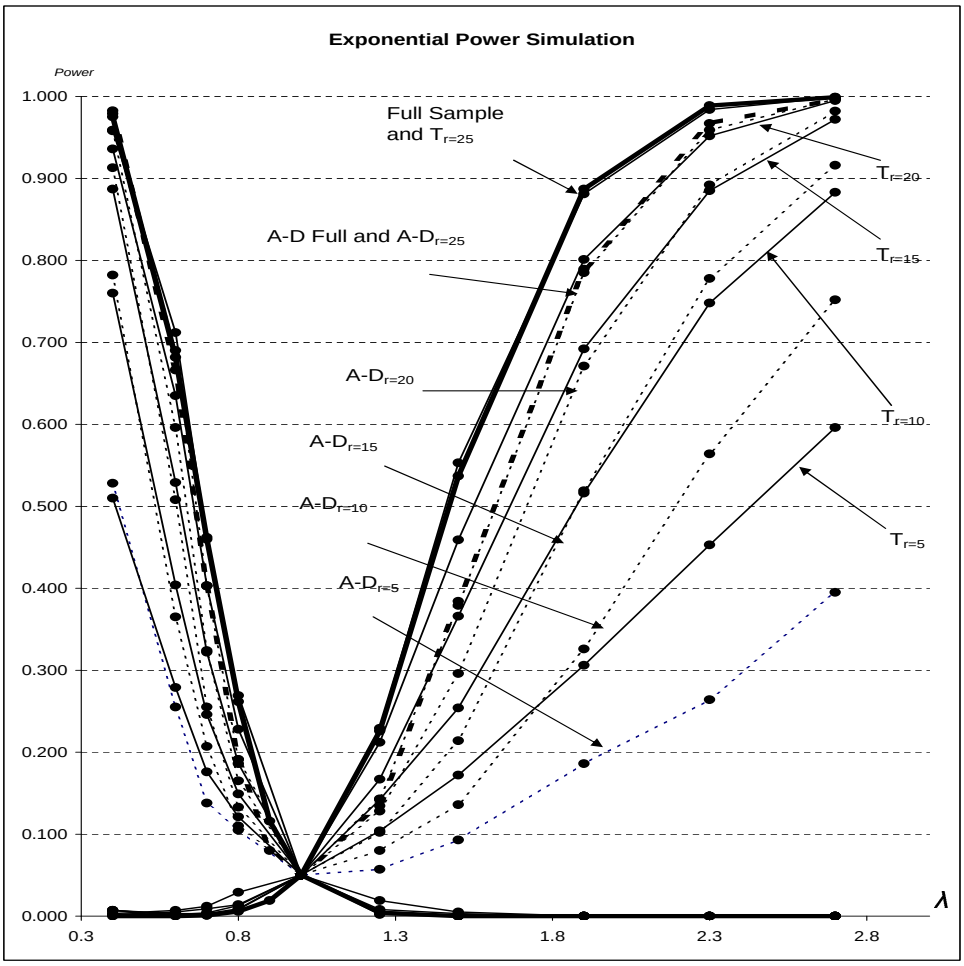

Figure 2: Results of Monte Carlo power simulation with underlying exponential distributed data with type I error $\alpha = 0.05$. Under H_0, the exponential distribution has parameter $\lambda = \frac{1}{\mu} = 1$. Thus, the upper tail test applies to the lower λ_a values. Notice that, like the Normal distribution, higher power is achieved for higher r values. Also notice how T_r achieves higher power than $A - D$, except for low values of λ_a (high values of μ_a).

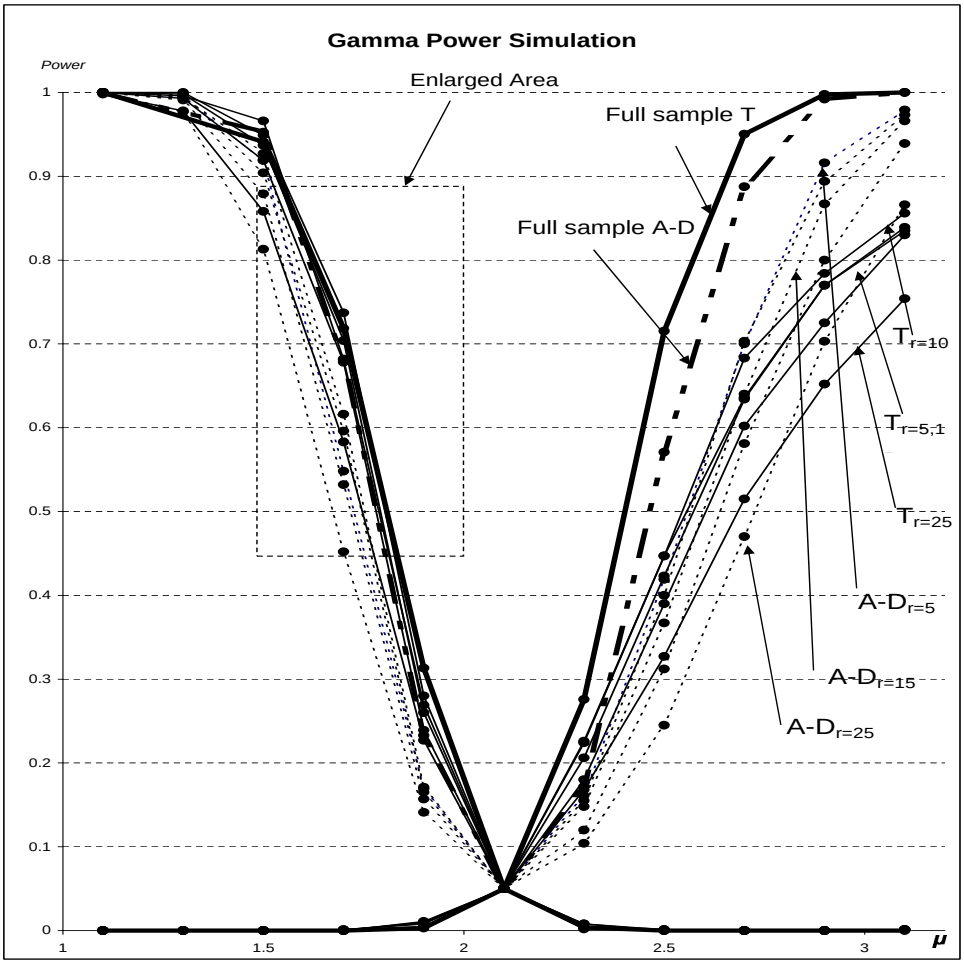

Figure 3: Results of Monte Carlo power simulation with underlying Gamma distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, the Gamma distribution has parameters $\alpha = \mu = 2.1$ and $\beta = 4.41$. Here the counter-intuitive result of higher power comes from $r = 10$ and then decreases as $r > 10$ for both the T_r and the $A - D$ test statistics. This phenomena is evident in the enlarged area shown in figure 4.

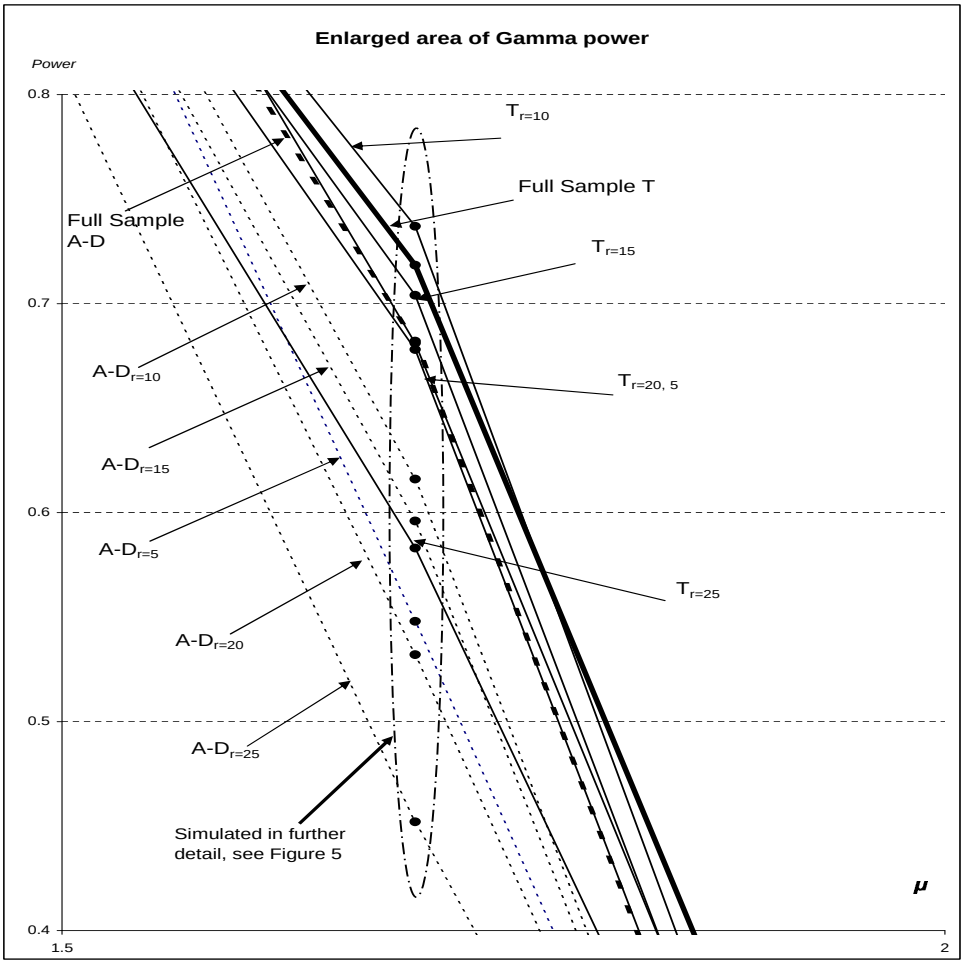

Figure 4: This enlarged area shows clearly the case that $T_{r=10}$ has higher power than even the full sample $T_{n=25}$. Thus we see the counter-intuitive result that under certain conditions an experiment can actually achieve higher power with a censored sample than with a full sample. Further investigation of this phenomenum at a higher resolution of r is found in Figure 5.

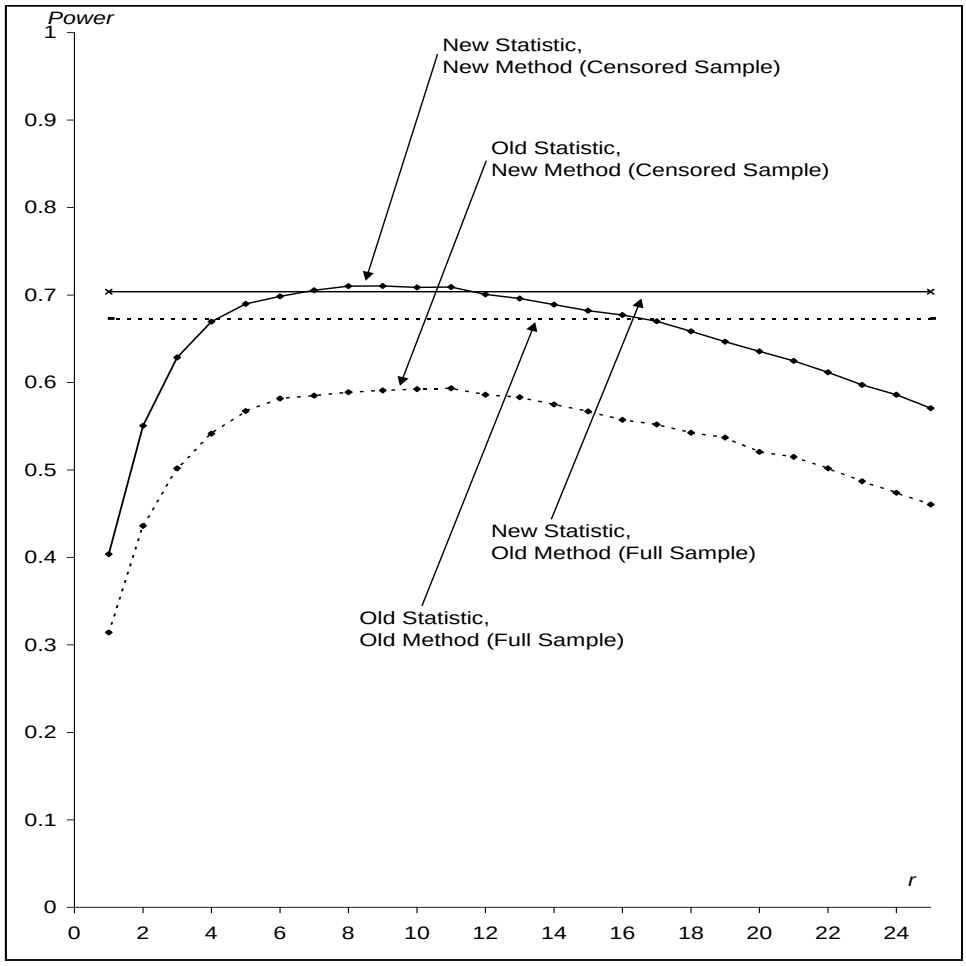

Figure 5: In an in depth experiment suggested from Figure 4, here is a plot of r versus power for each value of $r = 1$ (1) 25 for $\mu_a < \mu_0$. Note the two phenomena that 1) power increases on r then decreases for both statistics and 2) the special cases at $r = 6, 7, 8, 9,$ and 10 where higher power is achieved in a censored sample than with a full sample.