2017

Quantifying Similarity in Reliability Surfaces Using the Probability of Agreement

Nathaniel Stevens
University of San Francisco, ntstevens@usfca.edu

C. M. Anderson-Cook

Recommended Citation

Stevens, N.T., & Anderson-Cook, C.M. (2017). Quantifying Similarity in Reliability Surfaces Using the Probability of Agreement. Quality Engineering, 29(3), 395-408. http://dx.doi.org/10.1080/08982112.2017.1312004

Quantifying Similarity in Reliability Surfaces Using the Probability of Agreement

Nathaniel T. Stevens
University of San Francisco
San Francisco, California

Christine M. Anderson-Cook
Los Alamos National Laboratory
Los Alamos, New Mexico

Abstract:

When separate populations exhibit similar reliability as a function of multiple explanatory variables, combining them into a single population is tempting. This can simplify future predictions and reduce uncertainty associated with estimation. However, combining these populations may introduce bias if the underlying relationships are in fact different. The probability of agreement formally and intuitively quantifies the similarity of estimated reliability surfaces across a two-factor input space. An example from the reliability literature demonstrates the utility of the approach when deciding whether to combine two populations or to keep them as distinct. New graphical summaries provide strategies for visualizing the results.

Key Words: generalized linear models, reliability, equivalence testing, homogeneity of population characteristics, probability of agreement

1. INTRODUCTION

The importance of high-quality manufactured products is as important today as it was during World War II, when the reliability engineering discipline was born. Readers interested in the history of reliability engineering are referred to Bhamare et al. (2007) who describe the evolution of the field over the 20th century. More recently Meeker and Hong (2014) and Steinberg (2016) describe current challenges in reliability research in the context of Big Data, sensor technology and modern manufacturing processes. Although a rich literature exists for reliability analysis, much of it is devoted to effectively and accurately modeling product lifetimes and degradation. There has been little emphasis placed on the comparison of product reliabilities across different populations.

The question of whether different populations can be combined or treated as practically equivalent is one that occurs in a wide variety of applications. Equivalence testing (see Wellek, 2010, Szarka, 2014 and Anderson-Cook & Borror, 2016) provides a hypothesis testing based approach for comparison of static populations. A key difference between this approach and traditional hypothesis testing is where the emphasis of initial assumptions is placed. Traditional hypothesis testing assumes that the populations to be compared are equivalent until there is evidence to the contrary. Equivalence testing, on the other hand, starts with the assumption that the populations are different until there is sufficient evidence to assert that they are practically equivalent (Richter & Richter, 2002). For instance, bioequivalence has gained considerable traction in recent years with comparison of drugs and pharmaceutical treatments (see Borman et al., 2009 and Limentani et al., 2005 for details). Measurement system comparison also focuses on evaluating the similarities between systems (Barnett and Youden, 1970; Bland and Altman, 1999; Ludbrook, 2002; Barnhart et al., 2007) and there are opportunities to leverage principles and techniques from this literature as well.

In the context of comparing population reliabilities Stevens and Anderson-Cook (2016) adapt the probability of agreement (PA) (Stevens et al., 2015; 2017) and use it as a flexible means of assessing similarity between reliabilities. This was proposed as an alternative to hypothesis testing approaches that evaluate the significance of parameters in nested models to determine whether individual or pooled models are warranted (see e.g., Meeker and Escobar, 1998 or Lu and Anderson-Cook, 2015).

If interest lies in evaluating whether related populations of systems potentially share reliability characteristics as a function of age and usage, then the goal is to quantify the similarity of the modeled relationships between reliability and explanatory factors. An advantage of the PA approach over traditional hypothesis testing approaches is that the comparison of population reliabilities is based on user-specified acceptable tolerances for practically inconsequential differences in reliability. In short, the PA approach values practical significance over statistical significance. In the present article we extend this method and propose new graphical summaries to allow for assessment of a larger explanatory variable space in this article.

There are several scenarios where we might anticipate similar behavior to be observed for different populations or sub-populations. For example in the defense application described in Lu and Anderson-Cook (2015), a single population of complex munitions produced by a single manufacturer was distributed between the Army and Navy. The units were produced in overlapping time periods, but the observed reliability of the systems as a function of age differs between the two subpopulations. Lu and Anderson-Cook (2015) consider the number of transfers that each unit experiences as an available proxy for the usage that the system experiences. Combining age and usage information enhances the reliability model and facilitates examination of how differently the predicted reliabilities behave depending on the model used.

With recent developments in sensor and tracking technology, it is becoming increasingly common to have a detailed history of each unit. This knowledge can help surveillance programs by providing a more comprehensive understanding of the mechanisms driving change in reliability, as well as allowing individualized management of more expensive assets (see Anderson-Cook et al., 2015 for more details). The goal of managing the units is to remove them from service before they have a high probability of failure, but not too early as the expense of replacing units is substantial. For other systems, understanding the drivers of reliability changes can allow for more realistic warranty programs or better prediction of future maintenance needs.

In this paper, we re-examine the data of Lu and Anderson-Cook (2015) and consider the following questions:

1. Depending on the users' assessment of what constitutes a practically important difference in reliability, are the relationships between the response, reliability, and the explanatory variables, age and usage (number of transfers), sufficiently similar to be treated as the same?
2. If the usage rate (an increase or decrease in the number of annual transfers) in the future was going to be changed in one or both of the services, would there be a model that could adequately predict future reliability?
3. If a unit were to be transferred between the Army population and the Navy population, would there be a model that could be used to predict its future reliability?

To answer these questions we adapt the probability of agreement approach in Stevens and Anderson-Cook (2016) to accommodate the higher dimensionality of the explanatory variable space (now including both age and usage data), and we provide new graphical summaries for this setting. The approach generalizes to other scenarios where the goal is to compare the relationship between inputs and the predicted reliability for several populations based on what constitutes a practically important difference.

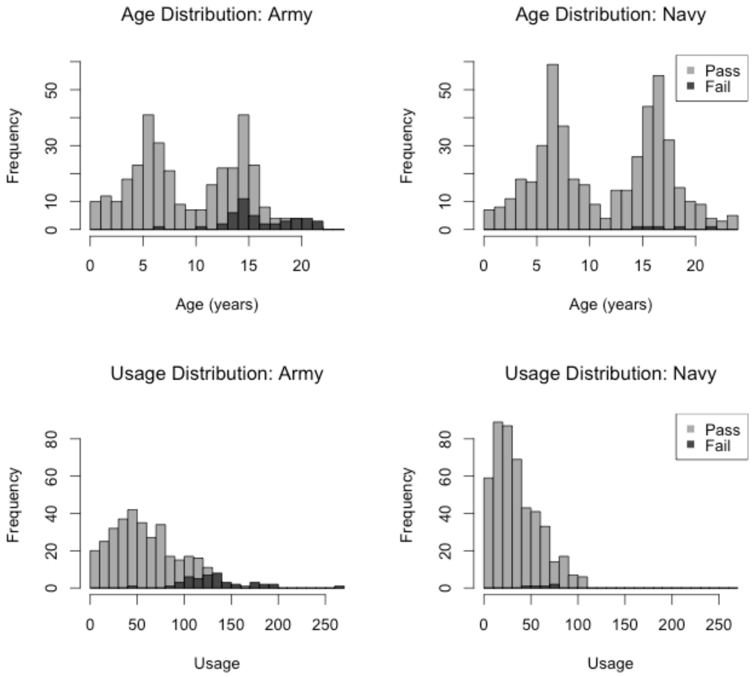


Figure 1: Age and Usage Distributions in Army and Navy Populations

We first revisit the results of Lu and Anderson-Cook (2015), where the observed passes and failures for each of the Navy and Army populations are shown in Figure 1, as both a function of age and usage. We see from the top row of Figure 1 that the age of units when tested is similar for the two populations (0-25 years), but there are substantially more observed failures for the Army system. In the bottom row of Figure 1, we see that in general the Army units experience many more transfers than most of the Navy units, and that these units with higher usage tend to be associated with the majority of the failures. Figure 2 is a scatterplot of age and usage for the two populations, indicating a much higher usage rate for the Army units than units of a similar age from the Navy population.

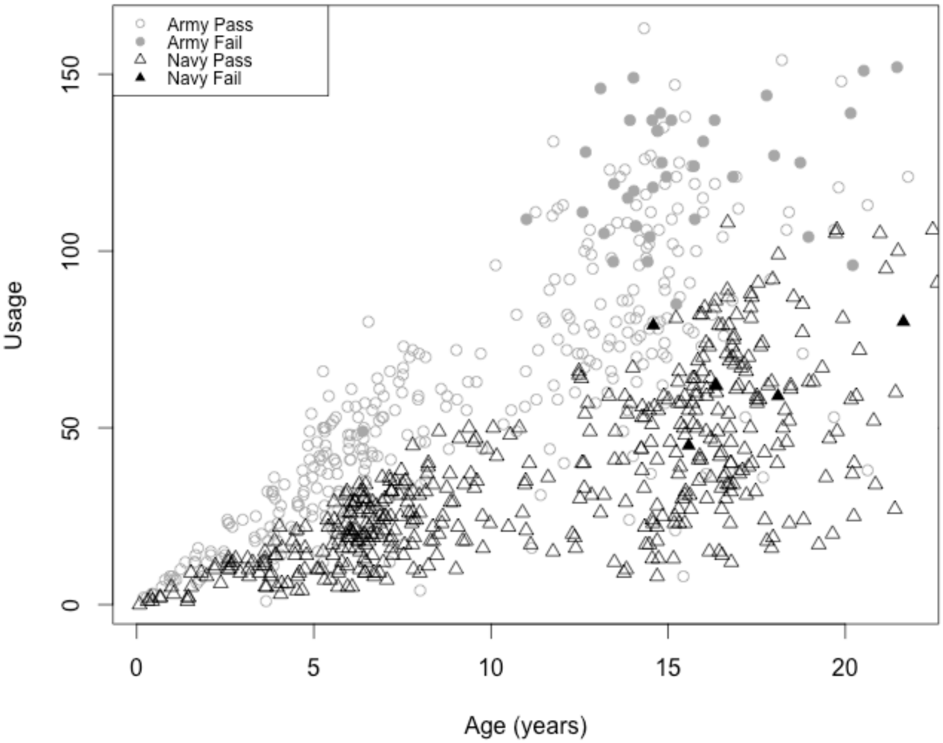


Figure 2: Scatterplot of Usage versus Age in Army and Navy Populations

In settings like this, reliability is assessed by destructively testing $i = 1,2,\dots,n$ units in each of $j = 1,2$ populations. Here $j = 1$ corresponds to the Army population and $j = 2$ corresponds to the Navy population. The response variable of interest is binary and can be described as follows:

$$Y_{ij} = \begin{cases} 1 \text{ if unit } i \text{ from population } j \text{ passes} \\ 0 \text{ if unit } i \text{ from population } j \text{ fails} \end{cases}$$

In population j, we assume that the reliability of unit i depends on its age and/or its usage, respectively denoted a_{ij} and u_{ij}. Conditional on the age and/or usage of the unit, we define the reliability of the system as the probability that the unit passes the destructive test, i.e. $P\left(Y_{ij} = 1\right)$, and denote this probability by π_{ij}.

To model the relationship between π_{ij} and a_{ij} and u_{ij}, Lu and Anderson-Cook (2015) use probit regression models of the form

$$\pi_{ij} = \Phi\big(\beta_{0,j} + \beta_{1,j} a_{ij}\big) \tag{1}$$

and

$$\pi_{ij} = \Phi\big(\beta_{0,j} + \beta_{1,j} a_{ij} + \beta_{2,j} u_{ij}\big) \tag{2}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Lu and Anderson-Cook (2015) estimate the model parameters from the data using a Bayesian analysis. For this paper, we fit the same models, but use a frequentist approach with maximum likelihood as the estimation procedure. In addition, since it is possible that the changes in reliability might be explained solely by the number of transfers, we also consider a usage-only model for prediction of reliability of the form

$$\pi_{ij} = \Phi\big(\beta_{0,j} + \beta_{2,j} u_{ij}\big) \tag{3}$$

Table 1: Summary of maximum likelihood parameter estimates with their associated 95% confidence intervals for the fitted models. Entries are displayed as MLE (95% CI).

Model	β_0	β_1	β_2
Army – age only	3.64 (2.79, 4.49)	-0.19 (-0.25, -0.13)	·
Navy – age only	3.71 (2.16, 5.27)	-0.10 (-0.19, -0.01)	·
Army – usage only	4.27 (3.32, 5.21)	·	-0.03 (-0.04, -0.02)
Navy – usage only	3.18 (2.29, 4.06)	·	-0.02 (-0.03, -0.004)
Army – age and usage	4.29 (3.12, 5.45)	-0.003 (-0.10, 0.10)	-0.03 (-0.04, -0.02)
Navy – age and usage	3.69 (2.14, 5.23)	-0.06 (-0.17, -0.06)	-0.01 (-0.03, 0.001)

Table 1 shows a summary of the estimated model parameters for each of the three sets of models, for the two populations of units. While there are differences in the observed estimates, it is difficult from this summary to assess the similarity of the curves/surfaces and how appropriate it might be to combine the curves into a single combined population with a common relationship for reliability. The PA methodology outlined in this article provides a formal, yet intuitive, means of quantifying such similarity and it helps to inform the decision of whether or not to combine populations based on what sized differences in reliability are considered important. We examine plots of the relationships in later sections of the paper.

The remainder of the article is organized as follows: In Section 2 we define the probability of agreement for this setting, and highlight adaptations from previous work. In particular, we describe new graphical methods of summarizing agreement in 3-dimensions. Then, in Section 3, we illustrate the use of this methodology by quantifying and displaying agreement between the Army and Navy munitions reliability surfaces, and we address questions 1-3 posed previously. In Section 4 we describe an R Shiny app (Shiny, 2016) that we have developed to aid practitioners in applying this methodology in a user-friendly, open source, environment. Finally in Section 5, we close with a summary and discussion of possible extensions.

2. PROBABILITY OF AGREEMENT

Stevens and Anderson-Cook (2016) propose using the probability of agreement to compare reliabilities in two different populations. In that work, a system's reliability was modeled as a function of age only, and the PA methodology quantified the similarity between reliability curves. In Section 2.2, we extend the PA methodology to the comparison of reliabilities when reliability is modeled as a function of more than one explanatory variable, here both age and usage. In this case the PA is used to quantify the similarity of 3-dimensional reliability surfaces. But first, in Section 2.1, we describe the statistical intuition behind the PA using a simple example involving the consideration of means from normally distributed populations.

2.1 Statistical Intuition

At its most basic level, the PA can be thought of as a metric that assesses the similarity of two distributions by quantifying the spread of the distribution of their differences. For example, consider drawing a random sample from each of two independent normal distributions: $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_i \sim N\left(\mu_y, \sigma_y^2\right)$ for $i = 1,2, \dots, n$. Interest may lie in the comparison of μ_x and μ_y. In particular, we may be interested in knowing what the true difference $\mu_x - \mu_y$ is. Of course since the true means are unknown, this difference must be inferred from sample data by comparing the estimates $\hat{\mu}_x$ and $\hat{\mu}_y$. The distribution of $\tilde{\mu}_x - \tilde{\mu}_y$ (i.e., $\bar{X} - \bar{Y}$) provides insight into what types of values $\hat{\mu}_x - \hat{\mu}_y$ can be expected and whether the true difference $\mu_x - \mu_y$ is large or small. Note that we overscore Greek letters with a circumflex (hat) to denote a parameter estimate, and we use a Greek letter overscored by a tilde to denote the corresponding estimator (a random variable).

Agreement would be indicated if the distribution of $\tilde{\mu}_x - \tilde{\mu}_y$ was narrow and centered at zero. We can formally quantify the spread of this distribution with a metric such as $\theta =$

$P(|\tilde{\mu}_x - \tilde{\mu}_y| < \delta)$ where δ is some tolerance. Large values of this probability indicate that the distribution is mostly contained in the interval $(-\delta, \delta)$ and hence that μ_x and μ_y are sufficiently similar. On the other hand, small values of this probability indicate that the spread of the distribution tends to be wider than this interval, and hence that μ_x and μ_y are sufficiently different. Given that $\tilde{\mu}_x = \bar{X} \sim N(\mu_x, \sigma_x^2/n)$ and $\tilde{\mu}_y = \bar{Y} \sim N(\mu_y, \sigma_y^2/n)$, we have $\tilde{\mu}_x - \tilde{\mu}_y \sim N(\mu_x - \mu_y, (\sigma_x^2 + \sigma_y^2)/n)$, and so θ can be written as

$$\theta = \Phi\left(\frac{\delta - (\mu_x - \mu_y)}{\sqrt{(\sigma_x^2 + \sigma_y^2)/n}}\right) - \Phi\left(\frac{-\delta - (\mu_x - \mu_y)}{\sqrt{(\sigma_x^2 + \sigma_y^2)/n}}\right)$$

where $\Phi(x)$ represents the standard normal cumulative distribution function evaluated at x. An estimate of this probability is found by substituting estimates of μ_x, μ_y, σ_x, and σ_y.

In Figure 3 we explore how estimates of θ are related to the true difference in population means, $|\mu_x - \mu_y|$, for a variety of sample sizes n. Note that the plot is strictly of $\hat{\theta}$ versus the standardized true difference $|\mu_x - \mu_y|/\sigma$, where for simplicity in this example we assume $\sigma_x = \sigma_y = \sigma$. For each of the values of n considered, we see that larger values of the true difference are associated with smaller values of $\hat{\theta}$, and small values of this difference correspond to larger values of $\hat{\theta}$. Larger sample sizes give a stronger signal of the result, but the overall pattern remains the same. Note that the inflection point depends on δ. This agrees with intuition: if μ_x and μ_y are truly similar/different $\hat{\theta}$ will be able to identify it. It is this intuition that is the basis for understanding how the PA quantifies agreement between population reliabilities.

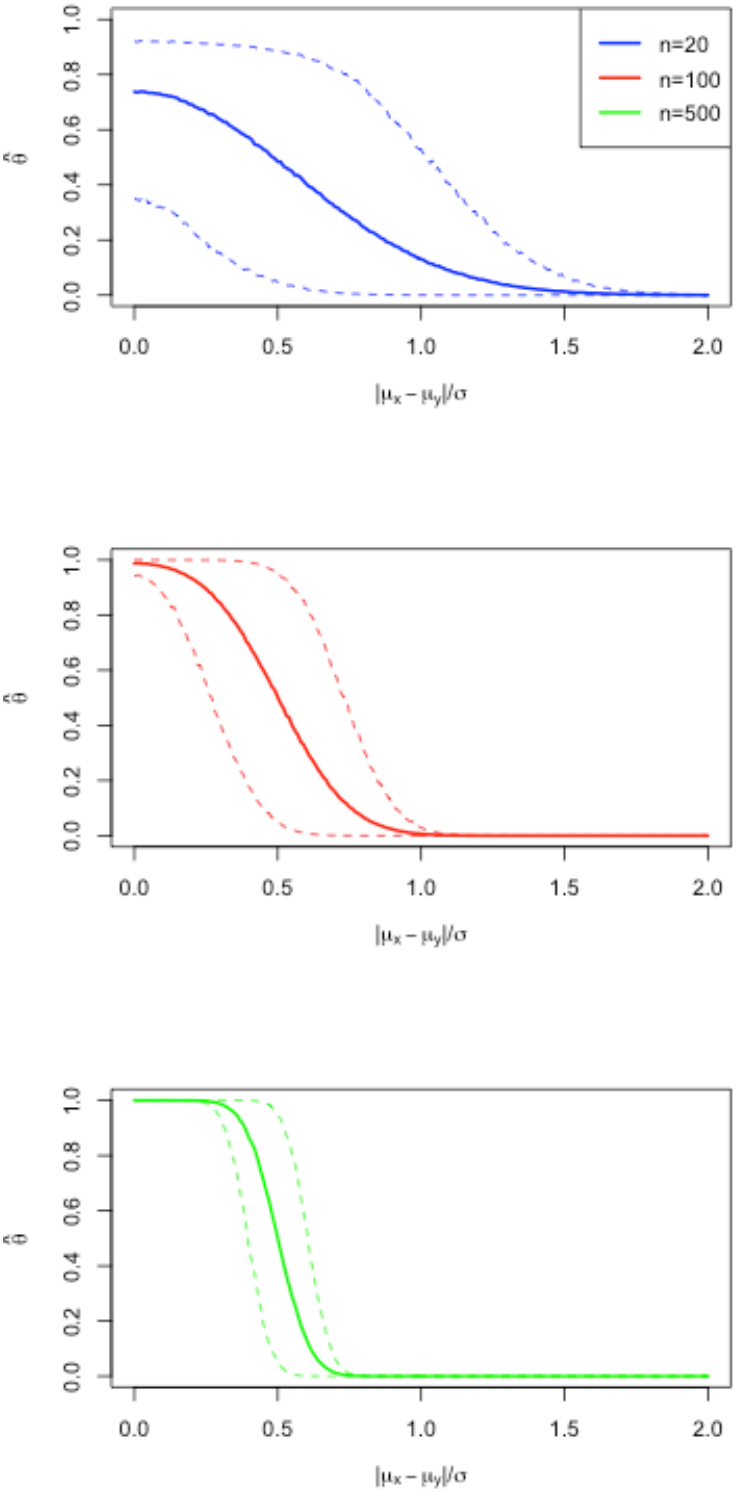


Figure 3: Estimated agreement probabilities versus true difference when $\delta = 0.5\sigma$. Solid and dashed lines respectively represent the mean and 5th and 95th percentiles of the $\tilde{\theta}$ distribution.

2.2 Extending the Probability of Agreement

The probability of agreement is a simple, yet powerful, method of quantifying agreement between two populations. In this context, it facilitates an informed assessment of the similarity between population reliabilities. A major advantage of the PA is that non-statisticians can easily interpret it, and this interpretation does not change even when the underlying reliability model is changed. For example, the PA is easily adapted to logistic, as opposed to probit, regression models. Furthermore, as illustrated in this article, even when the dimension of the explanatory variable space is increased, the PA can still be calculated and interpreted in the same way as the simpler situation explored by Stevens and Anderson-Cook (2016).

Another advantage of the PA is that it is a practical and contextually relevant method of comparison. Unlike a hypothesis testing approach to population comparison, the PA methodology emphasizes practical importance over statistical significance. Not only does the user specify the width of the indifference region they are comfortable with, the user also decides how large the PA should be to combine populations. In Section 3, we illustrate the PA methodology and this decision-making process with the munitions example introduced in Section 1.

Recall that the reliability of the system in population j, π_{ij}, is related to the system's age, a_{ij}, and usage, u_{ij}, with a generalized linear model via a probit link function: $\pi_{ij} = \Phi\big(\beta_{0,j} + \beta_{1,j}a_{ij} + \beta_{2,j}u_{ij}\big)$. See Myers et al. (2010) or McCullagh and Nelder (1989) for a review of generalized linear models. While the methodology discussed here uses a probit link for illustration, logistic regression with a logit link can also be easily accommodated. In either case maximum likelihood estimation is used to estimate $\beta_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$, and hence π_{ij}. In the case of probit regression we have $\hat{\pi}_{ij} = \Phi\big(\hat{\beta}_{0,j} + \hat{\beta}_{1,j}a_{ij} + \hat{\beta}_{2,j}u_{ij}\big)$. To compare two reliability surfaces from two different populations, we compare the fitted probit regression surfaces $\hat{\pi}_1$ and $\hat{\pi}_2$, and quantify the similarity between them. This is achieved by comparing the distribution of the corresponding estimators $\tilde{\pi}_1$ and $\tilde{\pi}_2$.

To assess the agreement between population reliabilities we quantify the similarity of two probit regression surfaces by extending the PA approach to the scenario in which reliability is modeled as a function of both age and usage. In particular, we use the following metric:

$$\theta(a_i, u_i) = P(|\tilde{\pi}_{i1} - \tilde{\pi}_{i2}| \leq \delta | a_i, u_i) \qquad (4)$$

where $\tilde{\pi}_{ij}$ denotes the reliability estimator for a system from population $j = 1,2$ at age a_i and with usage level u_i. Here $(-\delta, \delta)$ represents the interval within which differences in reliability are considered practically inconsequential. To interpret this metric, consider a value of 0.95: in this case we would expect 95% of the $\tilde{\pi}_{i1} - \tilde{\pi}_{i2}$ distribution to be contained within $(-\delta, \delta)$, which provides evidence to believe that the true difference in population reliabilities, $\pi_{i1} - \pi_{i2}$, is acceptably small. Note in using this approach it would be inappropriate to conclude that there is a 95% chance that the true difference in population reliabilities, $\pi_{i1} - \pi_{i2}$, is within $(-\delta, \delta)$. Such a statement would only be appropriate in a Bayesian framework. We discuss this further in Section 5.

Since an available R Shiny app may be used to implement this approach, we have moved some of the technical details of the method to the Appendix. As demonstrated in this Appendix the reliability estimators, $\tilde{\pi}_{ij}$, approximately follow a normal distribution. Consequently we have:

$$\theta(a_i, u_i) \cong \Phi\left(\frac{\delta - E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}{\sqrt{Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}}\right) - \Phi\left(\frac{-\delta - E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}{\sqrt{Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}}\right) \quad (5)$$

where, again, $\Phi(\cdot)$ is the standard normal cumulative distribution function and

$$E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2}) = E(\pi_{i1}) - E(\pi_{i2}) = \pi_{i1} - \pi_{i2}$$

$$Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2}) = Var(\tilde{\pi}_{i1}) + Var(\tilde{\pi}_{i2}) = \sigma_{i1}^2 + \sigma_{i2}^2$$

Note that the reliability estimators $\tilde{\pi}_{ij}$ for different populations are based on testing different units and hence assumed independent. We use σ_{ij}^2 to denote the asymptotic variance of $\tilde{\pi}_{ij}$. The interested reader is referred to the Appendix for a thorough discussion of the development of this quantity based on asymptotic likelihood theory.

The metric $\theta(a_i, u_i)$ takes on different forms depending on the link function used, but with the probit link it can be stated in terms of $\boldsymbol{\beta} = (\beta_{0,1}, \beta_{1,1}, \beta_{2,1}, \beta_{0,2}, \beta_{1,2}, \beta_{2,2})^T$:

$$\theta(a_i, u_i; \boldsymbol{\beta}) = \Phi(c_1(\boldsymbol{\beta})) - \Phi(c_2(\boldsymbol{\beta})) \quad (6)$$

where

$$c_1(\boldsymbol{\beta}) = \frac{\delta - \left(\Phi(\beta_{0,1} + \beta_{1,1}a_i + \beta_{2,1}u_i) - \Phi(\beta_{0,2} + \beta_{1,2}a_i + \beta_{2,2}u_i)\right)}{\sqrt{\boldsymbol{x}^T\left(\phi(\beta_{0,1} + \beta_{1,1}a_i + \beta_{2,1}u_i)^2 I^{-1}(\boldsymbol{\beta_1}) + \phi(\beta_{0,2} + \beta_{1,2}a_i + \beta_{2,2}u_i)^2 I^{-1}(\boldsymbol{\beta_2})\right)\boldsymbol{x}}}$$

$$c_2(\boldsymbol{\beta}) = \frac{-\delta - \left(\Phi(\beta_{0,1} + \beta_{1,1}a_i + \beta_{2,1}u_i) - \Phi(\beta_{0,2} + \beta_{1,2}a_i + \beta_{2,2}u_i)\right)}{\sqrt{\boldsymbol{x}^T \left(\phi(\beta_{0,1} + \beta_{1,1}a_i + \beta_{2,1}u_i)^2 I^{-1}(\boldsymbol{\beta_1}) + \phi(\beta_{0,2} + \beta_{1,2}a_i + \beta_{2,2}u_i)^2 I^{-1}(\boldsymbol{\beta_2})\right)\boldsymbol{x}}}$$

where $\boldsymbol{x}^T = (1 \quad a_i \quad u_i)$, $\boldsymbol{I}(\boldsymbol{\beta}_j)$ is the expected information matrix for $\boldsymbol{\beta}_j$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability density and cumulative distribution functions, respectively.

To use this approach to compare populations, an estimate of $\theta(a_i, u_i)$ is obtained by substituting the estimates of $\beta_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$ $(j = 1,2)$ into equation (6). This estimate, $\hat{\theta}(a_i, u_i)$, is then computed across a range of ages and usages and can be plotted as a 3-dimensional surface or a 2-dimensional contour plot which serves as a visual depiction of the agreement between the two reliability surfaces for different ages and usage levels. A lower confidence bound associated with each pointwise estimate is plotted similarly but separately to depict the uncertainty associated with estimation. This confidence limit is calculated using the standard error associated with the asymptotic normal distribution of the PA estimator $\tilde{\theta}(a_i, u_i)$. For full details on implementation, see the Appendix.

The perceived agreement between reliability surfaces depends critically on the definition of the indifference region $(-\delta, \delta)$. Narrower indifference regions mean that the user thinks that small differences matter and correspond to a reduction in the PA. On the other hand, wider indifference regions mean that the user feels that resultant conclusions are more robust to bigger differences between the surfaces, and correspond to higher values of the PA. The available R Shiny app has been created to automate the modeling, estimation and plotting associated with the proposed methodology. This app may also be used as an investigative tool for a practitioner to evaluate the sensitivity of the PA to the chosen value of δ. We discuss this Shiny app further in Section 4.

3. EXAMPLE

In Table 1 of Section 1, maximum likelihood estimates of equations (1), (2) and (3) for predicting reliability as a function of age, usage or both were provided. We now consider these relationships in combination with the probability of agreement methodology developed in the previous section to assess the similarity of the curves and surfaces between the Army and Navy populations.

As mentioned previously, a benefit of the PA approach to compare the reliability relationships is that there is flexibility for the decision-maker to determine a suitable threshold for what constitutes a practically important difference. After some discussion among the experts, there was a consensus that a difference of more than 5% (for example, reliability of 85% for one population versus 90% for the other) represented an important difference in terms of how the units might be managed. Using this threshold, Figure 4 shows both the Army and Navy reliability curves with their associated 95% confidence intervals based on equation (1), as well as the PA based on a fixed value of $\delta = 0.05$. In this case, we obtain the intuitive result that for the first 10 years, there is a very high PA for the two services, Army and Navy, but then between ages 10 and 13 years, the two curves diverge sharply and there is little chance of a common relationship summarizing the observed reliability pattern. In this case, it is unlikely that a formal method is needed to determine that the two curves should be treated separately.

Using the same threshold of $\delta = 0.05$, Figure 5 shows the estimated reliability curves based on equation (3) using only usage as a predictor as well as the PA plot. In this case, a different pattern emerges with the Navy reliability curve having huge uncertainty associated with it for usage (numbers of transfers) greater than 100. When we examine the bottom left plot of Figure 1 and Figure 2, it is clear why this is the case. There is almost no data available for the Navy population with observed numbers of transfers greater than 100. Hence, we should not be surprised that any model has difficulty determining what to expect for these values in the absence of information.

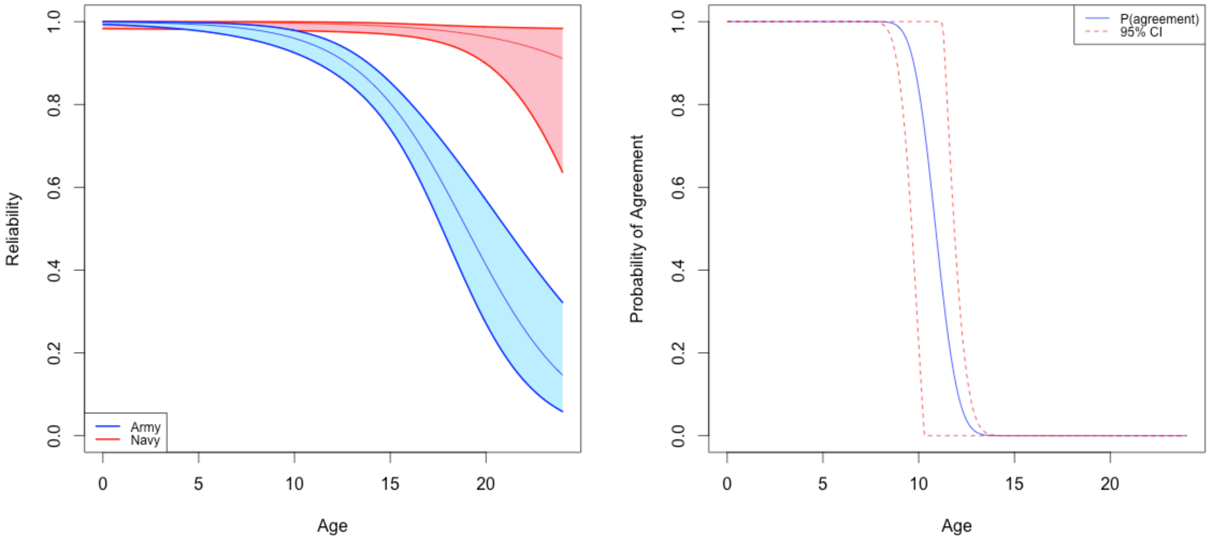


Figure 4: Comparing Army and Navy Reliability (which is a function of age only)
Left plot: Reliability as a function of age with 95% confidence intervals
Right plot: Probability of Agreement for $\delta = 0.05$

However, we might naively assume that since the Army curve with its associated 95% confidence interval is completely contained in the Navy 95% confidence interval, that it would be reasonable to treat the two populations as similar. When we examine the PA plot in the right side of Figure 5, we see that a formal assessment suggests that a common pattern of reliability as a function of usage is reasonable (with high PA) for usage values less than 90 transfers. However, for usage values in the range [90, 150], the PA drops sharply, and for many values in this range the PA is quite low. This suggests that there are many scenarios where the two curves would not agree. Although the confidence intervals overlap, when we look at the maximum likelihood curves themselves, these estimates are considerably more than 0.05 apart for the estimated reliability at a given usage value. The PA increasing for large numbers of transfers (>200) is a result of both curves approaching zero reliability, where they once again predict similarly. Hence, the PA is helpful as a tool to provide a formal quantification of differences between the populations that take into account both the values of the curves at different values of the explanatory variable, but also for incorporating the associated uncertainty with those estimates.

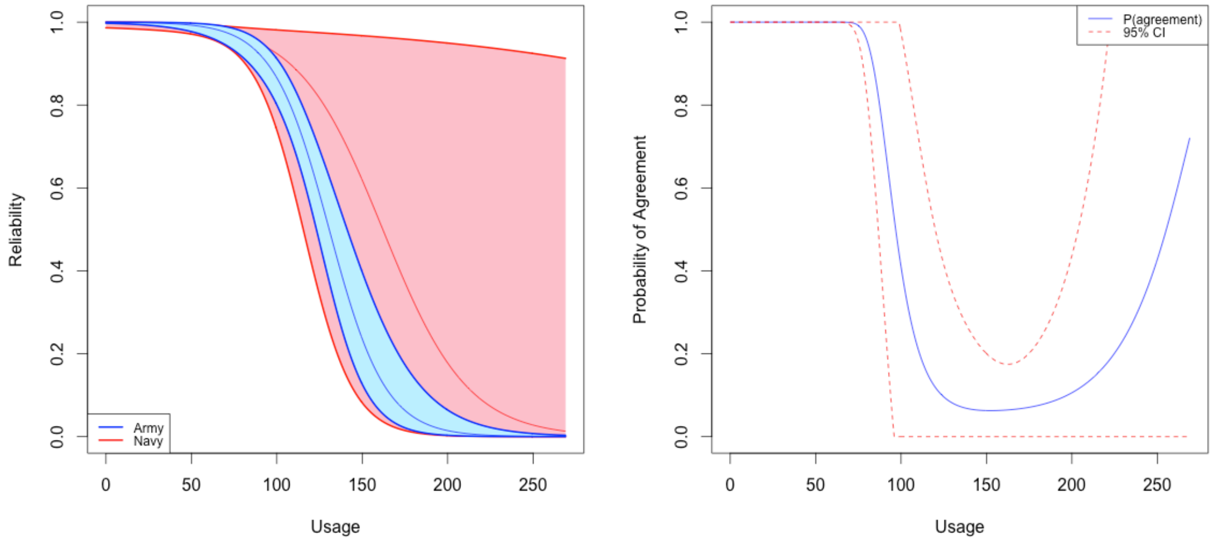


Figure 5: Comparing Army and Navy Reliability (which is a function of usage only)
Left plot: Reliability as a function of usage with 95% confidence intervals
Right plot: Probability of agreement for $\delta = 0.05$

Figure 6 shows the maximum likelihood estimates with their associated 95% confidence intervals for the combined age and usage reliability surfaces based on the models in equation (2). Recall from Figure 2 that age and usage are correlated for both the Army and Navy populations. The estimated model for Army shows that usage appears more influential in affecting reliability than changes in age. The 95% confidence interval around the estimated reliability surface has moderate uncertainty in the estimation throughout the region shown. In contrast, the estimated reliability surface for the Navy population shows more impact from both age and usage, with higher overall estimated values for reliability for much of the region. However, the 95% confidence interval around the surface is very wide and for a good portion of the region (where there is little or no observed data) the interval covers almost all values between 0 and 1. As with the usage only model shown in Figure 5, it is difficult to look at just the separate surfaces in Figure 6 and determine how reasonable it is to combine the two models into a single surface for both services.

Using the methodology developed in Section 2, Figure 7 shows a contour plot of the probability of agreement for the surfaces shown in Figure 6. The top plot shows a point estimate of the PA for $\delta = 0.05$ for each combination of age and usage throughout the region of interest, while the bottom plot shows the lower bound of the confidence interval for PA. For both plots, there is high agreement for ages between 0 and 15 years and for number of transfers (usage) between 0 and 50. For the point estimate of PA, the level of agreement drops from 100% to 0% between 50 and 120 transfers across the majority of the ages. For the lower confidence interval bound, the drop occurs much more quickly (as seen by how close the contour lines are) between 50 and 80 transfers, and most quickly for the oldest units in the range.

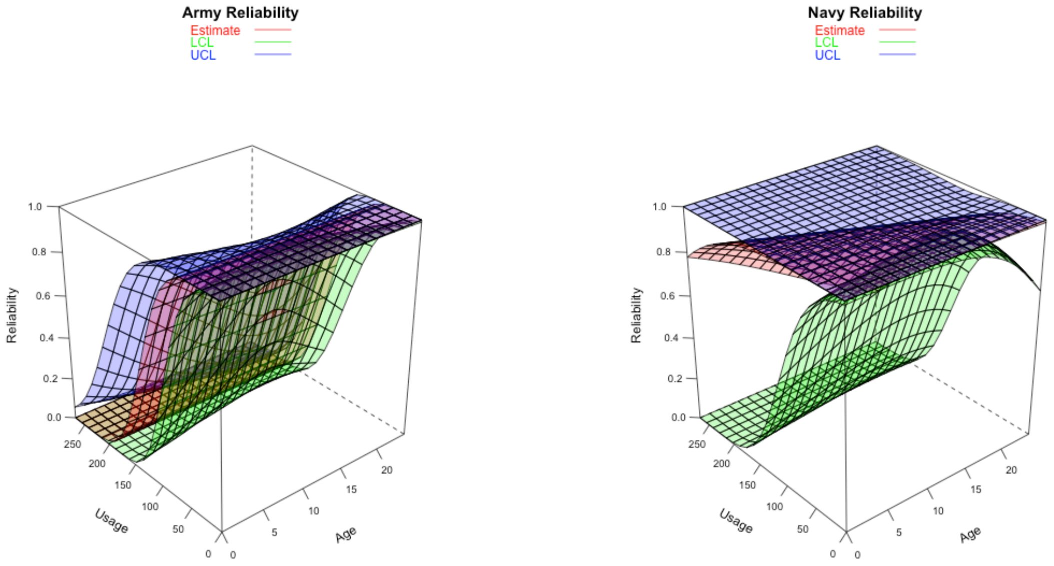
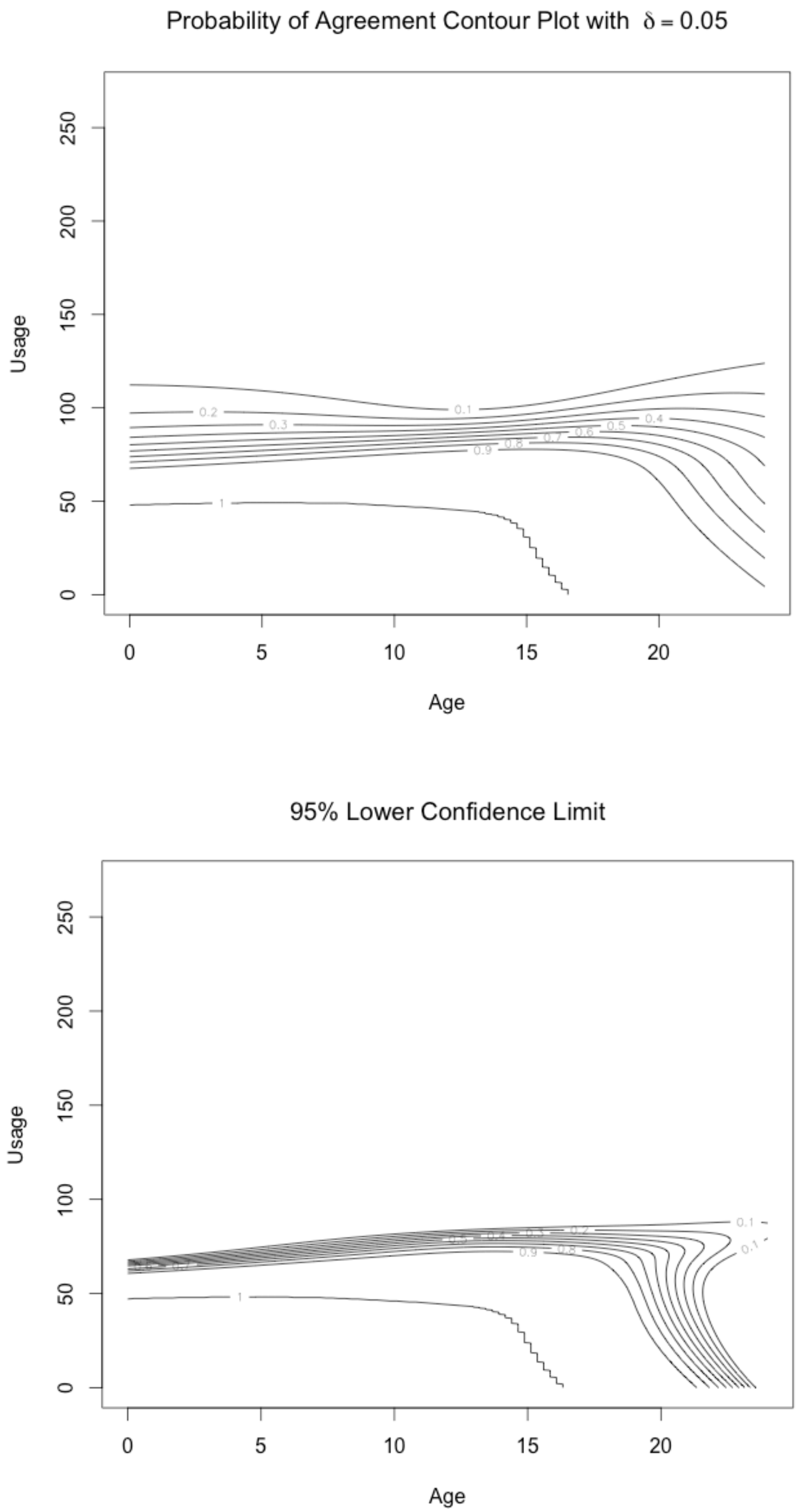


Figure 6: Army and Navy Reliability Surfaces

Comparing these results to the surfaces shown in Figure 6, we see that large values of the PA for low ages and usage arises due to the similarity of the reliability surfaces, and the small uncertainty, in this region. As well, low values of PA outside this region are unsurprising given the different shapes of the surfaces and the large uncertainty associated with the Navy reliability surface. Thus the PA contour plots provide a formal basis to conclude that there is very little agreement in reliability between the two services.



Probability of Agreement Contour Plot with $\delta = 0.05$

95% Lower Confidence Limit

Figure 7: Probability of agreement contour plots
Top plot: PA estimate with $\delta = 0.05$
Bottom plot: 95% Lower confidence limit

The *regions of sufficient agreement* plot shown in Figure 8 provides an alternative to the contour plots shown in Figure 7. In this plot, the user decides how large the PA should be to combine populations. In this example, the subject matter experts deliberated and then selected a θ threshold of 0.95, which means that they would be comfortable with combining the Army and Navy populations into a single population as long as the PA value was at least 0.95 for the chosen $\delta = 0.05$ value. Figure 8 shows two shaded regions – the blue region shows all combinations of age and usage for which $\theta \geq 0.95$. The red region shows the combinations of age and usage for which the lower confidence bound for θ is above 0.95.

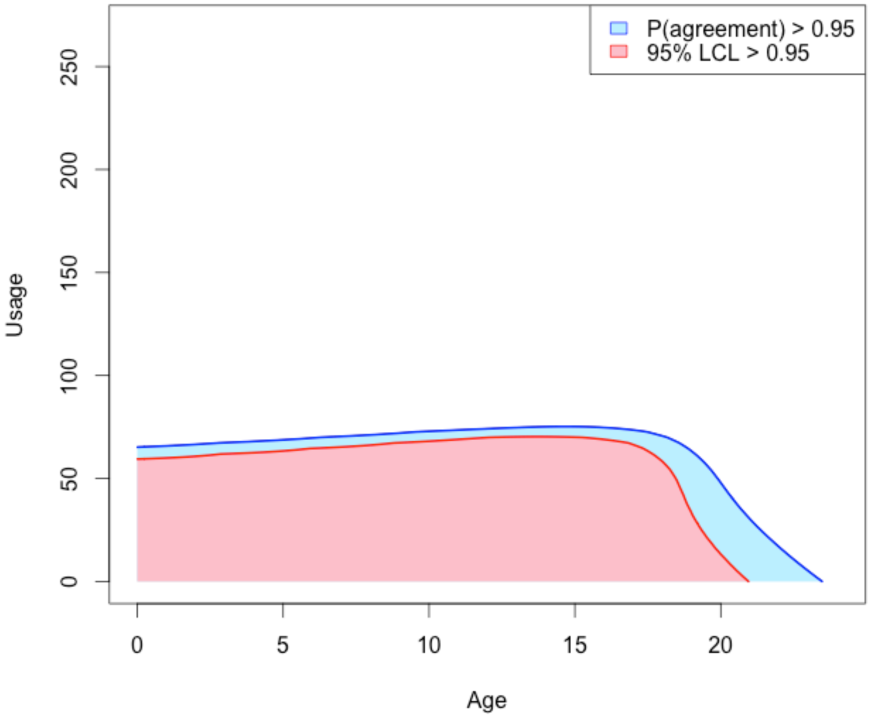


Figure 8: Regions of Sufficient Agreement

This single summary provides a straightforward view of all regions where the criteria specified by the decision-maker are satisfied for both the point estimate and lower bound on the confidence region of the PA. With the Shiny app described in the next section, the reader can experiment with different choices of δ and the threshold for θ to understand the impact of these choices on the decision. Decreasing δ shrinks the regions shown in Figure 8, since fewer combinations of age and usage will have a PA value within this more stringent requirement. Decreasing the threshold for θ means that the user is willing to accept a more lenient requirement.

Based on the results of the PA analysis for this case study, we now return to the three questions posed in Section 1. First, using the user's assessment that a difference in reliabilities greater than 5% ($\delta = 0.05$) was too large for combining the populations and the chosen threshold for the PA was 0.95, the best modeling strategy for estimating the reliability of the Army and Navy populations was to *use separate models* for the Army and Navy. This conclusion is based on the lack of sufficient agreement for large portions of the age and usage region of interest (age $\in [0,25]$ and usage $\in [0,200]$). It was hoped that since the Army and Navy populations had started as a homogeneous population at the time of manufacture, there might be an opportunity to use a unified model for both. However, given the nature of the data with little overlap in the age and usage combinations at later ages, current uncertainty in the models and estimated differences suggest that the combined model would not be suitable. It is also possible that there is an additional factor, such as environmental exposure (dry vs. humid) or handling procedures, which truly does change the relationship connecting reliability and the explanatory factors. Based on available data, this cannot be eliminated as a possibility.

Second, since the usage rate (i.e., the number of transfers per year) might change as a function of how the populations are manipulated, there was interest in having the ability to estimate reliability for different usage rates. This should be possible based on the two separate models. For example, suppose that the Army decided to increase or reduce the typical number of annual transfers in the future. This would require that predictions be made using the existing estimated model using equation (2), which would account for the changing relationship between age and usage. This is far superior to using either of the models in equations (1) or (3), which assume the same relationship between age and usage, where the single explanatory variable is used as a proxy for both contributions. Since the combined model was not considered appropriate, then these predictions should be made separately for each service.

Third, there was interest in being able to model the reliability for a unit that was moved between the services. For example, suppose that a 10-year-old Army unit with 50 transfers was shifted to the Navy. At the moment of the transfer, the Army model could be used to estimate the unit's current reliability. However, after a couple of years in the Navy, the units age and number of transfers would be known, but with separate models for each of the Army and Navy, it would be difficult to formally estimate the reliability of the unit without additional assumptions. Since one of the reasons the two population reliabilities were judged too different to be combined was lack of overlapping data in the age-usage space, this analysis should be repeated as new data become available. It is possible that future models (having been updated with additional data) would be similar enough to use a combined model for the Army and Navy populations. If this was true, then this model would allow for a formal estimate of the reliability of the transferred unit to be made.

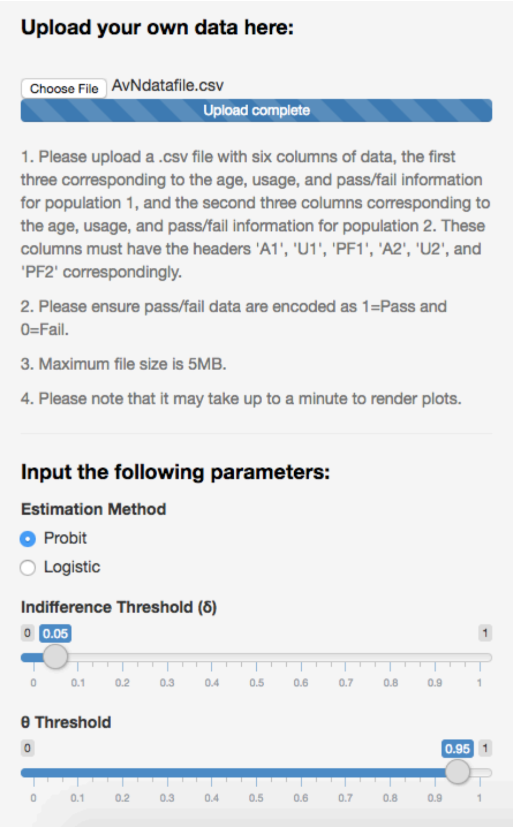
The process for exploring the probability of agreement for two populations of munitions based on age and usage has illustrated how to formalize these choices while incorporating the requirements of the decision-makers. While it is advantageous to combine the two populations when appropriate, it is beneficial to have the users articulate what sized differences in reliability are suitable for combining as well as the degree of certainty required for the value of the probability of agreement.

4. SHINY APP

As has been demonstrated, the probability of agreement is a useful tool for quantifying and summarizing the similarity between reliabilities in two populations. While its interpretation is intuitive and simple for a non-statistician to understand, the technical details associated with the methodology might hinder a practitioner from implementing it for their own use. To provide a user-friendly interface for this methodology, we have developed a ready-to-use R Shiny app that is freely available and can be accessed at the following URL: https://nathaniel-t-stevens.shinyapps.io/pagreement_app/. The purpose of this app is to provide a user with the opportunity to conduct a PA-based analysis without having to understand and implement the underlying technical details described in the Appendix.

To perform an analysis, a user need only upload a .csv file of data formatted in accordance with the instructions in the app. The munitions data from Section 3 may be downloaded from the Shiny app and used as an example of the formatting necessary to be compatible with the app. Once the data are uploaded, the analysis is automatically performed (after choosing either probit or logistic regression) with different analysis output displayed in different tabs. The output is organized as follows:

- Tab 1: Reliability Surfaces – two 3-dimensional plots are created that depict the reliability and uncertainty surfaces for each population as in Figure 6.
- Tab 2: PA Contours – the 2-dimensional contour plots of the PA and lower 95% confidence limit are created. Similar to Figure 7, these illustrate the dependence of the PA on age and usage.
- Tab 3: Agreement Regions – the regions of sufficient agreement plot is displayed here. Like Figure 8 it depicts a region of the explanatory variable space where the PA is deemed large enough for the two populations to be combined based on the user specified threshold.
- Tab 4: 2D Comparison – the comparison of 2-dimensional reliability curves based on age alone or usage alone is shown here. Plots of the reliability curves and the 2-dimensional PA (like Figures 4 and 5) are displayed.
- Tab 5: Raw Data – the raw data that is uploaded into the app is displayed here.



Upload your own data here:

Choose File AvNdatafile.csv

Upload complete

1. Please upload a .csv file with six columns of data, the first three corresponding to the age, usage, and pass/fail information for population 1, and the second three columns corresponding to the age, usage, and pass/fail information for population 2. These columns must have the headers 'A1', 'U1', 'PF1', 'A2', 'U2', and 'PF2' correspondingly.

2. Please ensure pass/fail data are encoded as 1=Pass and 0=Fail.

3. Maximum file size is 5MB.

4. Please note that it may take up to a minute to render plots.

Input the following parameters:

Estimation Method

● Probit

○ Logistic

Indifference Threshold (δ)

0 0.05 1

θ Threshold

0 0.95 1

Figure 9: User inputs for the R Shiny app with shown settings to match the results illustrated in Section 3

In addition to automating the PA analysis, with 'sliders' the R Shiny App provides the user with control over the width of the indifference region (δ), and the threshold of θ above which populations would be considered sufficiently similar. Figure 9 shows the left side of the app with the choices available to the user. Moving these sliders updates the corresponding thresholds, and hence the outputted plots in real time. This measure of control allows the user to investigate different values of the thresholds, which aids in the decision-making process. It is the hope of the authors that access to this app will minimize the barrier-to-entry and computational burden associated with implementing and using the methodology in practice.

5. SUMMARY & DISCUSSION

The probability of agreement provides an intuitive and practically useful means of comparing reliabilities in two populations. In this article we have extended the methodology proposed by Stevens and Anderson-Cook (2016) to account for the dependence of a system's reliability on its usage as well as age. While increasing the number of explanatory variables necessitates alternative ways of visualizing the PA (i.e., PA contours and the regions of sufficient agreement plot), the interpretation of the metric itself is unchanged. We have also developed an R Shiny App that facilitates straightforward application of the PA methodology where users need only know how to interpret the results, not the details of the theory that generates them.

We have illustrated these extensions and adaptations on the munitions example of Lu and Anderson-Cook (2015) and in this context have demonstrated the valuable insights it provides when making decisions about similarity between populations. In this example, given the specified level of practically inconsequential differences, the best choice was to continue to estimate the reliabilities of the Army and Navy populations separately. We demonstrated the methodology assuming that a constant difference in reliability was natural for this application. However, the choice of what difference to select for the PA process can be flexibly defined to change as a function of age or usage, or as a function of the uncertainty in the estimated surfaces. Incorporating these different forms for the indifference region (δ) would be straightforward to implement. We emphasize that the choice of indifference region as well as the threshold for acceptable sizes of the probability of agreement for each application is critical to the successful and meaningful use of this method.

The calculation of the PA can also be extended to higher numbers of explanatory variables in a straightforward way. However, adaptations for the graphical summaries to display the results are required. Taking slices of the explanatory variable space and showing contours of the PA can provide one option for viewing results in higher dimensions. In addition, the regions of sufficient agreement plot could be adapted by first selecting all combinations of the input space that satisfy the specified user requirements and then showing this region in selective 3-dimensional plots, where any additional explanatory variables are fixed at chosen values. These extensions warrant careful consideration in future work.

In Section 2.2 we noted that the probability of agreement as defined in (4) cannot be interpreted as the probability that the true difference in population reliabilities is within $(-\delta, \delta)$. For this interpretation to be valid the probability of agreement would have to have been defined as $P(|\pi_{i1} - \pi_{i2}| \leq \delta | a_i, u_i)$, which is only appropriate within a Bayesian framework. Recognizing the intuitive appeal of such an interpretation, we plan to consider a Bayesian probability of agreement (BPA) methodology in future work.

REFERENCES

1. Anderson-Cook, C.M. and Borror, C.M. (2015). The difference between ‘equivalent’ and ‘not different’. *Quality Engineering* 28: 249–262.
2. Anderson-Cook, C.M., Morzinski, J. and Blecker, K.D. (2015). Statistical model selection for better prediction and discovering science mechanism that affect reliability. *Systems* 3: 109-132.
3. Barnett, R.N. and Youden W.J. (1970). A revised scheme for the comparison of quantitative methods. *American Journal of Clinical Pathology* 54: 454-462.
4. Barnhart H.X., Haber M.J. and Lin L. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* 17: 529–569.
5. Bhamare, S.S., Yadav, O.P. and Rathore, A. (2007). Evolution of reliability engineering discipline over the last six decades: a comprehensive review. *International Journal of Reliability and Safety* 1(4): 377-410.
6. Bland, J.M., and Altman D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8: 135-160.
7. Borman, P., Chatfield, M., Damjanov, I. and Jackson, P. (2009). Design and analysis of method equivalence studies. *Analytical Chemistry* 81(24): 9849-9857.
8. Limentani, G.B., Ringo, M.C., Ye, F., Bergquist, M.L. and McSorley, M.L. (2005). Beyond the t-test: Statistical equivalence testing. *Analytical Chemistry* 77(11): 221-226.
9. Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology* 29: 527-536.

10. Lu, L. and Anderson-Cook, C.M. (2015). Improving reliability understanding, estimation and prediction with usage information. *Quality Engineering* 27: 304–316.
11. Maplesoft. (2016). Maple 18: Maple Inc., www.maplesoft.com.
12. Meeker, W.Q. and Escobar, L.A. (1998). *Statistical methods for reliability data*. John Wiley & Sons.
13. Meeker, W.Q. and Hong, Y. (2014). Reliability meets big data: opportunities and challenges. *Quality Engineering* 26(1): 102-116.
14. McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*. 2nd ed. Chapman & Hall / CRC Press.
15. Myers, R.H., Montgomery, D.C., Vining, G.G. and Robinson, T.J. (2010). *Generalized linear models with applications in engineering and the sciences*. 2nd ed. John Wiley & Sons.
16. Richter, S.J., and Richter, C. (2002). A method for determining equivalence in industrial applications. *Quality Engineering* 14: 375-380.
17. Shiny. (2016). Shiny: Easy web applications in R. RStudio Inc., http://shiny.rstudio.com/.
18. Steinberg, D.M. (2016). Industrial statistics: the challenges and the research. *Quality Engineering* 28(1): 45-59.
19. Stevens, N.T. and Anderson-Cook, C.M. (2016). Comparing the reliability of related populations with the Probability of Agreement. *Technometrics (in press)*. DOI: 10.1080/00401706.2016.1214180
20. Stevens, N.T., Steiner, S.H. and MacKay, R.J. (2015). Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Statistical Methods in Medical Research (in press)*. DOI: 10.1177/0962280215601133
21. Stevens, N.T., Steiner, S.H. and MacKay, R.J. (2017). Comparing heteroscedastic measurement systems with the probability of agreement. *Statistical Methods in Medical Research (in press).*
22. Szarka, J.L. (2014). Equivalence and Noninferiority Tests for Quality, Manufacturing and Test Engineers. *Journal of Quality Technology* 46: 378-380.
23. Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. 2nd ed. New York: CRC Press.

APPENDIX

In this section we provide technical details surrounding maximum likelihood theory as it relates to point and interval estimation of the probability of agreement. Recall that reliability is assessed by destructively testing $i = 1,2,\dots,n$ units in each of $j = 1,2$ populations. In this case the response variable Y_{ij} is binary, assuming a value of 1 if unit i from population j passes the destructive test and 0 if it fails. In population j, we relate the reliability of system i to its age and its usage, respectively denoted a_{ij} and u_{ij}. Conditional on the age and usage of the system, the reliability of the system is defined as the probability that the unit passes the destructive test: $\pi_{ij} = P\left(Y_{ij} = 1|a_{ij}, u_{ij}\right)$. In the context of probit regression, π_{ij} and a_{ij} and u_{ij} are related through the probit link function such that $\pi_{ij} = \Phi\left(\beta_{0,j} + \beta_{1,j}a_{ij} + \beta_{2,j}u_{ij}\right)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Maximum likelihood estimation is used to estimate $\beta_{0,j}$, $\beta_{1,j}$ and $\beta_{2,j}$, and hence π_{ij}. Consequently the estimator $\tilde{\boldsymbol{\beta}}_j = \left(\tilde{\beta}_{o,j}, \tilde{\beta}_{1,j}, \tilde{\beta}_{2,j}\right)^T \sim MVN\left(\boldsymbol{\beta}_j, \boldsymbol{I}\left(\boldsymbol{\beta}_j\right)\right)$ approximately, where $\boldsymbol{I}\left(\boldsymbol{\beta}_j\right)$ is the expected information matrix. By the Delta Method $\tilde{\pi}_{ij} = \pi_{ij}\left(\boldsymbol{\beta}_j\right) = \Phi\left(\tilde{\beta}_{0,j} + \tilde{\beta}_{1,j}a_{ij} + \tilde{\beta}_{2,j}u_{ij}\right)$ also follows an approximate distribution given by

$$\tilde{\pi}_{ij} \sim N\left(\pi_{ij}, \boldsymbol{\nabla}\pi_{ij}\left(\boldsymbol{\beta}_j\right)\boldsymbol{I}^{-1}\left(\boldsymbol{\beta}_j\right)\boldsymbol{\nabla}\pi_{ij}\left(\boldsymbol{\beta}_j\right)^T\right)$$

where $\boldsymbol{\nabla}\pi_{ij}\left(\boldsymbol{\beta}_j\right)$ denotes the 1×3 gradient vector given by

$$\begin{pmatrix} \frac{\partial \pi_{ij}\left(\boldsymbol{\beta}_j\right)}{\partial \beta_{0,j}} & \frac{\partial \pi_{ij}\left(\boldsymbol{\beta}_j\right)}{\partial \beta_{1,j}} & \frac{\partial \pi_{ij}\left(\boldsymbol{\beta}_j\right)}{\partial \beta_{2,j}} \end{pmatrix}$$

Note that all partial derivatives were taken symbolically using Maple (Maplesoft, 2016) to ease implementation and avoid coding errors.

Due to the asymptotic normality of each $\tilde{\pi}_{ij}$ the probability of agreement as defined in (4) is calculated as

$$\theta(a_i, u_i) \cong \Phi\left(\frac{\delta - E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}{\sqrt{Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}}\right) - \Phi\left(\frac{-\delta - E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}{\sqrt{Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2})}}\right)$$

where

$$E(\tilde{\pi}_{i1} - \tilde{\pi}_{i2}) = E(\pi_{i1}) - E(\pi_{i2}) = \pi_{i1} - \pi_{i2}$$

$$Var(\tilde{\pi}_{i1} - \tilde{\pi}_{i2}) = Var(\tilde{\pi}_{i1}) + Var(\tilde{\pi}_{i2}) = \sigma_{i1}^2 + \sigma_{i2}^2$$

with the asymptotic variance of $\tilde{\pi}_{ij}$ given by $\sigma_{ij}^2 = \boldsymbol{\nabla}\pi_{ij}\left(\boldsymbol{\beta}_j\right)\boldsymbol{I}^{-1}\left(\boldsymbol{\beta}_j\right)\boldsymbol{\nabla}\pi_{ij}\left(\boldsymbol{\beta}_j\right)^T$.

The probability of agreement $\theta(a_i, u_i)$ can be explicitly stated in terms of $\boldsymbol{\beta} = \left(\beta_{0,1}, \beta_{1,1}, \beta_{2,1}, \beta_{0,2}, \beta_{1,2}, \beta_{2,2}\right)^T$ as with equation (6) in Section 2.

By the asymptotic normality of $\boldsymbol{\beta}$ and the Delta Method, $\tilde{\theta}(a_i, u_i) = \theta\left(a_i, u_i; \tilde{\boldsymbol{\beta}}\right)$ also has an approximate multivariate normal distribution:

$$\tilde{\theta}(a_i, u_i) \sim MVN(\theta(a_i, u_i), \boldsymbol{\nabla}\theta(a_i, u_i; \boldsymbol{\beta})Cov(\boldsymbol{\beta})\boldsymbol{\nabla}\theta(a_i, u_i; \boldsymbol{\beta})^T)$$

where $\boldsymbol{\nabla}\theta(a_i, u_i; \boldsymbol{\beta})$ denotes the 1×6 gradient vector given by

$$\nabla\theta(a_i, u_i; \boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{0,1}} & \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{1,1}} & \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{2,1}} & \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{0,2}} & \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{1,2}} & \frac{\partial\theta(a_i, u_i; \boldsymbol{\beta})}{\partial\beta_{2,2}} \end{pmatrix}$$

and where

$$Cov(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{I}^{-1}(\boldsymbol{\beta_1}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}^{-1}(\boldsymbol{\beta_2}) \end{bmatrix}$$

Note that the off-diagonals in this block-diagonal matrix are zero because we assume that the estimators associated with different populations are independent and hence uncorrelated. Again we use Maple to automate taking the partial derivatives.

We then estimate $\theta(a_i, u_i)$ with $\hat{\theta}(a_i, u_i) = \theta\left(a_i, u_i; \hat{\boldsymbol{\beta}}\right)$ and pointwise confidence limits are calculated using the following standard error:

$$SE\left(\hat{\theta}(a_i, u_i)\right) = \sqrt{\nabla\theta\left(a_i, u_i; \hat{\boldsymbol{\beta}}\right) Cov\left(\hat{\boldsymbol{\beta}}\right) \nabla\theta\left(a_i, u_i; \hat{\boldsymbol{\beta}}\right)^T}$$

Stevens and Anderson-Cook (2016) thoroughly investigate the asymptotic properties of the probability of agreement estimator and conclude that it is unbiased when interpolating π_1 and π_2, may be slightly biased when extrapolating π_1 and π_2, and approximate confidence intervals as described here tend to be conservative.