Modeling Product Shelf Life

Approaches using Maximum Likelihood, Gamma Distributions, Cluster Analysis, and Linear Regression

Greco, Elizabeth
ecg83@cornell.edu

Hoare, Derek
hoared@kenyon.edu

Miao, Lin
miaol@kenyon.edu

Smith, Emily
smithem227@gmail.com

Advisor: Farnell, Elin
farnelle@kenyon.edu
Department of Mathematics and Statistics, Kenyon College

April 24, 2016

Abstract: Sproxil, Inc. is a company which produces PINs that manufacturers attach to products and which are then used by consumers to verify the authenticity of the product manufacturer. The goal of this work is to use existing PIN verification data provided by Sproxil to obtain a model for product shelf life: the length of time between PIN generation and product verification. We present several models that can be used to predict information about the shelf lives of various batches of products. We use maximum likelihood estimation to fit gamma distributions, which model the distributions of shelf lives. Cluster analysis is used to determine whether certain types of product batches have similar verification behavior. We find that the size of a product batch has an impact on how quickly verifications occur. Finally, regression analysis is used to find predictive relationships between variables related to shelf lives. We find that certain variables measuring how quickly the verification cycle begins are strong predictors of later stages in the verification cycle.

1 Introduction

Sproxil, Inc. helps consumers determine whether products they purchase are genuine or counterfeit. The company produces labels containing unique PINs (personal identification numbers), which are placed on products by manufacturers. Consumers can then send these codes (e.g. via text message) and receive a response from Sproxil such as "genuine," "fake," or "used."

The service that Sproxil provides helps to protect consumers from purchasing counterfeit products. There is a significant need for this service, particularly in emerging markets where counterfeit drugs are common and anti-counterfeiting measures are difficult to implement. Incidentally, Sproxil's data also provides insight into these emerging markets where little is known about the manner in which products sell. Our goal as defined by Sproxil was to use this data to produce a model for the shelf lives of products.

This work was completed as a semester-long project in a mathematical modeling course at Kenyon College in Gambier, Ohio, as part of the PIC Math program (Preparation for Industrial Careers in Mathematical Sciences[1]).

2 Research Problem

The data we received from Sproxil consist of 330 CSV (comma-separated value) files, each of which corresponds to a single product. We received data strictly from products verified in Nigeria and only those verifications that generated a "genuine" response. Each row within the files corresponds to a single unit of the product that was verified as genuine. Note that all client and product identification information was encrypted.

The columns contain data including the PIN generation date (the month, day, year, and time that the PIN was produced), the verification transaction date (the month, day, year, and time that the PIN was verified by the user), the product and client IDs (unique numbers

[1]Support for this Mathematical Association of America (MAA) and Society for Industrial and Applied Mathematics (SIAM) program is provided by the National Science Foundation (NSF grant DMS-1345499).

corresponding to the particular product and the client producing that product), and broad and specific industry categories.[2]

We used the verification transaction dates and PIN generation dates to create a new column of data containing the PIN shelf life. This is the difference, in days, between the PIN generation date and the verification date. We computed the shelf lives in Excel, but they can also be computed using the statistical computing language R [4], which we used for most of our later analysis.

Sproxil asked us to plot and model PIN shelf lives as well as PIN shelf lives grouped by PIN generation batch. In order to group shelf lives by PIN generation batch, we first converted the PIN generation date (which is formatted as a date and time) into a date only. We then separated each CSV file (which corresponds to a single product) into multiple files, one for each PIN generation date. Sproxil was interested in whether certain batches of PINs had different shelf life distributions than the rest. They were also interested in seasonal effects and the effect of the industry category on PIN shelf life.

3 The Gamma Distribution and Maximum Likelihood Estimation

3.1 The Gamma Distribution

When we plotted histograms of PIN shelf lives for a PIN generation batch, we found that they often followed a distribution which resembled a gamma distribution. We fit gamma distributions to our data, and we found by visual inspection that the gamma distribution often provides a good model for the distributions of the shelf lives.

The gamma distribution has two parameters: the shape parameter α and the scale parameter β, both greater than zero, and the distribution has the density function:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \text{ for } x > 0, \text{ where } \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx.$$

3.2 Maximum Likelihood Estimation

Maximum likelihood estimation (or MLE) is a method used to approximate the parameters of a distribution which best fits a given set of observations. We used an algorithm presented in [2] to write an R-script that computes the maximum-likelihood estimates of both parameters of a gamma distribution.

A likelihood function gives, roughly speaking, the probability of observing the data $x_1, \ldots, x_n$ given a probability model. The goal of maximum likelihood estimation is to find the parameters of the probability distribution which maximize the likelihood function. This maximizes the "agreement" of the observed data with the proposed probability model. If we model our data with a gamma distribution with shape parameter α and scale parameter

[2]Some industry categories are not known or not applicable. Examples of broad categories are "Anti-Infective" or "Cosmetics," and examples of specific categories are "Anti-Malarial" or "Body Cream."

β, the likelihood function is:

$$L(\alpha, \beta | x_1, \ldots, x_n) = \prod_{i=1}^{n} \left(\frac{x_i^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \exp\left(-\frac{x_i}{\beta}\right) \right) = \left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \right)^n \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x_i \right),$$

where $x_1, \ldots, x_n$ are the observed data. The method of maximum likelihood estimation says that the best model is the one with the parameters $\hat{\alpha}$ and $\hat{\beta}$ that maximize the likelihood function ($\hat{\alpha}$ and $\hat{\beta}$ denote the estimators of α and β, respectively). A more detailed explanation of the mathematics can be found in Appendix A.

3.3 Modeling PIN Shelf Life with Gamma Distributions

We used the method of maximum likelihood estimation to fit gamma distributions to the PIN shelf life data for each PIN generation batch. Often the minimum shelf life is quite large (several hundred days). When this is the case, the fitted gamma distribution overestimates the shelf life distributions before the minimum shelf life because the fitted model predicts that verifications begin before they actually do. We can see an example of this in Figure 1 where the minimum shelf life is 359 days.

In order to account for this, before fitting a distribution we first created a vector of "min-shifted" shelf lives. These are obtained by subtracting from the raw shelf lives the minimum shelf life among all items in the PIN generation batch. In this way, the distribution of min-shifted shelf lives begins at zero as does the gamma distribution.[3] The result of fitting a gamma distribution to the min-shifted shelf life is shown in Figure 2.

We assessed the fit of these gamma distributions by inspecting the histograms. We noticed that sometimes near the beginning of the verification cycle, there is a period of a few months in which no verifications occur. When this is the case, the gamma distribution still overestimates the shelf life distribution at the beginning even though the shelf lives are min-shifted. We can see an example of this in Figure 3. When there are long periods in which no verifications occur, the minimum shelf life is not a good measure of when the verification cycle truly begins.

We decided that it would be better to model the PIN shelf life *after* verifications begin to occur consistently. To make this change, we successively removed the smallest data point until the first and tenth data points were within 5 days of one another. This means that ten verifications occurred within five days. We then shifted the shelf lives again by the new minimum and re-fit a distribution. We called this new data the truncated shelf life, and we called the verification date corresponding to the smallest truncated shelf life the minimum truncated verification date. This date is a good indicator of when verifications begin to occur consistently. After removing outliers through this truncation process, the maximum likelihood gamma distribution appears to have a better fit, as shown in Figure 4.

[3]Note that the R-script that performs MLE takes the logarithm of each data point. Since the smallest min-shifted shelf life is zero and the logarithmic function is undefined at zero, we must truncate any zeros before passing our data to the function.

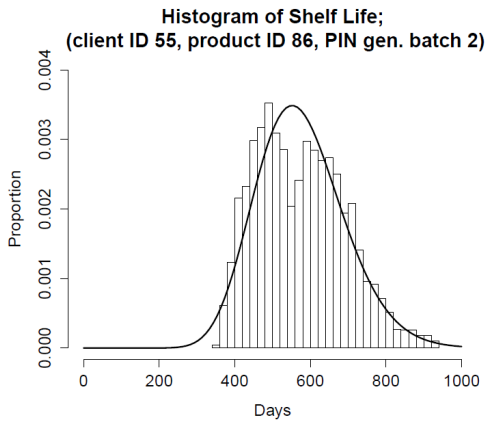

Figure 1: If the minimum shelf life is large, then the gamma distribution does not match the shelf life distribution well to the left of the minimum shelf life.

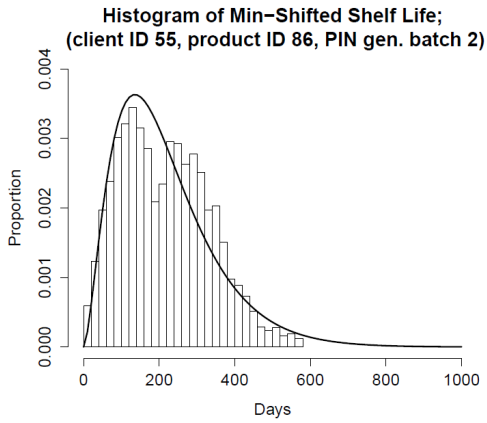

Figure 2: Fitting a gamma distribution to the min-shifted shelf lives generally gives a better fit than fitting a gamma distribution to the raw shelf lives.

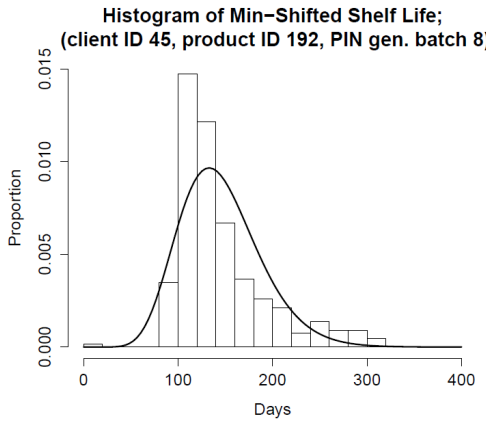

Figure 3: After the first verification, there is a period of 83 days in which no verifications occur.

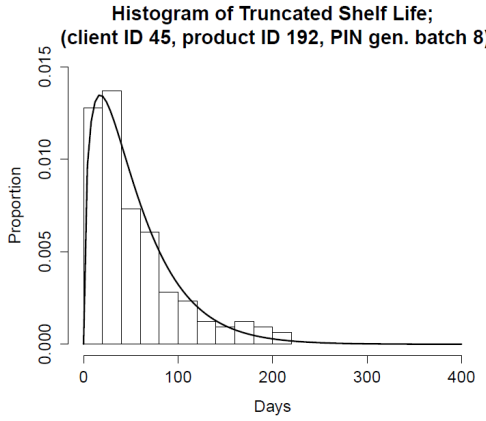

Figure 4: After removing the outlier, the maximum likelihood gamma distribution appears to provide a better fit for the shelf-life data.

3.4 Database of PIN Batch Variables

After implementing maximum likelihood estimation to approximate shape and scale parameters for each PIN generation batch, we compiled a database of our results. Each entry in the database contains information about one PIN generation batch. Our database includes statistics that were already known about a batch, such as batch size, as well as information generated from our modeling procedures. We used these characteristics to identify similar and dissimilar batches through cluster analysis and linear regression. Some of the variables we examined are listed in Table 1.

shape	The shape parameter of the gamma distribution fitted to a PIN batch
scale	The scale parameter of the gamma distribution fitted to a PIN batch
actual.50th *	The actual 50^{th} percentile of shelf life, i.e. the time in days by which 50% of PINs had been verified
estimated.50th *	The estimated 50^{th} percentile of shelf life, which is calculated by integrating the fitted gamma distribution function
diff.50 *	Actual 50^{th} percentile subtracted from the estimate
n.obs	Size of a PIN batch (prior to removal of outliers)
min.shelf.life	Minimum shelf life in a dataset
min.trunc.shelf.life	Minimum shelf life once outliers are excluded
min.ver.date	Date at which item with minimum shelf life was verified
min.trunc.ver.date	Date at which item with minimum truncated shelf life was verified
T_i **	The time (in days) after the min.trunc.ver.date by which $1000i$ units have been verified, where $i = 1, 2, 3, 4, 5$.

Table 1: Variables included in the database of PIN generation batches.

We also retrieved the season, year, and month of the PIN generation date, minimum verification date, and minimum truncated verification date.

The variables diff.50 and diff.90, which represent the respective errors of 50^{th} and 90^{th} percentile estimates, can be used as a preliminary assessment of the gamma distribution's fit to the shelf lives. The mean of diff.50 was -9.86 days with a standard deviation of 18.65 days, while diff.90 had a mean of 12.46 days with a standard deviation of 28.06 days. For reference, the average of actual.50th is 153.22 days, and the average of actual.90th is 274.83 days. Both percentile estimations tend to be quite accurate, approximating their corresponding actual percentiles within two weeks.

4 Cluster Analysis

4.1 Theory

Cluster analysis is an exploratory data analysis technique that groups numerical data into clusters of similar data points using a chosen distance metric. The clustering solution is found by one of a number of clustering algorithms that aim to have each data point be as similar as possible to other points in its assigned cluster and as different as possible from points in other clusters. Because the clustering algorithms utilize the distances between data points, one begins by generating a matrix M of the distances between every pair of points with respect to the chosen metric. While there are many different clustering techniques, the methods of clustering we focused on were k-means and fuzzy clustering.

As described in [1], k-means clustering algorithms take as input the distance matrix M described above, a positive integer k designating the number of clusters to create, and k initial cluster centers. The cluster centers will be the averages of the points in their cluster. One may specify initial cluster centers; however, many algorithms allow for starting with random cluster centers. The clustering algorithm then uses this information to partition the data set into the k "best" clusters for the data set. The "best" clustering solution is defined to be that which minimizes the within group sum of squares or WGSS. Essentially, the WGSS is the sum the squared Euclidean distance from each point to the average of its cluster. If our data has q parameters or coordinates and we want to partition it into k clusters denoted $C_1, \ldots, C_k$, we can denote the l^{th} cluster center as $\bar{x}^{(l)}$. Then the WGSS can be computed with the following formula:

$$\text{WGSS} = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \in C_l} (x_{ij} - \bar{x}_j^{(l)})^2.$$

A k-means clustering algorithm finds the clustering solution that minimizes the above expression by following these steps:

1. Assign each data point to the cluster whose initial center is closest.
2. Recalculate cluster centers to be the average of the points now in the cluster.
3. Calculate the change in the within group sum of squares that would result from moving a data point to a different cluster. Make those changes that would decrease the WGSS.
4. Repeat steps 2 and 3 until no change improves the WGSS.

In order to determine the number of clusters to use in k-means clustering, we plot the WGSS for the k-means solution against the number of clusters. The WGSS necessarily decreases as the number of clusters increases, but adding more clusters is not always useful. To determine an appropriate number of clusters, we look for a particularly sharp drop in the WGSS. For instance, if the WGSS decreases substantially when the number of clusters is increased from two to three but only slightly when the number of clusters is increased from three to four, then we would choose to use three clusters. More information on choosing the number of clusters can be found in [1].

Fuzzy clustering is very similar to k-means clustering. However, rather than considering a data point's membership in a particular cluster to be a binary state, the algorithm calculates the probability a data point belongs to each cluster. This process can still return a non-fuzzy or "crisp" clustering solution if we assign each data point to the cluster to which it is most likely to belong. When we ran k-means and fuzzy clustering on our data, we found that fuzzy clustering often gave more intuitive solutions.[4]

4.2 Implementation

We used cluster analysis on Sproxil's data by taking combinations of numerical data from the spreadsheet described in Section 3.4, computing the distance matrix for this data using the Euclidean distance metric, running k-means and fuzzy clustering on this matrix, and comparing the resulting clustering solutions to categorical variables. Before computing the distance matrix, we scaled the numerical data if different variables had significantly different ranges.

Sproxil expressed to us that they were interested in knowing whether categorical variables like product type and season of PIN generation affected a product's shelf life curve. We clustered on many combinations of numerical data, such as the shape and scale parameters of fitted gamma distributions, the actual 50^{th} and 90^{th} percentiles, the difference between the actual and estimated 50^{th} and 90^{th} percentiles, and the mean and standard deviation of the fitted gamma distributions. However, whenever we compared the results to one of the categorical variables mentioned, we did not find a strong correlation. For example, the following table is the result of comparing the two-cluster solution for the shape and scale parameters of the fitted gamma distributions to whether a product is pharmaceutical or non-pharmaceutical.

	Pharma	Non-Pharma
Cluster 1	116	33
Cluster 2	182	29

Table 2: Comparing cluster assignments to a categorical variable.

Note that the pharmaceutical and non-pharmaceutical products are mixed relatively evenly between the clusters. From this, we conclude that whether a product is pharmaceutical or non-pharmaceutical does not have a strong effect on the parameters of that product's fitted gamma distribution. Since we got similar results from every attempt to compare clustered data to type of product and season of PIN generation date, we concluded that these variables appear not to affect verification cycles as much as one might expect.

We did, however, find an interesting relationship using cluster analysis. We performed fuzzy cluster analysis using the variables T_1 through T_5 to produce three clusters of PIN generation batches. The results are shown in Figure 5.

[4]K-means clustering often grouped all outliers together.

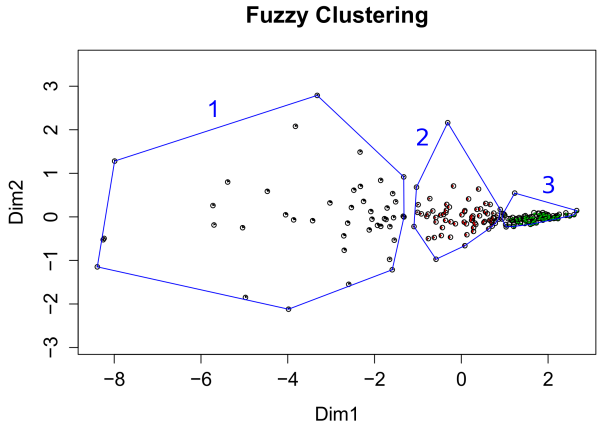

Figure 5: The result of running fuzzy clustering on the PIN generation batches using the variables T_1 through T_5.[5]

We compared the cluster assignments to the batch sizes to see if batch size has an effect on cluster assignment. The results are summarized in Table 3. We found that Cluster 1 tends to contain the smallest PIN generation batches, Cluster 2 tends to contain slightly larger PIN generation batches, and Cluster 3 tends to contain the largest PIN generation batches. In particular, almost all PIN generation batches of size 90,000 or greater were assigned to Cluster 3.

	Batch Size (in 10,000s)										
	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-36	Mean Batch Size
Cluster 1	23	14	3	2	1	1	1	0	0	0	14,396
Cluster 2	23	15	15	6	0	2	2	1	1	2	23,636
Cluster 3	3	10	9	10	12	5	5	2	7	21	66,915

Table 3: Comparing PIN generation batch size to cluster assignment.

The mean values of the variables T_1 through T_5 among batches in each cluster are largest for Cluster 1 and smallest for Cluster 3. This means, on average, batches in Cluster 1 take the longest to reach milestones in the verification cycle, while batches in Cluster 3 take the shortest.

The relationship between batch size and cluster assignment suggests that the size of a PIN generation batch could be used to predict a product's verification cycle. We explored this with multiple regression in Section 5.3.

[5]For this graph, we used multidimensional scaling (MDS) to place these data points in space with meaningful relative positions. An MDS algorithm uses a distance matrix to create coordinates for each item in a dataset. This allows the Euclidean distance between plotted points to approximate the distances between the original data points along multiple dimensions. For our data, we used T_1 through T_5 to generate the coordinates Dim1 and Dim2. More about MDS can be found in [5].

5 Regression Analysis

5.1 Linear Regression

Linear regression is a technique for exploring potential linear relationships between variables. Using this technique, we can construct equations that can be used to make predictions about the behavior of PIN shelf lives. We implemented this in two ways: we used simple linear regression, which uses one variable to make predictions, and we used multiple regression, which uses many variables to make predictions.

In simple linear regression, we get the following model for predicting y using x:

$$y = \beta_0 + \beta_1 x + \epsilon.$$

In this model, β_0 is the intercept, β_1 is the effect that the predictor x has on y, and ϵ is random deviation from the regression line. When constructing this model, we get the following statistics: R^2, which tells us the percent of the variation in y that is explained by x; and t-values and p-values, which indicate if the coefficients are significantly different from zero.

In multiple regression, we get models with more terms:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon.$$

In this model, p is the number of predictors, β_0 is the intercept, β_i is the effect that the predictor x_i has on y, and ϵ is the random deviation in the model. When using multiple regression, we look at adjusted R^2 (R^2_{adj}) values instead of R^2 values, because the R^2_{adj} values take into account the number of predictors and penalize models which use too many predictors. This model also gives us t-values and p-values with which we may determine whether predictors are significant.

We used stepwise regression with forward selection to generate a multiple regression model in which every predictor is significant. This procedure, which we automated in R, begins with a dependent variable y and a collection of possible predictors. We select the predictor that has the strongest linear relationship with y and we create a simple linear regression between the two variables. Next, we select the predictor out of the remaining variables that would bring about the greatest increase in our first model's predictive value. We create a new multiple regression model using both predictors; then, we check the significance of each predictor. If both are still significant, we choose a third predictor. This process continues until there remains no predictor that could significantly improve the model.

5.2 Linear Regression Models to predict actual.90th

5.2.1 Simple Linear Regression Model

There is a strong linear relationship between actual 50^{th} percentile and actual 90^{th} percentile. The least-squares regression is:

$$\widehat{\text{actual.90th}} = 78.9088 + 1.407(\text{actual.50th}).$$

Note that the units in this model are days; that is, this model predicts the number of days until 90 percent of a product has been sold using the number of days it took for half of the product to sell. We note here that a good predictive model for the 90^{th} percentile could potentially be useful to a manufacturer that wishes to predict appropriate timing for restocking a product.

The coefficient estimates, standard errors, t-values, and p-values (listed as Pr(>|t|)) are summarized in in Table 4.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	78.9088	9.0299	8.74	0.0000
actual.50th	1.4070	0.0444	31.71	0.0000

Table 4: Linear regression to predict actual.90th from actual.50th.

Both the intercept and the actual 50^{th} percentile are significant in the model since their p-values are approximately zero. This model has an $R^2 = 78.78\%$, meaning that 78.78% of the variation in actual 90^{th} percentile can be explained by the variation in actual 50^{th} percentile.

With linear regression, we are also able to look at subsets of the data based on parameters including, but not limited to, industry broad, industry specific, and season of PIN generation. We used this as another way to look for clusters of similar products. Sproxil was interested in determining whether pharmaceutical products and non-pharmaceutical products differ significantly in their shelf lives. We fit separate regression lines to the pharmaceutical data and the non-pharmaceutical data to investigate this.

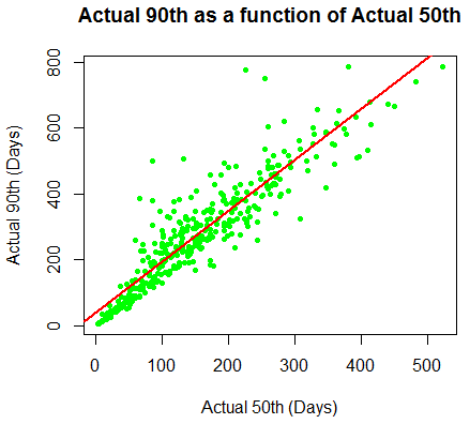

Figure 6: This is the relationship between actual 50^{th} and actual 90^{th} percentiles with a regression line.

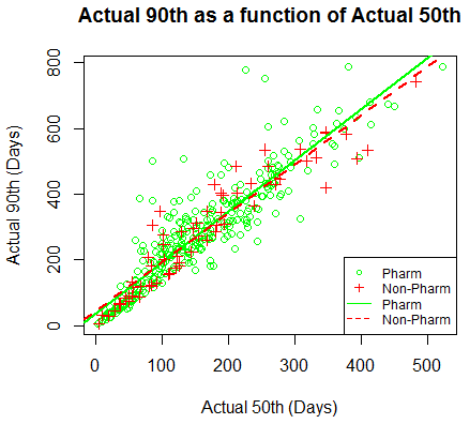

Figure 7: This is the same relationship as in Figure 6 with the data colored by whether the products are pharmaceuticals or non-pharmaceuticals.

In Figure 6, we see very little difference in the regression lines between the pharmaceuticals and non-pharmaceuticals. This is confirmed by the 95% confidence intervals for both

the slopes and intercepts of the models for pharmaceutical and non-pharmaceuticals, since the confidence intervals for the slopes and intercepts overlap.

We should also note that when conducting regression analysis on the shelf life percentiles we chose to disregard the data from 2014. This decision was made because we noticed that the data from 2014 exhibited different behavior from the rest of the data. When we observe the average batch size for products produced in each year in Table 5, we see that there is a significant drop in 2014.

	2010	2011	2012	2013	2014
Mean Batch Size	35356.00	27938.34	24176.74	20856.22	9440.57

Table 5: Mean batch size by year of production.

This led us to believe that many items from the 2014 batches had not been verified by the time we received data from Sproxil. If this is the case, then the 50^{th} and 90^{th} percentiles we computed are not the true 50^{th} and 90^{th} percentiles of the complete data set.

5.2.2 Multiple Regression

Having identified 50^{th} shelf life percentile as a strong predictor of 90^{th} shelf life percentile, we conducted stepwise regression to identify other predictors that could be added to this model. We initially experimented with incorporating categorical predictors in the form of indicator variables; however, we chose to restrict our search to numerical predictors in order to identify variables that had linear relationships with the 90^{th} percentile of shelf life.

We found that the five best predictors for the actual 90^{th} percentile shelf life data are the actual 50^{th} percentile, the actual 30^{th} percentile, minimum truncated verification date, batch size, and T_1. These variables were all significant at the $\alpha = 0.05$ significance level. These variables are used in the following predictive model:

$$\begin{aligned}\widehat{\text{actual.90th}} = 1489.3227 &+ 1.739(\text{actual.50th}) - 0.5116(\text{actual.30th}) \\ &-0.1236(T_1) - 0.00024(\text{n.obs}) - 0.09364(\text{min.trunc.ver.date}).\end{aligned}$$

In this model, the min.trunc.ver.date is measured as the number in days since January 1, 1970. The coefficients of this model are summarized in Table 6. The p-values are all less than 0.05, indicating that the predictors are significant.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	15843	191.4	8.279	0.00000
actual.50th	1.739	0.1328	13.093	0.00000
actual.30th	-0.5116	0.1738	-2.944	0.00353
T_1	-0.1236	0.05004	-2.471	0.01412
n.obs	-0.00024	0.000	-2.382	0.01792
min.trunc.ver.date	-0.09364	0.0119	-7.868	0.00000

Table 6: Multiple regression to predict actual.90th from the five best predictors.

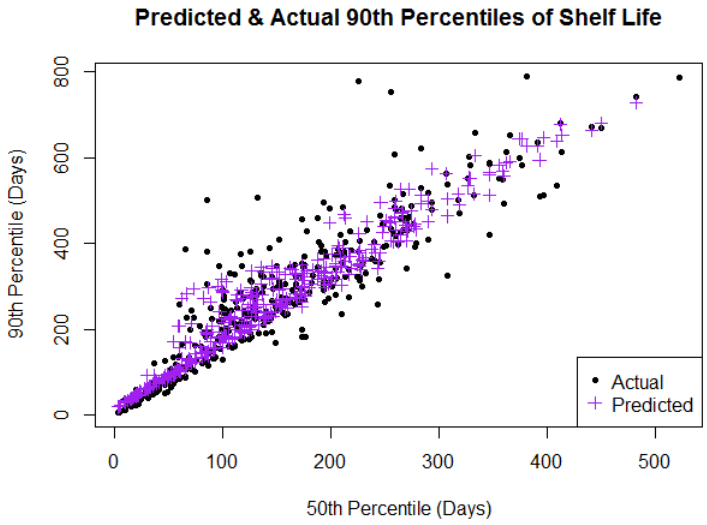

Figure 8: The purple plusses are the fitted values for actual.90th. We can see that the multiple regression model gives us more flexibility than a simple linear model.

This model has an R^2 value of 83.25%, meaning that 83.25% of the variation in actual 90^{th} percentile can be explained by the variations in actual 30^{th} and 50^{th} percentiles, batch size, T_1, and the minimum truncated verification date. This indicates that our model is strongly useful for predicting the actual 90^{th} percentile. The standard error of the residuals from this model is 62.02 days.

5.3 Linear Regression Models to Predict the T_i's

We also employed linear regression and multiple regression to establish predictive relationships involving the variables T_1 through T_5. Recall that these are the times it takes for 1000 through 5000 items to be verified, respectively. These offer another way to measure how quickly units are verified besides computing percentiles of the shelf life data. The advantage of using these variables over the percentiles is that the T_i's are still accurate even if we do not have the most recent verification data. As long as we have data on the first 5000 verifications, we can compute these variables to understand how verifications occur over time.

The first model predicts T_1 based on the batch size because our cluster analysis results suggested a relationship between batch size and how quickly verifications occur. The R^2 value is 13.61%, which indicates that it is difficult to accurately predict T_1 based on batch size alone. The other four models predict T_2, T_3, T_4, and T_5 based on the values of the previous T_i's. The simple regression model predicting T_2 based on T_1 has an R^2 value of 93.77%, and the multiple regression models for predicting T_3 through T_5 all have adjusted R^2 values above 97%, which means that knowledge of certain milestones in the verification cycle allows us to accurately predict future milestones.

Although batch size is a factor in how quickly verifications occur, it is difficult to predict what a verification cycle will look like before it begins. However, once the cycle begins,

data on how quickly verifications initially occur can be used to predict how quickly future verifications will occur. Like the 90^{th} percentile models, this information appears to be potentially useful for manufacturers working with Sproxil.

6 Conclusion

In this project, we used multiple techniques to model the shelf-life distribution of PIN batches. We aimed to identify qualities about a PIN batch that could be used to predict its shelf-life distribution.

We found that before a batch of a product begins to sell, few predetermined variables appear to predict the behavior of its shelf-life distribution. In particular, our cluster analysis and linear regression approaches did not find relationships between product type or season of PIN generation and shelf life, though further research may indicate that such relationships do exist. Among variables that are known before the verification cycle begins, batch size appeared to be a variable that influences shelf life. We also found that variables depicting a PIN batch's initial verification cycle pattern could strongly predict the cycle's overall duration. For example, 30^{th} and 50^{th} percentiles of shelf life had strong positive correlations with the 90^{th} percentile. Additionally, the time it takes for specific numbers of PINs to be verified, as measured by T_1 through T_5, could be used to predict the time needed for greater numbers of items to be verified.

The rate at which a batch begins to be verified has a strong relationship with the time needed for a batch's verification cycle to resolve. Given these findings, it is feasible to use data about initial verification patterns for a PIN generation batch in order to predict the duration of its cycle. Our models may be tested by implementing them to track verification cycles of currently circulating PIN batches. Since the strongest predictive relationships we found involved T_1 through T_5, Sproxil might investigate whether earlier predictors, such as milestones on the order of hundreds of verifications, have the same predictive power.

7 Acknowledgements

The authors thank the Preparation for Industrial Careers in Mathematical Sciences program, the Mathematical Association of America, the Society for Industrial and Applied Mathematics, and the National Science Foundation (NSF grant DMS-1345499). The course in which this work was completed was made possible by their support. In addition, the authors gratefully acknowledge the contributions that made the collaboration with Sproxil possible. In particular, Jennifer Campos, Director of Service Innovation at Sproxil, served as a liaison and provided valuable feedback throughout the collaboration. Ashifi Gogo, Chief Executive Officer of Sproxil, was instrumental in facilitating an initial connection between Sproxil and Kenyon College and was personally involved in regular communication between Sproxil and the authors.

A Appendix: Maximum Likelihood Estimation

We will describe the procedure presented in [2] for finding the values of the parameters $\hat{\alpha}$ and $\hat{\beta}$ which maximize the likelihood function (introducd in Section 3.2). These can be found most easily by maximizing the log likelihood function, which is equivalent to maximizing the likelihood function. The log likelihood function simplifies to

$$\log(L(\alpha, \beta | x_1, \ldots, x_n)) = n(\alpha - 1)\overline{\log x} - n \log \Gamma(\alpha) - n\alpha \log \beta - \frac{n}{\beta}\overline{x} \tag{1}$$

where $\overline{\log x}$ denotes the mean of the logarithm of the x_i's and $\overline{x}$ denotes the mean of the x_i's.

To maximize the log likelihood function with respect to β, we differentiate (1) with respect to β and set the derivative equal to zero. There is a critical point $\hat{\beta} = \overline{x}/\alpha$ (where α is the true parameter value, not the estimator $\hat{\alpha}$); this is the value of β which maximizes (1). We substitute this value back into (1) and simplify to get

$$\log(L(\alpha, \beta | x_1, \ldots, x_n)) = n(\alpha - 1)\overline{\log x} - n \log \Gamma(\alpha) - n\alpha \log \overline{x} + n\alpha \log \alpha - n\alpha. \tag{2}$$

Now we must maximize (2) with respect to α. We cannot directly solve for the critical value of α, so we must approximate it. A fast algorithm for iteratively approximating $\hat{\alpha}$ is presented in [2]. This makes use of a generalized Newton's method.

We begin with an initial guess α_0 for the value of $\hat{\alpha}$ that maximizes (2). This guess is improved by approximating the log likelihood by a function of the form

$$g(\alpha) = c_0 + c_1\alpha + c_2 \log(\alpha)$$

and finding the value of α which maximizes $g(\alpha)$. This is taken as our new guess for $\hat{\alpha}$. The values of the constants c_0, c_1, and c_2 are chosen so that (2) and $g(\alpha)$ have the same function value and first and second derivative at α_0. If α_{old} is our initial guess, then our updated guess α_{new} obtained by maximizing $g(\alpha)$ satisfies the following:

$$\frac{1}{\alpha_{\text{new}}} = \frac{1}{\alpha_{\text{old}}} + \frac{\overline{\log x} - \log \overline{x} + \log \alpha_{\text{old}} - \Psi(\alpha_{\text{old}})}{\alpha_{\text{old}}^2(1/\alpha_{\text{old}} - \Psi'(\alpha_{\text{old}}))} \tag{3}$$

where $\Psi(x)$ is the digamma function, $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$, and $\Psi'(x)$ is the trigamma function, $\Psi'(x) = \frac{d}{dx}\Psi(x)$ [2].

According to [2], this method converges in about four iterations. We wrote an R-script to iterate this process until we get successive α-values that are within 0.001 of each other. For the initial guess α_0 we use

$$\hat{\alpha} \approx \alpha_0 = \frac{0.5}{\log \overline{x} - \overline{\log x}} \tag{4}$$

which is a good initial approximation obtained from Stirling's approximation for the gamma function [2].

References

[1] B. EVERITT and T. HOTHORN, An Introduction to Applied Multivariate Analysis with R: Use R!, Springer Science+Business Media, LLC, New York, NY, 2011.

[2] T. P. MINKA, Estimating a Gamma distribution, http://research.microsoft.com/en-us/um/people/minka/papers/minka-gamma.pdf, 2002.

[3] J. OKSANEN, Cluster Analysis: Tutorial with R, http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf, 2014.

[4] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2013.

[5] F. WICKELMAIER, An Introduction to MDS, https://homepage.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf, 2003.