Statistical Model Selection for TID Hardness Assurance

R. Ladbury, J. L. Gorelick, S. S. McClure

Abstract— We investigate model dependence of bounding estimates of TID degradation as a function of sample size and statistical model and develop a method for selecting the model with greatest predictive power.

Index Terms—radiation effects, reliability estimation, quality assurance

I. Introduction

Strategies that allow Radiation Lot Acceptance Testing (RLAT) to be waived based on a part's historical performance can be among the most controversial aspects of Radiation Hardness Assurance (RHA) methodologies for Total Ionizing Dose (TID) degradation. In part, this is because historical data are usually less representative of flight-lot performance than RLAT data. In some cases (e.g. process changes, multiple foundries, bad lots), historical data may even be misleading. While risks associated with bad lots of parts and process changes can be estimated using historical data,[1] only RLAT offers high assurance of detecting such threats. However, even when historical data are representative of flight lots, strategies to compensate for the use of less representative data can themselves lead to inconsistent results.

A common strategy when using historical data for qualification is to increase the required radiation design margin (RDM=failure dose/mission dose) above the level usually required when qualification is based on RLAT. For instance, if RLAT requires RDM≥2, qualification using historical data would require a larger RDM>α>>2.

Methods for determining the minimum allowable RDM, α, split into two categories. Design Margin Breakpoint (DMBP)[2] methods base α on engineering judgment. While the choice of α is subjective, these methods are the only option for small data sets that may not fully reflect part-to-part and lot-to-lot variation. For larger datasets, statistical methods (e.g. the Parts Characterization Criterion (PCC)[3] and Aggregate[1] methods) base α on a fit of historical failure doses to an assumed form—usually normal or lognormal. Usually, α is chosen so that we have substantial confidence (for example, 90%) that a given percentage (at least 99%, say) of parts will pass at the mission dose. (In other words we want the mission dose to coincide with the dose at which 99% of parts will pass with 90% confidence—$D_{99/90}$.). Usually, DMBP RDMs must bound those from statistical methods, since the statistically determined α's apply to a specific part, while α for DMBP methods must apply for a broad range of different parts.

However, RHA methods can differ on the appropriate value for α, on RDM definition, on historical data quality and quantity requirements, or even on the definition of failure. These differences can lead to controversy for large flight projects with multiple partners, especially when each partner has a significant investment in product heritage. The situation is exacerbated by a lack of systematic studies of how the above differences affect the efficacy of the different methodologies, of what margins are adequate or of how to optimize methods for determining such margins.

In this work, we examine how dataset size and statistical model affect bounding estimates for various RHA methods and develop techniques for comparing the performance of different methodologies. We do so by examining part-to-part and lot-to-lot variability in common linear bipolar microcircuits for which we have long time series of RLAT data. We will see that while lot-to-lot variation can be bounded by moderate RDM ($3<\alpha<4$); some parts exhibit large lot-to-lot variability and seem to require α significantly larger than those typical of many DMBP methodologies. However, additional analysis reveals that while in some cases this high variability results from pathological behavior in the part's failure distribution, in other cases, the part's failure behavior is simply more complicated than can be reflected by the simple statistical models used by most RHA regimes. In the latter case, a more sophisticated statistical model can yield a much lower α.

II. Methodology

We focus on a few part types that illustrate the behaviors discussed above. Our time series data for these parts come from the Boeing Space and Intelligent Systems (BSIS) radiation database. Data were gathered during normal RLAT. Consistent test methods per Mil-Std. 883 Method 1019 allowed analysis across lots, and consistent procurement

Manuscript received July 20, 2009. The authors thank NASA's Electronic Parts and Packaging (NEPP) program and the Landsat Data Continuity Mission for support of this research.

R. Ladbury is with NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA (phone: 301-286-1030; fax: 301-286-4699; e-mail: Raymond.L.Ladbury.1@gsfc.nasa.gov).

J. L. Gorelick is an indeptndent consultant in Los Angeles, CA USA (e-mail: j.l.gorelick@att.net).

S. S. McClure is with NASA Jet Propulsion Laboratories, Pasadena, CA 91109 (telephone: 818 354 0482, email: Steven.S.Mcclure@jpl.nasa.gov)

specifications minimized extraneous sources of variability (e.g. different fabrication facilities, etc.). Table I lists the parts used in this study. The RH1014 is an example of a part with well behaved lot-to-lot variability. The Analog Devices AD590 temperature transducer (packaged in a flat pack) and OP484 quad op amp show large lot-to-lot variation. Finally, the National Semiconductor LM111 voltage comparator and Analog Devices OPO7 op amp show large lot-to-lot variation, but with fewer pathologies than the OP484 and AD590. All parts were tested at standard dose rates except the AD590, which was tested at 0.01 rads(Si)/s. Dose rates were consistent over all tests. For the RH1014, parametric shifts are reported for device 4 in the quad package, while for the OP484, parametric shifts reported are for device 1 in the quad package.

TABLE I: PARTS USED IN THE STUDY

Manufacturer	Part	# of Lots	Guarantee krad(Si)	Dose Levels krad(Si)	Sensitive Parameter
Linear Technologies	RH1014	38	100	60,100,200	I_B (# 4)
Analog Devices	OP484	9	100	100, 300	I_B
Analog Devices	AD590	10	N/A	5, 10, 15, 20 (low doserate)	Ambient Error
National Semi.	LM111	9	100	30, 60	I_B
Analog Devices	OP07	11	100	100, 300	I_B (#1)

We define α as the product of two factors, RDM_{RLAT} and RDM_{HIST}. The former factor is the RDM that we would require if we had RLAT for the flight lot (we take $RDM_{RLAT}=2$ for concreteness). The latter factor accounts for the lot-to-lot variation in the historical data. Thus, α is the total martin to be applied to historical data to avoid compromising reliability compared to the normal RDM_{RLAT} applied to RLAT data.

We next look at simple ways of defining RDM_{HIST}. In essence, we ask what value of RDM_{HIST} we would need to bound degradation for our softest lot if we based our failure estimate on the mean failure level of the hardest lot. This leads to widely ranging estimates of α from one part type to another. We then explore the causes underlying the variability for different part types and whether more sophisticated statistical analyses can yield more consistent results.

III. LOT-TO-LOT VARIABILITY AND RDM

We can measure lot-to-lot variability in TID degradation in a number of ways. Perhaps the simplest is to look at the ratio of the TID induced shift of a parameter for one lot to that for another. This does not even require that we define a criterion for parametric failure. If we do define such a criterion, we can look at the dose at which each part in a lot fulfills this failure criterion and define the mean failure level (dose) for the lot. The ratio of the mean failure levels then provides another measure of lot-to-lot variation. In Table II, we use such ratios to estimate RDM_{HIST} for each of our parts. Column 2 gives the ratio of the mean parametric degradation for the softest lot (SL) to that for the hardest lot. Column 3 lists the criteria we used to define parametric failure, and column 4 gives the ratio of the mean doses where these failure criteria were met (this time for the hardest lot to the softest lot). In a similar manner, one can use the ratios of other statistical quantities—e.g. the first (lowest-dose) failure for a lot, $D_{99/90}$ for a lot, etc. However, these metrics yield similar trends.

TABLE II: LOT-TO-LOT VARIATION

Part #	Mean Parametric Δ Ratio: SL to HL	Failure Criterion	Mean Failure Level Ratio: HL to SL
RH1014	~1.9 (@100 krad(Si))	ΔIbias>50 nA	1.6
OP484	~9 (@100 krad(Si))	ΔIin>2 μA	9.3
AD590	~10 (@15 krad(Si))	Amb. Err> 5°C	10
LM111	~3.4 (@60 krad(Si))	ΔIb_{in+}>300 nA	3.4
OP07	~3.4 (@300 krad(Si))	ΔIin>40 nA	4.2

Multiplying the values in columns 2 or 4 by $RDM_{RLAT}=2$, this suggests that $3<\alpha<4$ would be adequate to bound degradation for the RH1014. The OP07 and LM111 would require α~8, and the AD590 and OP484 would require α~20.

Figure 1 shows that basing our mean failure level on 2 or 3 lots of data can reduce the spread in our estimates for α for the OP07, LM111 and even the AD590. However, it does not help appreciably for the RH1014 or the OP484. This raises the question of what accounts for the widely varying values we obtain for α with our simple statistical analysis, and whether a more sophisticated analysis could produce results that were more consistent from part type to part type. To examine this question we began by carrying out an exploratory data analysis (EDA) of the lot-to-lot variation for the different parts in our study.

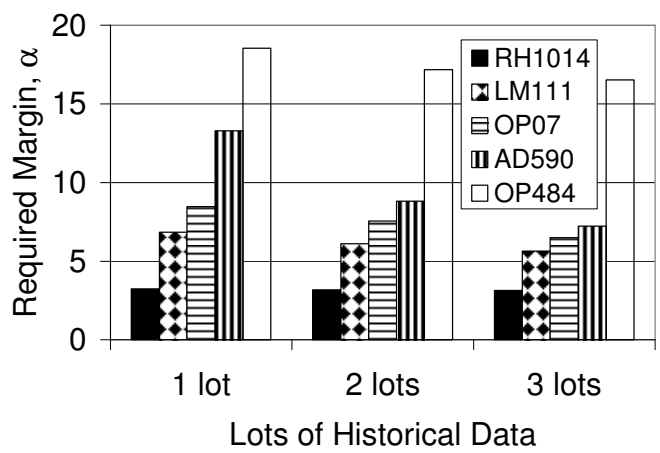

Fig.1 Required margin, α decreases as more lots of historical data are added unless the data are already sufficient to characterize lot-to-lot variation (e.g. RH1014) or lot-to-lot variation is pathological (e.g. OP484).

IV. EXPLORATORY DATA ANALYSIS

Our exploratory data analysis sought to elucidate the factors underlying the large spread in our estimated α for the parts in our study. We looked in the data time series for correlations, time dependence or other trends, as well as failure distribution pathologies (e.g. bimodality or thick tails). Except for the RH1014, each part poses challenges that lead to systematic errors if ignored in RHA analyses. (See figures 2-6.)

Failure levels for the RH1014 are unimodal with no

evidence of bad lots or outliers. The distributions of both mean lot failure levels and standard deviations give a good fit to a lognormal distribution. This implies that this part would be amenable to most statistical analyses and that qualification using historical data would pose minimal risks.

The lots in the LM111 sample show bimodality from lot to lot, though not from part-to-part in a lot.[5] Attempting to fit the data to a unimodal distribution exaggerates the distribution width, and this factor is largely responsible for the high values of α for this part. Moreover, if we lump data across lots to a single distribution, we will get a poor fit, since the bimodality manifests across lots rather than within a lot. The OP07 shows a linear trend between mean lot failure level μ and standard deviation σ. Since fitting all data to a single distribution obscures this correlation, this also will overestimate α. The role of systematic errors due to the simplicity of the statistical models in producing large estimates of lot-to-lot variation suggests that more complicated models could yield better bounds on α.

In contrast, lot-to-lot variability of the OP484 and AD590 is clearly pathological. As noted in reference [1], the OP484 shows bimodality in its failure distribution not just from lot to lot, but from part to part within a lot. Moreover, in 2 of 9 lots Iin saturated below the lowest test level of 100 krad(Si), while the other 7 lots do not show saturation even at 300 krad(Si). This makes it difficult even to define failure criteria that are relevant across all lots. The AD590 also exhibits large lot-to-lot variability in its degradation, although the data appear thick-tailed rather than bimodal.

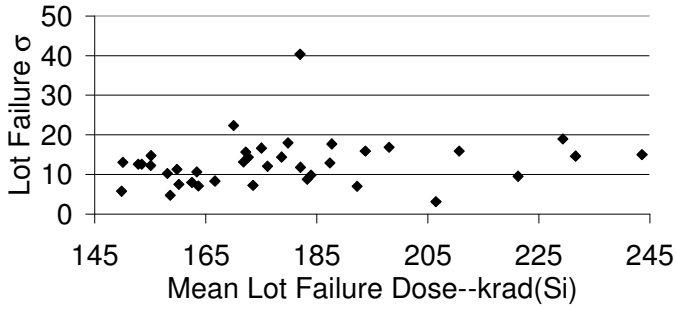

Fig. 2 Means and standard deviations (σ) of RH1014 lot failure doses are well behaved and uncorrelated.

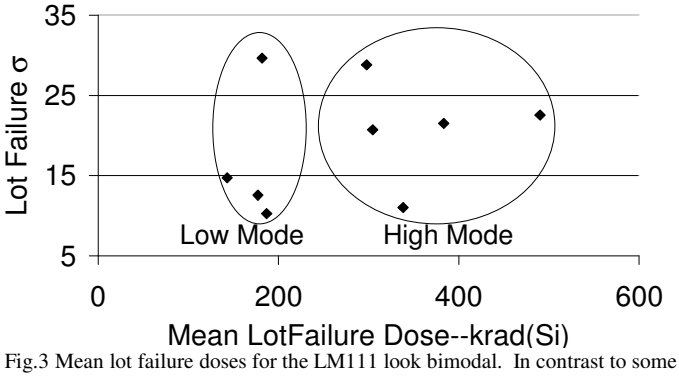

Fig.3 Mean lot failure doses for the LM111 look bimodal. In contrast to some studies[4], failures in each lot were unimodal.

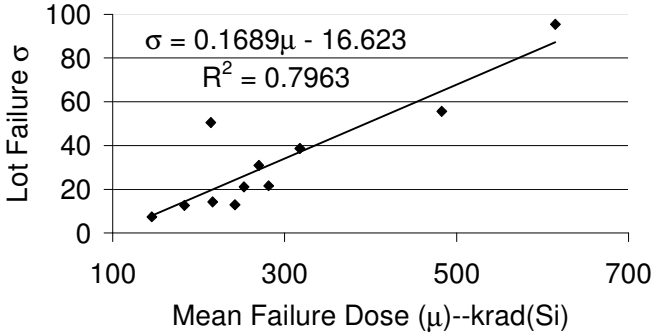

Fig.4 Lot mean failure dose μ and standard deviation σ are linearly correlated for the OP07.

To take into account the trends and features revealed by our EDA, we need a more complicated statistical model—one that can reflect lot-to-lot bimodality of the LM111 and the μ-σ correlation seen for the OP07. This raises the question of how we determine whether we have sufficient data to constrain such a model and how we determine whether the performance of this model is sufficiently improved to merit the added complexity. However, since we will be comparing models with different levels of complexity (e.g. number of parameters), our performance metric will have to go beyond simple "goodness of fit" criteria such as Least-Squares, $\chi 2$ or likelihood. We next consider 3 different models for our data and how to select the model that best explains our lot-to-lot variability and maximizes predictive power.

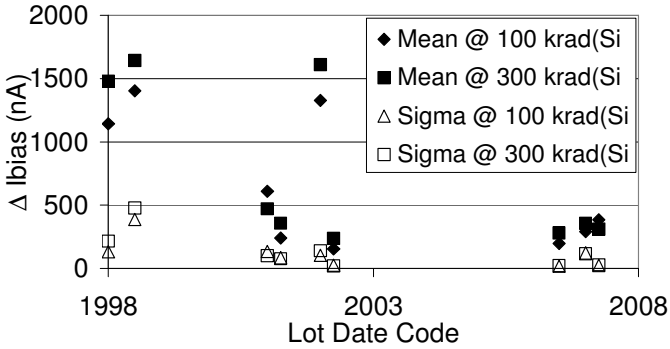

Fig. 5 Highly variable parametric degradation of OP484s prohibits even defining a failure criterion that is meaningful over all lots.

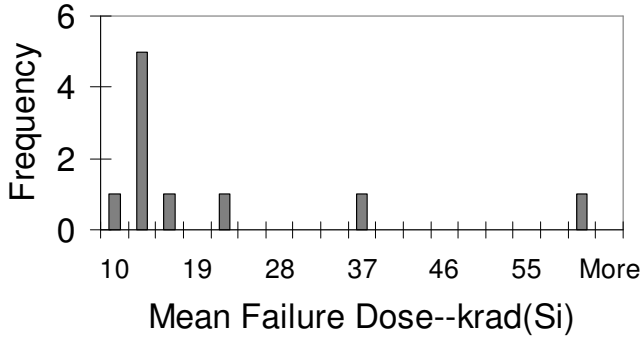

Fig.6 AD590 mean lot failures occur from 8.8-59 krad(Si). The distribution is skewed right, so a $D_{99/90}$ to the three best lots bounds the WC lot.

V. Statistical Methods

We consider 3 different models for our data:

1) Failures for each lot are distributed lognormally, but mean and standard deviation are unrelated from lot to lot.[5] This means historical data cannot predict future lot behavior and so each lot requires RLAT. This model requires 2n parameters to describe n lots.
2) Failures across all lots follow a single, unimodal lognormal distribution. This is equivalent to the PCC[3] or Aggregate[1] method. Two parameters describe the data, regardless of the number of lots n.
3) Each lot's failures are distributed lognormally as in model 1). The resulting fit parameters in turn follow well behaved (lognormal) distributions. Four parameters (2 each for the lognormal mean and standard-deviation distributions) describe the data for all n lots. Model 3 is flexible enough that we can impose correlations found in EDA as constraints.

To find the vector of parameters ($\mathbf{\Theta}_{BF}$) for each model that best fits the failure data **X,** we maximize the likelihood

$$\mathrm{L} = \prod_{i=1}^{n} \mathrm{P}_{\mathrm{Fi}}(\mathbf{\Theta}, \mathbf{X}) \tag{1}$$

for the failure model with respect to the parameter vector $\mathbf{\Theta}$ for given failure doses $\mathbf{X}$. Since the likelihood for a given parameter vector $\mathbf{\Theta}$ is just the product of the probabilities of our data given those parameters, the vector that maximizes the likelihood can be thought of as the most probable for that distribution form given our data.

However, the rapidity with which the likelihood decreases as we move away from the maximum also gives us information. In particular we can define confidence intervals for the distribution fit parameters using the relation between the χ^2 distribution and the likelihood ratio:

$$\ln(\mathrm{R}) = \ln(\mathrm{L}(\Theta, X)/\mathrm{L}(\Theta_{\mathrm{best}}, \mathrm{X})) \approx 0.5\chi^2(\mathrm{CL}, k) \tag{2}$$

where the number of degrees of freedom for the χ^2 distribution is equal to the number of parameters in the model, k.

For concreteness, we define $\mathrm{RDM}_{\mathrm{HIST}}$ as the ratio of the mode of the failure distribution (D_{Fmode}) to $D_{99/90}$ given by the model, so that:

$$\alpha_{\mathrm{mod}} \equiv 2 \times RDM_{HIST} = 2 \times D_{F\,\mathrm{mod}\,e} \big/ D_{99/90} \tag{3}$$

This definition is analogous to that used for the PCC method and allows direct comparison of α for the 3 models. However, in reality, the exact definition is not critical, since reliability begins to drop rapidly once α drops below some critical value.

For Model 2, our use of likelihood means we will be using the Aggregate methodology.[1] However, we have determined that this yields equivalent results to the PCC method.

As mentioned above, although we use likelihood to determine best-fit values and confidence intervals for our models, likelihood can not be used to compare models with different numbers of parameters, k. More complicated models will usually fit the data better than simpler ones whether or not the added complexity adds to our understanding. (For example, even if 3 data points trend linearly, a quadratic form gives an exact fit while actually giving us less useful information.) To address this model selection problem, we use Hirotugu Akaike's information criterion (AIC)[6]:

$$\mathrm{AIC} = -2\log[\mathrm{L}(\Theta, \mathrm{X})] + 2k \tag{4}$$

Since datasets in RHA tend to be small (<100 lots), we will use the form of AIC corrected for small sample size[7]:

$$\mathrm{AICc} = \mathrm{AIC} + 2k(k+1) \big/ (n-k-1) \tag{5}$$

AICc is comprised of 2 terms. The first term, proportional to the negative log likelihood, decreases as the model's goodness of fit to the data improves. The other term increases roughly linearly with the number of model parameters, k. Thus, models with small AICc account for the data economically (i.e. fewer parameters), while those with large AIC either give a poor fit or are unnecessarily complicated.

To justify using a more complicated model (higher k), the goodness of fit (likelihood) must increase exponentially in k. Thus, AIC expresses Occam's razor, limiting the complexity of the model to that needed to explain the data. However, AIC actually has a more fundamental origin: It is an unbiased estimator for the Kullback-Liebler (K-L) divergence[8]—a measure of a model's deviation from the true generating model. Minimizing the K-L divergence (or AIC) identifies the model that is closest to the real one, and this model will have the greatest predictive power.

Since its introduction, AIC has become a standard technique of model selection, with applications ranging from astronomy to microbiology and from software reliability to satellite design.[9] An idea of the usefulness of AIC can be gained from the fact that a Google Scholar search on 9/5/2009 yielded over 9800 citations of Akaike'a 1974 paper (reference 6).

Because the RH1014 dataset is the best behaved and the largest, we begin by applying our model-selection methodology to this dataset to see if it yields reasonable results. Figure 7 shows that AICc begins to favor model 3 (> 99% significance) for as few as 4 lots of data. Moreover, as we add data, the average AICc per lot added decreases. This means that the predictive power of the model is increasing. In contrast, there is no significant trend in Model 1 or Model 2. This is easy to understand for Model 1, where we fit each lot individually, making no use of past history. In contrast, because Model 2 assumes a single, unimodal failure distribution for all lots, it is not sufficiently sophisticated to use all of the information as we add data for more lots.

For the RH1014, the required margin α for model 3 converges to ~3—about the same level as for the Aggregate Method—but does so more quickly. (See figure 8.) Thus, although we could have determined an appropriate value for α by either model 2 or model 3, AICc indicates that model 3 more closely captures the part-to-part and lot-to-lot variation for the RH1014. Model 3 also gives us additional information, including the degree to which the mean failure level and the width about the mean (e.g. standard deviation) vary from lot to

lot. This important diagnostic could alert us to process changes, bad lots, etc. as we add more data to our sample.

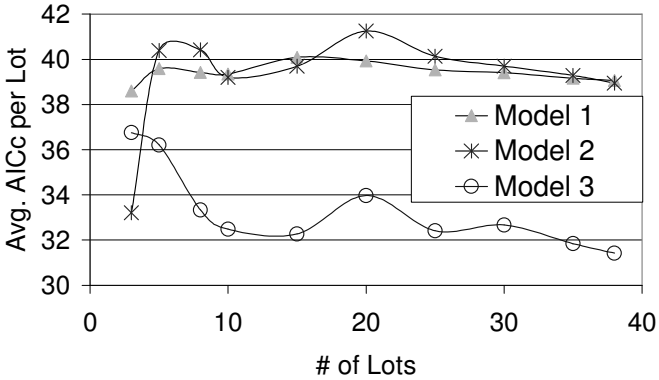

Fig. 7 Even for 4 lots of RH1014 data, Model 3's AICc is significantly less than for the other models, indicating that it will have more predictive power than the other models. Model 3's AICc also decreases as we add data, indicating it is better able to use the information in the data.

The advantage of model 3 is more evident for parts with complicated failure distributions—e.g. the OP07 with its roughly linear trend between mean lot failure level and standard deviation. Starting with model 3 applied to the lognormal fit parameters for individual OP07 lots, we capture the linear trend between lot mean failure and standard deviation seen in figure 4 as a constraint on the distribution of σ_{ln}. Specifically, we require σ_{ln} to be centered such that the trend holds on average, but allow the width about this central tendency to vary to give the best fit to the data. The resulting 3-parameter model significantly outperforms both Model 1 and Model 2 based on AICc.

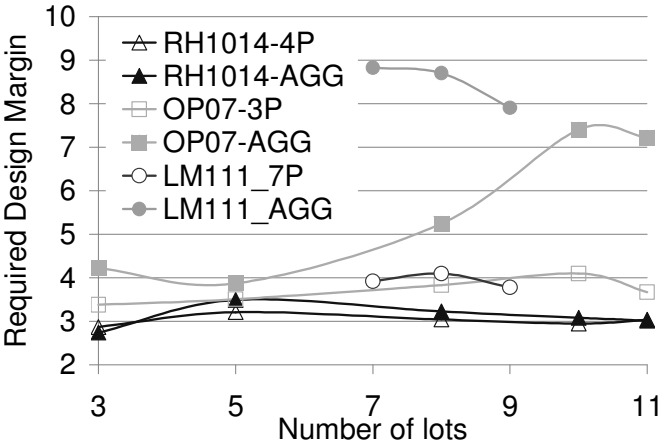

Fig. 8 For parts with well behaved failure distributions (e.g. RH1014), the Aggregate and 4-parameter modes yield equivalent values for α. However, for parts with complicated failure distributions, the more complicated model can reduce the estimate for α by ~2x from 7-9 to about 4.

To assess the bimodality of the LM111 we need a more complicated model. The EDA reveals two modes, but there is no clear difference in the mode widths. As such, we require 7 parameters for this model—two for the lognormal mean of each mode, two for the lognormal standard deviation and one parameter to specify the relative probabilities of each mode. This model, which reflects the bimodal lot-to-lot variability for LM111s, outperforms the aggregate model. The size of the softer mode is not less than a quarter and not more than twice the higher mode at the 90% CL based on the 9 lots of data available. For purposes of hardness assurance, we must assume that the flight lot belongs to the lower mode. Taking the ratio of the $D_{99/90}$ to the lower mode as our measure of lot-to-lot variation, we estimate α~4—roughly a factor of 2 less than the Aggregate method.

Although inference of lot-to-lot variability using the Aggregate and PCC methods suggested α>8-9 would be needed to bound degradation for the LM111 and OP07, more sophisticated statistical models show that α~4 is a more reasonable lower bound on margin needed to achieve Ps=99% @ 90% confidence. Figures 9 and 10 show that for α<4, Ps at the 90% confidence level and the confidence level we have in Ps =99% can decrease rapidly.

The introduction of model 3 allows us to handle not just well behaved data like that for the RH1014, but also data that pose challenges for conventional RHA methodologies. We can generalize model 3 by adding constraints or by adding additional modes (indeed, model 1 above can be viewed as a generalization where we add a mode for each additional lot). Similarly, if we had enough data, we could even treat failure distributions like that for the OP484 that have multiple modes within a wafer lot.

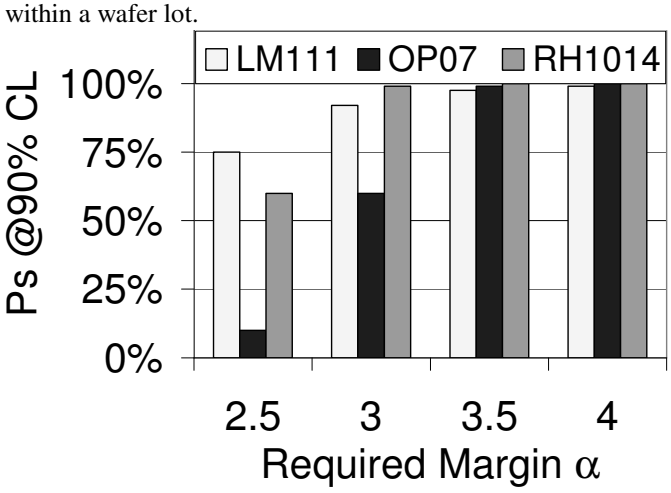

Fig. 9 For margins below 4, probability of success (Ps) (or confidence) decrease rapidly with decreasing required margin.

VI. Data in the Real World

The techniques and models outlined above yield insight into the importance of part-to-part and lot-to-lot variability for the parts studied. However, in most part qualification efforts, our historical data will be much more limited. The historical data that do exist may have been gathered using variable test methods and conditions, and the criteria for failure may be inconsistent from test to test. This raises the question of how we can apply these methods to generic data and whether the conclusions we have drawn (e.g. α_{min}~4) remain valid. To investigate the feasibility of statistical analysis of generic data, we conducted a meta-analysis of data for Analog Devices Op amps from the GSFC Radhome database.[10]

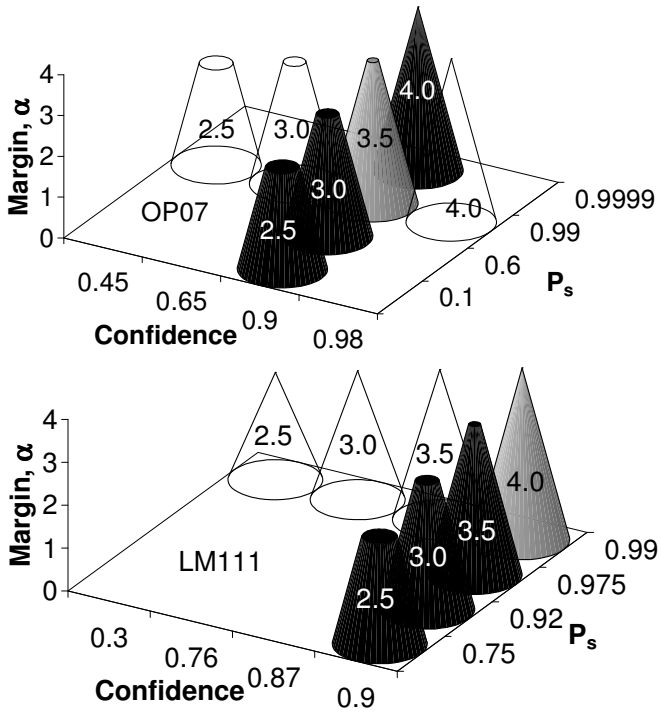

Fig. 10 Decreasing design margin (cone heights) decreases both probability of success (P_s) and the confidence we have in that probability. The figures above present the effects of decreasing margin on Ps at the 90% CL (dark cones) and the confidence level for Ps=99% (white cones) for the OP07 (upper) and LM111 (lower). The 99/90 PL is shown in grey.

The first task was to adopt a definition of failure dose we could apply unambiguously across all tests. The two most generally applicable definitions are dose of functional failure and the dose at which the first parameter for the device falls out of manufacturer's specifications. Since most devices are not tested to functional failure, we adopted the definition of failure dose as that where the first device parameter falls out of manufacturer's specifications. One potential disadvantage of this definition is that because it looks at absolute values of the parameters rather than their changes, one is sensitive to pre-radiation variability as well as radiation response. As such it is important to check for pre-rad variability.

Table III lists the 3 op amps with at least 3 test reports from a search of the Radhome database. Because we have limited statistics and limited confidence in the consistency of our test method from lot to lot, we estimate RDM_{HIST} by taking the ratio of mean failure levels for the hardest to the softest lot for each part (as in section III). By this criterion, only the OP400 variation would require α>4. However, with a mean lot failure dose of 5.3 krad(Si) and a standard deviation of 2.3, this part would be an unlikely candidate for historical data qualification. Other studies have also seen first parametric failure for the OP400 at between 3 and 7 krad(Si).[11]

TABLE III: ESTIMATED MINIMUM RDM α

Part Type	# lots	Mean Lot Fail. Dose μ	Inter-lot σ of μ	Ratio: Worst to Best μ	α based on means
OP27	3	29.5 krad(Si)	9	1.85	3.7
OP400	4	5.3 krad(Si)	2.3	3.2	6.4
OP77	3	16.4 krad(Si)	5.4	2	4

VII. MODEL SELECTION AND MODEL AVERAGING

In our current study, AICc strongly favored Model 3 over the other models. However, this is not always the case. When two or models have comparable values of AICc for a dataset, which one should we use. One approach is to take the model that gives the more conservative bound.

However, since AICc estimates a deviation (Kullbach-Liebler information) from the true model, a situation in which more than one model achieves roughly the same AICc indicates that each model captures part of the truth. One way[9] to capture as much information as possible is to average the results for all the theories, weighting the result for model *i* according to its Akaike weight, w_i

$$w_i = \exp(-\Delta_i / 2) / (\sum_{j=1}^{m} \exp(-\Delta_j / 2)) \quad (6)$$

where $\Delta_j = AIC_j - AIC_{min}$ gives the AIC distance. Such a weighted average often outperforms the results of any single model. If both distributions have the same number of parameters, the average reduces to weighting the results by their likelihood. However, Akaike weights allow combination of results for models with different levels of complexity.

VIII. FUTURE WORK AND CONCLUSIONS

Although waiving RLAT based on historical data always poses some risks, some TID RHA regimes are more effective than others at mitigating these risks. In this study, we have looked at one aspect of such mitigation—increasing the design margin for parts qualified with historical data to compensate for the less representative nature of historical data compared to RLAT data. In particular, we sought to determine the minimum margin α needed to ensure that we would not compromise reliability by use of historical data and to see how α changes as we add more data. In the process, we also found that the statistical model used to analyze the data could change our answer by up to a factor of 2. This led us to develop a method for selecting the best-performing model for our dataset. The results of our efforts suggest the following guidelines for TID RHA methodologies allowing qualification based on historical data:

1) Historical qualification should be attempted only when lot-to-lot and part-to-part variations are well behaved. Data for several lots (>3) are necessary to identify failure distribution pathologies and trends as well as for reliable estimation of lot-to-lot variability.
2) Minimum RDM, α, can be viewed as the product of two parts: the usual value used when RLAT data are

available RDM_{RLAT} (taken here to be 2) and a term quantifying lot-to-lot variability of historical data.

3) For DMBP methodologies, allowing RLAT to be waived for RDM<4 can significantly increase risk and decrease confidence for the parts so qualified.
4) Parts that have well behaved failure distributions but seem to exhibit high lot-to-lot variation may require analysis with a more complicated statistical model capable of reflecting trends or pathologies in the lot-to-lot variation of the parts.
5) Even modest samples of historical data (≥ 4 lots) can sometimes support more complicated statistical analyses than those typically carried out in RHA methodologies. This increases the information that can be inferred from the historical data.
6) Parts with low mean or broadly varying TID hardness are unsuitable candidates for qualification using historical data alone.
7) Provenance of historical data is crucial to effective historical qualification. Unless we resolve questions of foundry origin and process changes, the historical data may not represent of flight-lot performance.

In addition to the above guidelines, the analyses presented here demonstrate that even for modest datasets (3-4 lots), several statistical models can be applied and the appropriate model selected using criteria such as AIC. More complicated models offer greater flexibility to reflect features of real failure distributions and trends revealed during exploratory data analysis. They can also provide separate estimates of lot-to-lot and part-to-part variation, and so could be important for quality assurance and facilitate dynamical modeling of TID degradation. In addition, the use of likelihood not only facilitates model selection, but also model averaging and investigation of model dependence of RHA results.

Perhaps most important, the techniques discussed here offer a means of comparing RHA strategies with very different statistical properties. For instance, how does a strategy requiring 5× margin on the worst-performing part in a 3 lot sample compare to one requiring 4× margin on the 99/90 PL for a 4 lot sample? This facilitates consistent risk reduction across subsystems designed by different vendors with different RHA methodologies.

Finally, likelihood techniques are also amenable to application of Bayesian RHA methodologies. Because Bayesian methods represent the most efficient way of combining diverse types of information pertaining to a part's radiation performance, they are promising candidates for improving risk reduction and efficiency for TID RHA.

REFERENCES

[1] Ladbury, R. and Gorelick, J.L, "Statistical methods for large flight lots and ultra-high reliability applications," IEEE Trans. Nucl. Sci., Vol. 52, No. 6, pp. 2630 – 2637, 2005.

[2] "Ionizing dose and neutron hardness assurance guidelines for microcircuits and semiconductor devices", 2009. MIL-HDBK-814.

[3] R. L. Pease, "Microelectronic Piece Part Radiation Hardness Assurance for Space Systems, Atlanta, GA, July, 19. 2004 NSREC Short Course.

[4] A. Namenson, "Lot uniformity and small sample sizes in hardness assurance," IEEE Trans. Nucl. Sci., vol. 35, no. 6, pp. 1506–1511, Dec. 1988.

[5] J. Krieg *et al.*, "Hardness assurance implications of bimodal total dose response in a bipolar linear voltage comparator," *IEEE Trans. Nucl Sci.*, vol. 46, no.6, pp. 1627-1632. Dec. 1999.

[6] H. Akaike, "A New Look at the Statistical Model Identification," IEEE Trans. Auto Control AC-19:716-723, 1974.

[7] N. Sugiura, "Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections," Communications in Statistics, Theory and Method, A7: 13-26. 1974.

[8] S. Kullback and R. A. Leibler, *On Information and Sufficiency,* Ann. Math. Statist. Volume 22, Number 1 (1951), 79-86.

[9] Burnham, Kenneth P. and David R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. 2d ed. New York: Springer-Verlag.

[10] http://radhome.gsfc.nasa.gov/radhome/RadDataBase/RadDataBase.htm.

[11] P. Layton, L. Longden and E. Patnaude, "Compendia of Radiation Test Results," IEEE Radiation Effects Data Workshop 2007.