Optimizing the Creditworthiness Threshold of a Bivariate Distribution

Malachi Jones[1*], Victoria Knutson[2*], Benjamin Stockton[3*], and Advisor: Hui Gong[4]

[1]Colgate University
[2]St. Olaf College
[3]University of Wisconsin-Whitewater
[4]Valparaiso University

July 30, 2018

Abstract: Financial institutions must evaluate credit applications when deciding to issue credit. Creditworthiness varies amongst the applicants. Creditors must decide which applications to accept in order to maximize profit. For this paper, we assume applicants are divided into Good and Bad populations. We found optimal threshold values that maximized the creditor's profit under varying assumptions of Normal, χ^2, and Γ- distributions. To do so, we optimized the profit function with respect to the threshold value and we ran simulations to find the threshold value that maximizes the profit.

Keywords: Distribution, creditworthiness, profit function, threshold.

---

* Authors contributed equally to the paper.

1 Introduction

Financial institutions are constantly looking at how to maximize profit through their decisions to either accept or reject a credit application or reapplication. To do this, the creditor will look at the applicant's creditworthiness, or the measure of likelihood that the applicant will not default on their loans. To determine whether to accept or reject based on this value, the creditor requires an optimal threshold, above which the lender will accept the application.

The optimization of a Profit Function to determine the value of this optimal threshold allows for the derivation of a flexible formula that can be easily altered for changes in cost, revenue, and assumptions of the population. By considering a Profit Function that has elements of cost and revenue, the creditor is able to find an optimal threshold that both maximizes profit and minimizes the risk of the acceptance of a bad status applicant. We consider four classification probabilities based on the good or bad status of the applicant (whether they will in actuality repay their loans) and the decision of the creditor which will be correct or incorrect due to the applicant's status.

Throughout this paper, we have evaluated the optimal thresholds that maximize the profit function for different assumptions of the two populations: good and bad status applicants. This analysis of optimal thresholds is built off the research done by Chen et. al [1], which considers the population distribution model of Normal-Normal with the σ values of both populations equal to 1. In this paper, we considered varying assumptions of Normal, χ^2, and Γ distributions. We present the methods we used to find the optimal threshold in more depth in Section 2. Our findings of the optimal threshold under Normal-Normal distribution assumptions of the populations are in Section 3 and our findings of the optimal threshold under Gamma-Gamma distribution assumptions of the populations are in Section 4. Finally, we explore a few examples in Section 5 with the conclusion of the paper in Section 6.

2 Methods

We considered customers reapplying for credit. The general setup for this paper comes from a paper done by Chen et. al [1]. Let D be a binary variable that reflects the status of a customer. The value of D can be either Good(G) or Bad(B). In this paper, it is assumed a Good customer will pay back the loan and a Bad customer will not pay back the loan. Each customer is assigned a score, y^*, that represents the creditworthiness of the customer. This creditworthiness is an attempt to measure the likelihood of the customer paying back the loan. Then a threshold value, k is chosen for the creditworthiness scores that decides whether a customer is accepted or rejected. For a creditworthiness score, y^*, and a threshold value, k, let Y_k be the binary decision variable with the values $A =$ Acceptance and $R =$ Rejection. This could also be denoted:

		Decision (Y_k)	
		Accept (A)	Reject (R)
Status	Good (G)	True Positive (TP)	False Negative (FN)
(D)	Bad (B)	False Positive (FP)	True Negative (TN)

Table 1: Conditional Probabilities for Status and Decision

$$Y_k = \begin{cases} A(Accept\ application) & if\ y^* > k, \\ R(Reject\ application) & if\ y^* \leq k \end{cases}$$

Next, we will define the four conditional probabilities (shown in Table 1) for our two binary variables, D and Y_k. These conditional probabilities are the True Positive (TP) and the True Negative (TN) and their compliments, the False Negative (FN) and the False Positive (FP). True Positive is the probability of correctly accepting an application of a Good status applicant, and True Negative is the probability of correctly rejecting an application of a Bad status applicant. Likewise, False Positive is the probability of incorrectly accepting an application of a Bad status applicant, and False Negative is the probability of incorrectly rejecting an application of a Good status applicant. Additionally, $TP + FN = 1$ and $FP + TN = 1$. The four conditional probabilities for a specific threshold can be written as follows:

- $TP_k = P(y^* > k|D = G) = P(Y = G|D = G)$,
- $TN_k = P(y^* \leq k|D = B) = P(Y = B|D = B)$,
- $FN_k = P(y^* \leq k|D = G) = P(Y = B|D = G)$,
- $FP_k = P(y^* > k|D = B) = P(Y = G|D = B)$.

Each of these conditional probabilities are associated with either a cost or a revenue. The revenues and costs of these probabilities have realistic interpretations. The revenue of a True Positive, denoted r_G is the interest earned from the repayment of the loan. The revenue of a True Negative is the application fees earned on the applications that were correctly rejected. This is represented by r_B. The costs of a False Positive and False Negative are symbolized by c_B and c_G respectively. The c_B value is the cost of the unpaid loan while c_G is the processing cost of the application along with the opportunity cost of the interest that could have been made had the application been accepted. We assume that all of these constants are greater than zero. These revenues and costs can be shown in a payoff matrix, such as Table 2.

Now, we will combine the four conditional probabilities with their associated costs and revenues to define the Profit Function. Let $R(k)$ be the Profit

		Decision (Y)	
		Accept (G)	Reject (B)
Status	Good (G)	r_G	c_G
(D)	Bad (B)	c_B	r_B

Table 2: Payoff Matrix

Function defined as

$$R(k) = (P_G * r_G * TP_k + P_B * r_B * TN_k) - (P_G * c_G * FN_k + P_B * c_B * FP_k)$$

where P_G is the proportion of the applicants who are of Good status and P_B is the proportion of those who are of Bad status with $P_G + P_B = 1$. $R(k)$ gives us the profit by taking the revenues minus the costs. The objective of our paper is to find the optimal threshold, k^* that maximizes the expected profit $R(k)$.

To maximize expected profit, we used two different methods: optimization of a multivariate function and simulations. We used both these methods for each of the different assumptions we made about the population distribution. First, we optimized the profit function by taking its derivative with respect to k and setting it equal to zero. By solving for k, we are able to get an extreme point. Using the Second Derivative Test, we were able to show that our extreme point was a maximum. Next, we used simulations to confirm the results we had derived. To do this, we generated a random sample of creditworthiness scores using the distribution we were analyzing at the time. Then we created a list of possible threshold values from the minimum creditworthiness score to the maximum creditworthiness score in our sample using small increments. By passing each of these through the Profit Function, we were able to approximate where the threshold should be set to maximize profit. By comparing this value to the one derived earlier, we could confirm that our threshold value is correct.

3 Normal Distribution

We begin by assuming a Normal-Normal model for population distributions. While it is not quite realistic that both distributions are distributed Normally, it is likely that there are clumps of the population centered around mediocre values of good and bad creditworthiness, so considering a Normal-Normal model is valid. And because analyzing a symmetric distribution such as the Normal distribution is easier in some ways, we begin with this population assumption.

Let $Y^* \sim N(\mu_G, \sigma_G^2)$ for the good population and $Y^* \sim N(\mu_B, \sigma_B^2)$ for the bad population. We assume that $\mu_G > \mu_B$ and that the good and bad population are independent.

$$f_G(y^*|\mu_G, \sigma_G^2) = \frac{1}{\sqrt{2\pi\sigma_G^2}} \exp\left[-\frac{(y^* - \mu_G)^2}{2\sigma_G^2}\right]$$

$$f_B(y^*|\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left[-\frac{(y^* - \mu_B)^2}{2\sigma_B^2}\right]$$

Under these assumptions, we have the following probabilities:

$TP_k = P(y^* > k|D = G) = 1 - F_G(k) = \int_k^\infty f_G(y^*)dy^*$

$FP_k = P(y^* > k|D = B) = 1 - F_B(k) = \int_k^\infty f_B(y^*)dy^*$

3.1 Case 1: $\sigma_G = \sigma_B$

We begin with the assumption that $\sigma_G = \sigma_B$ in order to make the computations easier, though this is not a very realistic assumption of the two populations. In the following two theorems, let k_0 be the optimal threshold value satisfying $R'(k_0) = 0$. Theorem 1 derives this k_0 value and Theorem 2 gives the second derivate test where k_0 proves to be a maximum.

Theorem 1: Under the assumption of Normal-Normal model and $\sigma_G = \sigma_B = \sigma$, k_0 can be shown to be unique and expressible in a closed-form expression involving P_G, P_B, c_G, c_B, r_G, r_B, μ_G, and μ_B.

$k_0 = \frac{\sigma^2 \ln C}{\mu_G - \mu_B} + \frac{\mu_G + \mu_B}{2}$
where
$C = \frac{P_B}{P_G} \frac{r_B + c_B}{r_G + c_G}$.

Proof: Given the profit function, $R(k)$:

$$\begin{aligned} R(k) &= P_G * r_G * TP_k + P_B * r_B * TN_k - (P_G * c_G * FN_k + P_B * c_B * FP_k) \\ &= P_G * (r_G + c_G) * TP_k - P_B * (r_B + c_B) * FP_k + (P_B * r_B - P_G * c_G) \end{aligned}$$

Then, under the Normal-Normal model assumptions and $\sigma_G = \sigma_B = \sigma$, we have
$TP_k = \int_k^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^* - \mu_G)^2}{2\sigma^2}\right] dy^*$
and
$FP_k = \int_k^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^* - \mu_B)^2}{2\sigma^2}\right] dy^*$

Taking the first derivative of $R(k)$ with respect to k, we have

$$\begin{aligned}
R'(k) &= P_B * (r_B + c_B) * f_B(k) - P_G * (r_G + c_G) * f_G(k) \\
&= P_B * (r_B + c_B) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k-\mu_B)^2}{2\sigma^2}\right] - P_G * (r_G + c_G) * \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k-\mu_G)^2}{2\sigma^2}\right] \\
&= P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)}\right)
\end{aligned}$$

where $C = \frac{P_B}{P_G}\frac{r_B+c_B}{r_G+c_G}$. Then $R'(k) = 0$ if and only if
$P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)}\right) = 0$

Since P_G, r_G, c_G, and $f_B(k)$ are all positive, the previous equation is equivalent to the following condition.

$$\begin{aligned}
C &= \frac{f_G(k_0)}{f_B(k_0)} \\
C &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k_0-\mu_G)^2}{2\sigma^2}\right]}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k_0-\mu_B)^2}{2\sigma^2}\right]} \\
C &= \exp\left[\left(-\frac{(k_0-\mu_G)^2}{2\sigma^2}\right) - \left(-\frac{(k_0-\mu_B)^2}{2\sigma^2}\right)\right] \\
\ln C &= \left(-\frac{(k_0-\mu_G)^2}{2\sigma^2}\right) - \left(-\frac{(k_0-\mu_B)^2}{2\sigma^2}\right) \\
2\sigma^2 \ln C &= (k_0-\mu_B)^2 - (k_0-\mu_G)^2 \\
2\sigma^2 \ln C &= k_0^2 - 2k_0\mu_B + \mu_B^2 - k_0^2 + 2k_0\mu_G - \mu_G^2 \\
2\sigma^2 \ln C &= 2k_0(\mu_G - \mu_B) + (\mu_B^2 - \mu_G^2) \\
2\sigma^2 \ln C + (\mu_G^2 - \mu_B^2) &= 2k_0(\mu_G - \mu_B) \\
k_0 &= \frac{\sigma^2 \ln C}{\mu_G - \mu_B} + \frac{\mu_G + \mu_B}{2} \qquad \blacksquare
\end{aligned}$$

Theorem 2: Assume a Normal-Normal distribution for the two populations with $N(\mu_G, \sigma^2)$ and $N(\mu_B, \sigma^2)$ with $\mu_G > \mu_B$. Let $C = \frac{P_B}{P_G}\frac{(r_B+c_B)}{(r_G+c_G)}$ and k_0 be defined as in Thm. 1. Then:

1. $R'(k) \geq 0$ when $k \leq k_0$.
2. $R'(k) \leq 0$ when $k \geq k_0$.
3. $R''(k) < 0$.

Proof: (1) and (2) can be easily proved by changing the equal sign in Thm. 1 to $\geq$ and $\leq$, respectively.
To prove (3), we take the second derivative of $R(k)$ with respect to k:

$$\frac{\partial^2 R(k)}{\partial k^2} = -\frac{(\mu_G - k)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k-\mu_G)^2} P_G(r_G+c_G) + \left[\frac{(\mu_B - k)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k-\mu_B)^2}\right] P_B(r_B+c_B)$$

Considering the local maximum for which $R''(k) < 0$, we look at:

$$\frac{(\mu_B - k)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k-\mu_B)^2} P_B(r_B + c_B) < \frac{(\mu_G - k)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k-\mu_G)^2} P_G(r_G + c_G)$$

$$\frac{\mu_B - k_0}{\mu_G - k_0} C < \exp\left[\frac{1}{2\sigma^2}((\mu_B - k_0)^2 - (\mu_G - k_0)^2)\right]$$

$$2\sigma^2(\ln(\mu_B - k_0) - \ln(\mu_G - k_0) + \ln C) < (\mu_B - k_0)^2 - (\mu_G - k_0)^2)$$

$$2\sigma^2 \ln(\mu_B - k_0) - 2\sigma^2 \ln(\mu_G - k_0) + 2\sigma^2 \ln C < \mu_B^2 - 2\mu_B k_0 - \mu_G^2 + 2\mu_G k_0$$

Substituting in $k_0 = \frac{\sigma^2 \ln C}{\mu_G - \mu_B} + \frac{\mu_G + \mu_B}{2}$ on the right hand side:

$$2\sigma^2 \ln(\mu_B - k_0) - 2\sigma^2 \ln(\mu_G - k_0) + 2\sigma^2 \ln C < \mu_B^2 - \frac{2\mu_G\sigma^2 \ln C}{\mu_G - \mu_B} - \mu_G\mu_B - \mu_B^2 - \mu_G^2 + \frac{2\mu_B\sigma^2 \ln C}{\mu_G - \mu_B} + \mu_G\mu_B + \mu_G^2$$

$$2\sigma^2 \ln(\mu_B - k_0) - 2\sigma^2 \ln(\mu_G - k_0) + 2\sigma^2 \ln C < \frac{2\sigma^2 \ln C(\mu_G - \mu_B)}{(\mu_G - \mu_B)}$$

$$2\sigma^2 \ln(\mu_B - k_0) - 2\sigma^2 \ln(\mu_G - k_0) < 0$$

$$\mu_B < \mu_G$$

Thus we get $R''(k_0) < 0$ if and only if $\mu_B < \mu_G$, which is an assumption and therefore $R''(k_0) < 0$.
Theorem 9 below follows immediately from Theorem 2. ■

3.2 Case 2: $\sigma_G \neq \sigma_B$

Next we consider the case in which $\sigma_G \neq \sigma_B$. This is a more realistic assumption under the Normal-Normal model, because as previous consultation on the paper by Chen et. al [1] and literature suggests that these populations are uniquely distributed and independent from each another. In the following two theorems, let k_0 be the optimal threshold value satisfying $R'(k_0) = 0$. Theorem 3 derives this k_0 value and Theorem 4 gives the second derivate test where k_0 proves to be a maximum.

Theorem 3: Under the assumption of Normal-Normal model and $\sigma_G \neq \sigma_B$. We then define $\sigma_B = \sigma$ and $\sigma_G = m\sigma$ where $m > 0$ and $m \neq 1$. Then k_0 can be shown to be unique and expressible in a closed-form expression involving P_G, P_B, c_G, c_B, r_G, r_B, σ, m, μ_G, and μ_B.

$k_0 = \frac{(m^2\mu_B-\mu_G)\pm\sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln Cm)}}{m^2-1}$
and
$C = \frac{P_B}{P_G}\frac{r_B+c_B}{r_G+c_G}$.

Proof: Given the profit function $R(k)$:

$$\begin{aligned} R(k) &= P_G * r_G * TP_k + P_B * r_B * TN_k - (P_G * c_G * FN_k + P_B * c_B * FP_k) \\ &= P_G * (r_G + c_G) * TP_k - P_B * (r_B + c_B) * FP_k + (P_B * r_B - P_G * c_G) \end{aligned}$$

Under the Normal-Normal model assumptions and $\sigma_G \neq \sigma_B$, we have
$TP_k = \int_k^\infty \frac{1}{\sqrt{2\pi m^2\sigma^2}} \exp\left[-\frac{(y^*-\mu_G)^2}{2m^2\sigma^2}\right] dy^*$
and
$FP_k = \int_k^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^*-\mu_B)^2}{2\sigma^2}\right] dy^*$

Take the first derivative of the Profit Function with respect to k, we have

$$\begin{aligned} R'(k) &= P_B * (r_B + c_B) * f_B(k) - P_G * (r_G + c_G) * f_G(k) \\ &= P_B * (r_B + c_B)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(k-\mu_B)^2}{2\sigma^2}\right] - P_G * (r_G + c_G) * \frac{1}{\sqrt{2\pi m^2\sigma^2}} \exp\left[-\frac{(k-\mu_G)^2}{2m^2\sigma^2}\right] \\ &= P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)}\right) \end{aligned}$$

where $C = \frac{P_B}{P_G} * \frac{r_B+c_B}{r_G+c_G}$. Then $R'(k) = 0$ if and only if
$P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)}\right) = 0$
Since P_G, r_G, c_G, and $f_B(k)$ are all positive, the previous equation is equivalent to the following condition.

$$
\begin{aligned}
C &= \frac{f_G(k_0)}{f_B(k_0)} \\
C &= \frac{\frac{1}{\sqrt{2\pi(m\sigma)^2}}\exp\left[-\frac{(k_0-\mu_G)^2}{2(m\sigma)^2}\right]}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(k_0-\mu_B)^2}{2\sigma^2}\right]} \\
C &= \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi m^2\sigma^2}}\exp\left[\frac{(k_0-\mu_B)^2}{2\sigma^2}-\frac{(k_0-\mu_G)^2}{2m^2\sigma^2}\right] \\
C &= \frac{1}{m}\exp\left[\frac{(k_0-\mu_B)^2}{2\sigma^2}-\frac{(k_0-\mu_G)^2}{2m^2\sigma^2}\right] \\
\ln(Cm) &= \frac{(k_0-\mu_B)^2}{2\sigma^2}-\frac{(k_0-\mu_G)^2}{2m^2\sigma^2} \\
2m^2\sigma^2\ln(Cm) &= m^2(k_0-\mu_B)^2-(k_0-\mu_G)^2 \\
2m^2\sigma^2\ln(Cm) &= m^2k_0^2-2m^2\mu_Bk_0+m^2\mu_B^2-k_0^2+2\mu_Gk_0-\mu_G^2 \\
2m^2\sigma^2\ln(Cm) &= (m^2-1)k_0^2+(2\mu_G-2m^2\mu_B)k_0+m^2\mu_B^2-\mu_G^2 \\
0 &= (m^2-1)k_0^2+2(\mu_G-m^2\mu_B)k_0+m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln(Cm)
\end{aligned}
$$

Note that when $m = 1$, the quadratic term goes to zero. In this case, $\sigma_G = \sigma_B$ which is in Section 3.1. Using the quadratic formula:

$$
\begin{aligned}
k_0 &= \frac{-2(\mu_G-m^2\mu_B)\pm\sqrt{[(2)(m^2\mu_B-\mu_G)]^2-4(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln(Cm))}}{2(m^2-1)} \\
k_0 &= \frac{2(m^2\mu_B^2-\mu_G^2)\pm\sqrt{4(m^2\mu_B-\mu_G)^2-4(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln(Cm))}}{2(m^2-1)} \\
k_0 &= \frac{(m^2\mu_B-\mu_G)\pm\sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln(Cm))}}{m^2-1}
\end{aligned}
$$

Theorem 4: Assume a Normal-Normal distribution for the two populations with $N(\mu_G,(m\sigma)^2)$ and $N(\mu_B,\sigma^2)$ with $\mu_G > \mu_B$ and $\sigma_G = m\sigma_B$. Let $C = \frac{P_B}{P_G}\frac{(r_B+c_B)}{(r_G+c_G)}$ and k_0 be defined as in Thm. 3. Then:

1. $R'(k) \geq 0$ when $k \leq k_0$.
2. $R'(k) \leq 0$ when $k \geq k_0$.
3. $R''(k_0) < 0$ when $k_0 = \frac{(m^2\mu_B-\mu_G)+\sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln(Cm))}}{m^2-1}$.

Proof: (1) and (2) can be easily proved by changing the equal sign in the Thm. 3 to $\geq$ and $\leq$, respectively. To derive (3), we take the second derivative of $R(k)$

with respect to k:

$$\frac{\partial^2 R(k)}{\partial k^2} = -\frac{(\mu_G - k)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k-\mu_G)^2} P_G(r_G+c_G) + \frac{(\mu_B - k)}{m^3\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2m^2\sigma^2}(k-\mu_B)^2} P_B(r_B+c_B)$$

Considering the local maximum for which $R''(k) < 0$ we look at:

$$\frac{(\mu_B - k_0)}{m^3\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2m^2\sigma^2}(k_0-\mu_B)^2} P_B(r_B + c_B) < \frac{(\mu_G - k_0)}{\sigma^3\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(k_0-\mu_G)^2} P_G(r_G + c_G)$$

$$m^3 \frac{\mu_B - k_0}{\mu_G - k_0} C < exp\left[\frac{(\mu_B - k_0)^2}{2\sigma^2} - \frac{(\mu_G - k_0)^2}{2m^2\sigma^2}\right]$$

$$\begin{aligned}
2\sigma^2 m^2 \left(\ln(\mu_B - k_0) + \ln(m^3) - \ln(\mu_G - k_0)\right) + \mu_B - k_0 \ln(C) <& m^2(\mu_B - k_0)^2 - (\mu_G - k_0)^2 \\
2\sigma^2 m^2 \left(\ln(\mu_B - k_0) + \ln(m^3) - \ln(\mu_G - k_0)\right) + \mu_B - k_0 \ln(C) <& (m^2 - 1)k_0 + 2(\mu_G - \mu_B m^2)k_0 \\
&+ (\mu_B^2 m^2 - \mu_G^2)
\end{aligned}$$

Completing the square with the quadratic from Thm. 3, we get:

$$\begin{aligned}
2\sigma^2 m^2 \left(\ln(\mu_B - k_0) + \ln(m^3) - \ln(\mu_G - k_0)\right) + \mu_B - k_0 \ln(C) <& - m^2\mu_B^2 + \mu_G^2 + 2m^2\sigma^2 \ln(Cm) \\
&+ \mu_B^2 m^2 - \mu_G^2
\end{aligned}$$

$$\begin{aligned}
\ln(\mu_B - k_0) - \ln(\mu_G - k_0) <& - \ln(Cm) + \ln(m^3) - \ln(C) \\
\frac{(\mu_B - k_0)}{(\mu_G - k_0)} <& \frac{1}{m^2}
\end{aligned}$$

This only holds true if $k_0 = \frac{(m^2\mu_B - \mu_G) + \sqrt{(m^2\mu_B - \mu_G)^2 - (m^2-1)(m^2\mu_B^2 - \mu_G^2 - 2\sigma^2 m^2 \ln(Cm))}}{m^2-1}$, so only the one k_0 value from Thm. 3 proves a local maximum where $R''(k) < 0$. Theorem 9 below follows immediately from Theorem 3.

4 Gamma Distribution

Once we considered both of the previously examined cases for the Normal distribution, we decided to consider what a threshold may be for asymmetric distributions. Asymmetric distributions are likely to be more appropriate models for the distribution of Y^* as the scores will likely be clustered around the means of the Good or Bad status with a skew toward the center of the joint distribution. This would be a skew right for the Bad population and a skew left for the Good population. We initially considered the Γ-Distribution as it is defined for $y^* > 0$.

4.1 Case 3: $\beta_G = \beta_B = 2$

A special case of $\beta_G = \beta_B$ is one where Y^* is $\chi^2 - \chi^2$ distributed [2]. This is when $\beta_G = \beta_B = 2$, $\alpha_G = \frac{\nu_G}{2}$, and $\alpha_B = \frac{\nu_B}{2}$.

Let $Y^* \sim \chi^2(\nu_G)$ for the good population and $Y^* \sim \chi^2(\nu_B)$ for the bad population.

$$f_G(y^*|\nu_G) = \frac{1}{\Gamma(\frac{\nu_G}{2})2^{\nu_G/2}} y^{*\frac{\nu_G}{2}-1} e^{-\frac{y^*}{2}}$$

$$f_B(y^*|\nu_B) = \frac{1}{\Gamma(\frac{\nu_B}{2})2^{\nu_B/2}} y^{*\frac{\nu_B}{2}-1} e^{-\frac{y^*}{2}}$$

Under these assumptions, we have:

$TP_k = P(y^* > k|D = G) = 1 - F_G(k) = \int_k^\infty f_G(y^*)dy^*$

and

$FP_k = P(y^* > k|D = B) = 1 - F_B(k) = \int_k^\infty f_B(y^*)dy^*$

Theorem 5: Under the assumption of χ^2-χ^2 model. k_0 can be shown to be expressible in a closed-form expression involving P_G, P_B, c_G, c_B, r_G, r_B, ν_G, and ν_B.

$k_0 = 2(CD)^{\frac{2}{\nu_G - \nu_B}}$

where

$C = \frac{P_B}{P_G}\frac{r_B + c_B}{r_G + c_G}$

and

$D = \frac{\Gamma(\frac{\nu_G}{2})}{\Gamma(\frac{\nu_B}{2})}$.

Proof: Given the profit function $R(k)$:

$$\begin{aligned} R(k) &= P_G * r_G * TP_k + P_B * r_B * TN_k - (P_G * c_G * FN_k + P_B * c_B * FP_k) \\ &= P_G * (r_G + c_G) * TP_k - P_B * (r_B + c_B) * FP_k + (P_B * r_B - P_G * c_G) \end{aligned}$$

Under the χ^2-χ^2 model assumptions, we have

$TP_k = \int_k^\infty \frac{1}{\Gamma(\frac{\nu_G}{2})2^{\nu_G/2}} y^{*\frac{\nu_G}{2}-1} e^{-\frac{y^*}{2}} dy^*$

and

$FP_k = \int_k^\infty \frac{1}{\Gamma(\frac{\nu_B}{2})2^{\nu_B/2}} y^{*\frac{\nu_B}{2}-1} e^{-\frac{y^*}{2}} dy^*$

Take the first derivative of the Profit Function with respect to k, we have

$$\begin{aligned} R'(k) &= P_B * (r_B + c_B) * f_B(k) - P_G * (r_G + c_G) * f_G(k) \\ &= P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)}\right) \end{aligned}$$

where $C = \frac{P_B}{P_G} * \frac{r_B + c_B}{r_G + c_G}$. Then $R'(k) = 0$ if and only if

$P_G * (r_G + c_G) * f_B(k_0) * \left(C - \frac{f_G(k_0)}{f_B(k_0)}\right) = 0$

Since P_G, r_G, c_G, and $f_B(k)$ are all positive, the previous equation is equivalent to the following condition.

$$C = \frac{f_G(k_0)}{f_B(k_0)} = \frac{\frac{1}{\Gamma(\frac{\nu_G}{2})2^{\nu_G/2}} k_0^{\frac{\nu_G}{2}-1} e^{-\frac{k_0}{2}}}{\frac{1}{\Gamma(\frac{\nu_B}{2})2^{\nu_B/2}} k_0^{\frac{\nu_B}{2}-1} e^{-\frac{k_0}{2}}}$$

$$C = \frac{\Gamma(\frac{\nu_B}{2})}{\Gamma(\frac{\nu_G}{2})} 2^{\nu_B/2-\nu_G/2} k_0^{\nu_G/2-1-\nu_B/2+1} e^{-k_0/2+k_0/2}$$

Let $D = \frac{\Gamma(\frac{\nu_G}{2})}{\Gamma(\frac{\nu_B}{2})}$.

$$CD = 2^{\frac{1}{2}(\nu_B-\nu_G)} k_0^{\frac{1}{2}(\nu_G-\nu_B)}$$

$$\ln(CD) = \frac{1}{2}(\nu_B-\nu_G)\ln(2) + \frac{1}{2}(\nu_G-\nu_B)\ln(k_0)$$

$$\ln(CD) + \frac{1}{2}(\nu_G-\nu_B)\ln(2) = \frac{1}{2}(\nu_G-\nu_B)\ln(k_0)$$

$$\frac{2}{\nu_G-\nu_B}\ln(CD) + \ln(2) = \ln(k_0)$$

$$2(CD)^{\frac{2}{\nu_G-\nu_B}} = k_0$$

Theorem 6: Assume a χ^2 distribution for the two populations with $\chi^2(\nu_G)$ and $\chi^2(\nu_B)$ with $\nu_G > \nu_B$. Let $C = \frac{P_B}{P_G}\frac{(r_B+c_B)}{(r_G+c_G)}$, $D = \frac{\Gamma(\frac{\nu_G}{2})}{\Gamma(\frac{\nu_B}{2})}$, and k_0 be defined as in Thm. 5. Then:

1. $R'(k) \geq 0$ when $k \leq k_0$.
2. $R'(k) \leq 0$ when $k \geq k_0$.
3. $R''(k) < 0$.

Proof: (1) and (2) can be easily proved by changing the equal sign in Thm. 5 to $\geq$ and $\leq$, respectively. To derive the (3), we take the second derivative of $R(k)$ with respect to k:

$$\begin{aligned}\frac{\partial^2 R(k)}{\partial k^2} =& - P_G(r_G+c_G)\frac{1}{\Gamma(\frac{\nu_G}{2})2^{\frac{\nu_G}{2}}}\left(-\frac{1}{2}k^{\frac{\nu_G}{2}-1}e^{\frac{-k}{2}} + (\frac{\nu_G}{2}-1)e^{\frac{-k}{2}}k^{\frac{\nu_G}{2}-1}\right)\\ &+ P_B(r_B+c_B)\frac{1}{\Gamma(\frac{\nu_B}{2})2^{\frac{\nu_B}{2}}}\left(-\frac{1}{2}k^{\frac{\nu_B}{2}-1}e^{\frac{-k}{2}} + (\frac{\nu_B}{2}-1)e^{\frac{-k}{2}}k^{\frac{\nu_B}{2}-1}\right)\end{aligned}$$

Considering the local maximum, for which $R''(k) < 0$ we look at:

$$P_G(r_G+c_G)\frac{1}{\Gamma(\frac{\nu_G}{2})2^{\frac{\nu_G}{2}}}(k_0^{\frac{\nu_G}{2}-1}+(2-\nu_G)k_0^{\frac{\nu_G}{2}-1}) < P_B(r_B+c_B)\frac{1}{\Gamma(\frac{\nu_B}{2})2^{\frac{\nu_B}{2}}}(k_0^{\frac{\nu_B}{2}-1}+(2-\nu_B)k_0^{\frac{\nu_B}{2}-1})$$

$$(k_0^{\frac{\nu_G}{2}-1}+(2-\nu_G)k_0^{\frac{\nu_G}{2}-1}) < CD2^{\frac{\nu_G-\nu_B}{2}}(k_0^{\frac{\nu_B}{2}-1}+(2-\nu_B)k_0^{\frac{\nu_B}{2}-1})$$
$$k_0^{\frac{\nu_G}{2}-1}(1+(2-\nu_G)k_0^{-1}) < CD2^{\frac{\nu_G-\nu_B}{2}}k_0^{\frac{\nu_B}{2}-1}(1+(2-\nu_B)k_0^{-1})$$

Substituting for k_0:

$$CD2^{\frac{\nu_G-\nu_B}{2}}(1+(2-\nu_G)k_0^{-1}) < CD2^{\frac{\nu_G-\nu_B}{2}}k_0^{\frac{\nu_B}{2}-1}(1+(2-\nu_B)k_0^{-1})$$
$$(1+(2-\nu_G)k_0^{-1}) < (1+(2-\nu_B)k_0^{-1})$$
$$-\nu_G < -\nu_B$$
$$\mu_G = \nu_G > \nu_B = \mu_B$$

4.2 Case 4: $\beta_G \neq \beta_B$

Let $Y^* \sim \Gamma(\alpha_G, \beta_G)$ for the good population and $Y^* \sim \Gamma(\alpha_B, \beta_B)$ for the bad population [3].

$f_G(y^*) = \frac{1}{\Gamma(\alpha_G)\beta_G^{\alpha_G}} y^{*\alpha_G-1} e^{\frac{-y^*}{\beta_G}}$

$f_B(y^*) = \frac{1}{\Gamma(\alpha_B)\beta_B^{\alpha_B}} y\, *^{\alpha_B-1}\, e^{\frac{-y^*}{\beta_B}}$

Under these assumptions, we have:

$TP_k = P(y^* > k|D = G) = 1 - F_G(k) = \int_k^\infty f_G(y^*)dy^*$

and

$FP_k = P(y^* > k|D = B) = 1 - F_B(k) = \int_k^\infty f_B(y^*)dy^*$.

Recall the profit function:

$$\begin{aligned} R(k) &= P_G * r_G * TP_k + P_B * r_B * TN_k - (P_G * c_G * FN_k + P_B * c_B * FP_k) \\ &= P_G * (r_G + c_G) * TP_k - P_B * (r_B + c_B) * FP_k + (P_B * r_B - P_G * c_G) \end{aligned}$$

Theorem 7: Under the assumptions, k_0 can be shown to be unique and expressible in a closed-form expression involving P_G, P_B, c_G, c_B, r_G, r_B, α_G, α_B, and β_G, and β_B.

$$k_0 = \frac{(\alpha_B - \alpha_G)\beta_G\beta_B}{\beta_B - \beta_G} W\left(\frac{(\beta_B - \beta_G)(CD)^{\frac{-1}{\alpha_B - \alpha_G}}}{(\alpha_B - \alpha_G)\beta_B\beta_G}\right)$$
$$C = \frac{P_B}{P_G}\frac{r_B + c_B}{r_G + c_G}$$
$$D = \frac{\Gamma(\alpha_G)\beta_G^{\alpha_G}}{\Gamma(\alpha_B)\beta_B^{\alpha_B}}$$

Proof: Given the profit function $R(k)$:

$$\begin{aligned} R(k) &= P_G * r_G * TP_k + P_B * r_B * TN_k - (P_G * c_G * FN_k + P_B * c_B * FP_k) \\ &= P_G * (r_G + c_G) * TP_k - P_B * (r_B + c_B) * FP_k + (P_B * r_B - P_G * c_G) \end{aligned}$$

Under the assumptions, we have

$TP_k = \int_k^\infty \frac{1}{\Gamma(\alpha_G)\beta_G^{\alpha_G}} y^{*\alpha_G - 1} e^{\frac{-y^*}{\beta_G}} dy^*$

and

$FP_k = \int_k^\infty \frac{1}{\Gamma(\alpha_B)\beta_B^{\alpha_B}} y *^{\alpha_B - 1} e^{\frac{-y^*}{\beta_B}} dy^*$

Take the first derivative with respect to k, we have,

$$\begin{aligned} R'(k) &= P_B * (r_B + c_B) * f_B(k) - P_G * (r_G + c_G) * f_G(k) \\ &= P_G * (r_G + c_G) * f_B(k) * \left(C - \frac{f_G(k)}{f_B(k)} \right) \end{aligned}$$

where $C = \frac{P_B}{P_G} * \frac{r_B + c_B}{r_G + c_G}$. Then $R'(k) = 0$ if and only if
Since P_G, r_G, c_G, and $f_B(k)$ are all positive, the previous equation is equivalent to the following condition.

$$\begin{aligned} P_G * (r_G + c_G) * f_B(k_0) * \left(C - \frac{f_G(k_0)}{f_B(k_0)} \right) &= 0 \\ C &= \frac{f_G(k_0)}{f_B(k_0)} \\ C &= \frac{\frac{1}{\Gamma(\alpha_G)\beta_G^{\alpha_G}} k_0^{\alpha_G - 1} e^{\frac{-k_0}{\beta_G}}}{\frac{1}{\Gamma(\alpha_B)\beta_B^{\alpha_B}} k_0^{\alpha_B - 1} e^{\frac{-k_0}{\beta_B}}} \\ C &= \frac{\Gamma(\alpha_B)\beta_B^{\alpha_B}}{\Gamma(\alpha_G)\beta_G^{\alpha_G}} k_0^{\alpha_G - \alpha_B} e^{-k_0\left(\frac{1}{\beta_G} - \frac{1}{\beta_B}\right)} \\ C \frac{\Gamma(\alpha_G)\beta_G^{\alpha_G}}{\Gamma(\alpha_B)\beta_B^{\alpha_B}} &= k_0^{\alpha_G - \alpha_B} e^{-k_0\left(\frac{1}{\beta_G} - \frac{1}{\beta_B}\right)} \\ \text{Let } D = \frac{\Gamma(\alpha_G)\beta_G^{\alpha_G}}{\Gamma(\alpha_B)\beta_B^{\alpha_B}}. & \\ CD &= k_0^{\alpha_G - \alpha_B} e^{-k_0\left(\frac{1}{\beta_G} - \frac{1}{\beta_B}\right)} \end{aligned}$$

After some math...(Mathematica),

$$k_0 = \frac{(\alpha_B - \alpha_G)\beta_G\beta_B}{\beta_B - \beta_G} W \left(\frac{(\beta_B - \beta_G)(CD)^{\frac{-1}{\alpha_B - \alpha_G}}}{(\alpha_B - \alpha_G)\beta_B\beta_G} \right)$$

where W is the *Lambert* W function. The *Lambert* W function is the function that satisfies $z = W(z)e^{W(z)}$ or the inverse of $f(W) = We^W$ [4]. ■

Theorem 8: Assume a Γ- distribution for the two populations with $\Gamma(\alpha_G, \beta_G)$ and $\Gamma(\alpha_B, \beta_B)$ with $\frac{\mu_G - k_0}{\beta_G} > \frac{\mu_B - k_0}{\beta_B}$. Let $C = \frac{P_B}{P_G}\frac{(r_B + c_B)}{(r_G + c_G)}$, $D = \frac{\Gamma(\alpha_G)\beta_G^{\alpha_G}}{\Gamma(\alpha_B)\beta_B^{\alpha_B}}$, and k_0 be defined by Thm. 7. Then:

1. $R'(k) \geq 0$ when $k \leq k_0$.
2. $R'(k) \leq 0$ when $k \geq k_0$.
3. $R''(k) < 0$.

Proof: (1) and (2) can be easily proved by changing the equal sign in Thm. 7 to $\geq$ and $\leq$, respectively. To derive the (3), we take the second derivative of $R(k)$ with respect to k:

$$\begin{aligned}\frac{\partial^2 R(k)}{\partial k^2} =& - P_B(r_B + c_B)\frac{(\frac{1}{\beta_B})^{\alpha_B}}{\Gamma(\alpha_B)}\left(\frac{-1}{\beta_B}k^{\alpha_B - 1}e^{\frac{-k}{\beta_B}} + (\alpha_B - 1)k^{\alpha_B - 2}e^{\frac{-k}{\beta_B}}\right) \\ &+ P_G(r_G + c_G)\frac{(\frac{1}{\beta_G})^{\alpha_G}}{\Gamma(\alpha_G)}\left(\frac{-1}{\beta_G}k^{\alpha_G - 1}e^{\frac{-k}{\beta_G}} + (\alpha_G - 1)k^{\alpha_G - 2}e^{\frac{-k}{\beta_G}}\right)\end{aligned}$$

Considering the local maximum, for which $R''(k) < 0$ we look at:

$$\frac{1}{\beta_G}k_0^{\alpha_G - 1}e^{\frac{-k_0}{\beta_G}} + (1 - \alpha_G)k_0^{\alpha_G - 2}e^{\frac{-k_0}{\beta_G}} < CD\left(\frac{1}{\beta_B}k_0^{\alpha_B - 1}e^{\frac{-k_0}{\beta_B}} + (1 - \alpha_B)k_0^{\alpha_B - 2}e^{\frac{-k_0}{\beta_B}}\right)$$

$$\begin{aligned}\frac{1}{\beta_G}k_0^{\alpha_G - 1} + (1 - \alpha_G)k_0^{\alpha_G - 2} &< CD\left(\frac{1}{\beta_B}k_0^{\alpha_B - 1} + (1 - \alpha_B)k_0^{\alpha_B - 2}\right)e^{\frac{-k_0}{\beta_B} + \frac{k_0}{\beta_G}} \\ k_0^{\alpha_G - 1}\left(\frac{1}{\beta_G} + (1 - \alpha_G)k_0^{-1}\right) &< CD\left(\frac{1}{\beta_B} + (1 - \alpha_B)k_0^{-1}\right)k_0^{\alpha_B - 1}e^{\frac{-k_0}{\beta_B} + \frac{k_0}{\beta_G}} \\ k_0^{\alpha_G - \alpha_B}e^{\frac{-k_0}{\beta_B} + \frac{k_0}{\beta_G}}\left(\frac{1}{\beta_G} + (1 - \alpha_G)k_0^{-1}\right) &< CD\left(\frac{1}{\beta_B} + (1 - \alpha_B)k_0^{-1}\right)\end{aligned}$$

And since $e^{\frac{-k_0}{\beta_B} + \frac{k_0}{\beta_G}} = CD$ then:

$$\begin{aligned}\frac{1}{\beta_G} + (1 - \alpha_G)k_0^{-1} &< \frac{1}{\beta_B} + (1 - \alpha_B)k_0^{-1} \\ \frac{1}{\beta_G} - \frac{\alpha_G}{k_0} &< \frac{1}{\beta_B} - \frac{\alpha_B}{k_0} \\ \alpha_G - \alpha_B &> \frac{-k_0}{\beta_B} + \frac{k_0}{\beta_G} \\ \alpha_G - \frac{k_0}{\beta_G} &> \alpha_B - \frac{k_0}{\beta_B} \\ \frac{\mu_G - k_0}{\beta_G} &> \frac{\mu_B - k_0}{\beta_B} \quad ■\end{aligned}$$

5 Maximized Profit Value

In this section, we will discuss the three possible relations k_0 can have to the minimum and maximum creditworthiness score.

Theorem 9: Let $R(k)$ be defined for $k \in [y^*_{min}, y^*_{max}]$, and k_0 be defined by Theorem 1. The maximization of $R(k)$ can be broken into three cases.

1. If $k_0 \leq y^*_{min}$, then $R(k)$ is maximal when $k^* = y^*_{min}$. If y^*_{min} is small enough, then $R(y^*_{min}) \approx P_G \cdot r_G - P_B \cdot c_B$.

2. If $k_0 \geq y^*_{max}$, then $R(k)$ is maximal when $k^* = y^*_{max}$. If y^*_{max} is big enough, then $R(y^*_{max}) \approx P_B \cdot r_B - P_G \cdot c_G$.

3. If $y^*_{min} < k_0 < y^*_{max}$, then $R(k)$ is maximal when $k^* = k_0$ with

$$\begin{aligned} R(k_0) = & P_B \cdot (r_B + c_B) \cdot (1 - F(k_0)) \\ & - P_G \cdot (r_G + c_G) \cdot (1 - F(k_0)) + P_G \cdot r_G - P_B \cdot c_B \end{aligned}$$

Proof:

1. Since $k \in [k_1, k_2]$, if $k_0 \leq k_1$, then $k_0 \leq k$. By Theorem 2 (2), $R'(k) \leq 0$. Hence, $R(k)$ is a decreasing function over $k \in [k_1, k_2]$, and then the maximum of $R(k)$ occurs at $k = k_1$.

   $R(k) = P_G \cdot (r_G + c_G) \cdot TP_k - P_B \cdot (r_B + c_B) \cdot FP_k + P_B \cdot r_B - P_G \cdot c_G$. As $k \to -\infty$, $TP_k \to 1$ and $FP_k \to 1$, so

$$\begin{aligned} R(k) & \to P_G r_G + P_G c_G - P_B r_B - P_B c_B + P_B r_B - P_G c_G \\ & \to P_G r_G - P_B c_B \end{aligned}$$

2. Since $k \in [k_1, k_2]$, if $k_0 \geq k_2$, then $k_0 \geq k$. By Theorem 2 (1), $R'(k) \geq 0$. Hence, $R(k)$ is an increasing function over $k \in [k_1, k_2]$, and then the maximum of $R(k)$ occurs at $k = k_2$.

   As $k \to \infty$, $TP_k \to 0$ and $FP_k \to 0$, so $R(k) \to P_B r_B - P_G c_G$

3. If $k_1 < k_0 < k_2$, then, by Theorem 2, $R(k)$ is maximal at $k = k_0$. ■

The same argument holds for the Normal-Normal case where $\sigma_G \neq \sigma_B$ and the two different Gamma-Gamma cases.

6 Examples

Now we will briefly discuss one example of a calculated threshold and its corresponding simulation. We only include Case 4, Γ-Distribution with $\beta_G \neq \beta_B$, here for brevity, although we will discuss the basic set up for the other three cases as well. The other three pairs of simulations and calculations can be found in the appendix. Each simulation followed the same procedure as outlined in Section 2. Then we calculated the optimal threshold using the formulas derived in Sections 3 & 4.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
r_G	\$1,400	\$1,400	\$1,400	\$1,400	\$1,400	\$1,400
r_B	\$280	\$280	\$280	\$280	\$280	\$280
c_G	\$560	\$560	\$560	\$560	\$560	\$560
c_B	\$42	\$1,680	\$2,800	\$4,900	\$7,000	\$105,000

Table 3: Different Cases for Revenue and Cost Values

6.1 Background

As seen in the paper by Chen, Das, and Gong [1], the cost revenue matrix consists of six different cases along with consistent means and proportions for the good and bad populations. These values were used by Chen et al [1]. We used the values as theoretical values to calculate and simulate our optimal thresholds regardless of their validity. For the calculations and simulations with the Normal Distribution, we used the Case 2 values for r_G, r_B, c_G, and c_B along with the Chen et al.'s values [1] for μ_G, μ_B, P_G, and P_B. For our calculations and simulations with the Γ-Distribution, we again used the Case 2 values for r_G, r_B, c_G, and c_B along with the Chen et al.'s values for P_G, and P_B. We arbitrarily chose values for α_G, α_B, β_G, and β_B to create distributions with reasonably realistic shapes. Since the Γ- Distribution's support is only positive, we must transform by adding the minimum y^* value to get negative values. Let $X \sim \Gamma(\alpha, \beta)$, such that $x > 0$ and let Y^* be a psuedo-Γ distribution, such that $y^* > y^*_{min}$. $Y^* = X + Y^*_{min}$ transforms X such that it follows a psuedo-Γ distribution such that $X > Y^*_{min}$ and is the same distribution as Y^*.

6.2 Case 4 Simulation

In each of the four simulations, we used a sample size of $n = 20,000$ and $P_G = P(Status = Good) = 0.735$ and $P_B = P(Status = Bad) = 0.265$ as seen in the work done by Chen et al [1]. We simulated data and thresholds to find an optimal threshold. To do so we randomly generated two sets of values for the Good and Bad populations using the appropriate distributions and parameters.

For the Γ-Distribution case, we have arbitrarily chosen $\alpha_G = \frac{10}{3}$, $\alpha_B = \frac{5}{2}$, $\beta_G = 3$, and $\beta_B = 2$. These values satisfy our assumed inequality $\frac{\mu_G - k}{\beta_G} > \frac{\mu_B - k}{\beta_B}$. With these distributions there is a good deal of overlap, so we can expect a high TP and moderately high FP. All of the data lie in the interval $[y^*_{min}, y^*_{max}] = [0.144, 44.353]$.

6.3 Case 4 Calculations

Next, we calculated the optimal threshold. Note that for each of the four cases, the value of C will not change as the variables it consists of remain

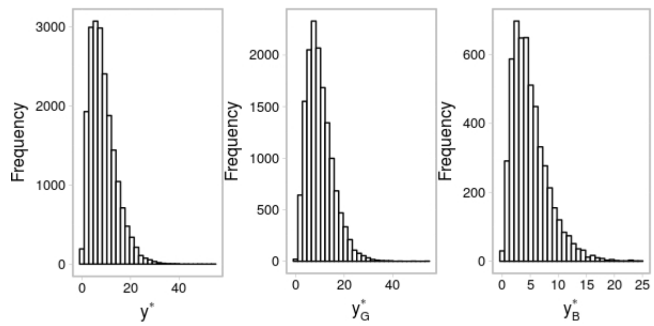

Figure 1: Histograms of $Y_G^* \sim \Gamma(\alpha_G = 3.33, \beta_G = 3)$ and $Y_B^* \sim \Gamma(\alpha_B = 2.5, \beta_B = 2)$

constant.

$$C = \frac{P_B}{P_G}\left(\frac{r_B + c_B}{r_G + c_G}\right) = \frac{0.265}{0.735}\left(\frac{280 + 1680}{1400 + 560}\right) = 0.361$$

Since we considered the case of two Γ-Distributions, our optimal threshold is calculated by Thm. 7.

$$D = \frac{\Gamma(\alpha_G)\beta_G^{\alpha_G}}{\Gamma(\alpha_B)\beta_B^{\alpha_B}} = \frac{\Gamma(10/3)3^{10/3}}{\Gamma(5/2)2^{5/2}} = 14.3863$$

$$\begin{aligned} k_0 &= \frac{(\alpha_B - \alpha_G)\beta_G\beta_B}{\beta_B - \beta_G} W\left(\frac{(\beta_B - \beta_G)(CD)^{\frac{-1}{\alpha_B - \alpha_G}}}{(\alpha_B - \alpha_G)\beta_B\beta_G}\right) \\ &= \frac{(5/2 - 10/3)(3)(2)}{2 - 3} W\left(\frac{(2-3)(0.361(14.3863))^{\frac{-1}{5/2 - 10/3}}}{(5/2 - 10/3)(3)(2)}\right) \\ &= 3.547 \end{aligned}$$

Since $k_0 \in [y^*_{min}, y^*_{max}]$, by Thm. 8, $k_0 = k^*$ and is our optimal threshold.

$$\begin{aligned} R(k^*) &= P_G \cdot TP_{k^*}(r_G + c_G) - P_B \cdot FP_{k^*}(r_B + c_B) + P_B r_B - P_G c_G \\ &= 0.735 \cdot 0.922(1400 + 560) - 0.265 \cdot 0.616(280 + 1680) + 0.265(280) - 0.735(520) \\ &= 670.85 \end{aligned}$$

Evaluating $R(k^*)$, we can see that the maximized profit for this example is \$670.85. Comparing our simulated data and calculated threshold, we can see that the calculated threshold value does maximize the profit function.

See Appendices B & C for more example simulations and calculations respectively.

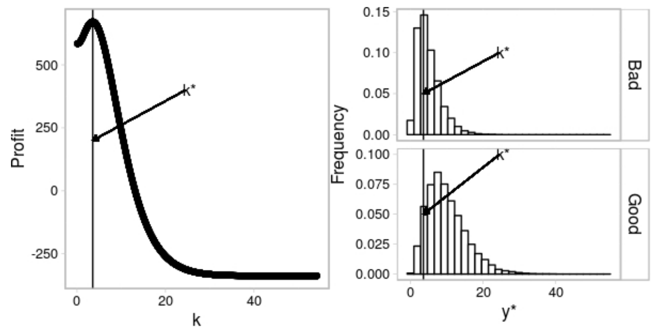

Figure 2: Simulated Data with Optimal Threshold k^* for the case where $\alpha_G = 3.33$, $\alpha_B = 2.5$, $\beta_G = 3$, and $\beta_B = 2$

7 Conclusion

Throughout the financial sector, at banks and other lenders, it is necessary for institutions to assess the credit applications and analyze the risk of each case, in the form of creditworthiness, in order for the institution to maximize profit. By considering at the Profit Function, we are able to derive a formula for the optimal creditworthiness threshold values that maximize profit that can be easily altered for changes in cost, revenue, and population assumptions.

We were able to find optimal creditworthiness threshold values for three different distributions with varying assumptions. In addition to these three distributions, we also explored threshold values for Student's t distribution and mixed distributions, such as the Good population being distributed Normal and the Bad being distributed Γ. Closed form expressions do not exist for these cases.

Future directions to the research of an optimal creditworthiness threshold could include running simulations to approximate the creditworthiness value at which the profit is maximized when the populations are different distributions. The research could also be expanded beyond the financial application by modifying the Profit Function to applications in other fields.

Acknowledgements

Thank you to the NSF for funding us (NSF grant DMS-1559912), to our research advisor, Dr. Gong, and to Valparaiso University for hosting us this summer.

References

[1] Chen J., Das J., and Gong H. Selection of optimal threshold using cost and revenue matrix. *Model Assisted Statistics and Applications*, 9:97–107, 2014.

[2] Chi-squared distribution — Wikipedia, the free encyclopedia, June 2018. https://en.wikipedia.org/wiki/Chi-squared_distribution.

[3] Weisstein Eric W. Gamma Distribution From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/GammaDistribution.html.

[4] Weisstein Eric W. Lambert W-Function From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/LambertW-Function.html.

Appendix

A Location of k_0 with respect to μ_G & μ_B

A.1 Case 1: Normal with $\sigma_G = \sigma_B$

Let $k^* = k_0$ be the optimal threshold given in Theorem 1. Then

1. If $\ln C \leq -\frac{(\mu_G-\mu_B)^2}{2\sigma^2}$, then $k_0 \leq \mu_B$.
2. If $\ln C \geq \frac{(\mu_G-\mu_B)^2}{2\sigma^2}$, then $k_0 \geq \mu_G$.
3. If $-\frac{(\mu_G-\mu_B)^2}{2\sigma^2} < \ln C < \frac{(\mu_G-\mu_B)^2}{2\sigma^2}$, then $\mu_B < k_0 < \mu_G$.

Proof:

1. If $\ln C \leq -\frac{(\mu_G-\mu_B)^2}{2\sigma^2}$, then $\frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2} \leq \mu_B$.
   Given $k_0 = \frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2}$, then $k_0 \leq \mu_B$.
2. If $\ln C \geq \frac{(\mu_G-\mu_B)^2}{2\sigma^2}$, then $\frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2} \geq \mu_G$.
   Given $k_0 = \frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2}$, then $k_0 \geq \mu_G$.
3. If $\frac{(\mu_G-\mu_B)^2}{2\sigma^2} < \ln C < \frac{(\mu_G-\mu_B)^2}{2\sigma^2}$,
   then $\frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2} > \mu_B$ and $\frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2} < \mu_G$.
   Therefore, when $k_0 = \frac{\sigma^2 \ln C}{\mu_G-\mu_B} + \frac{\mu_G+\mu_B}{2}$, then $\mu_B < k_0 < \mu_G$. ■

A.2 Case 2: Normal with $\sigma_G \neq \sigma_B$

Let $k^* = k_0$ be the optimal threshold given in Theorem 3. Then

1. If $\ln C \leq \frac{2\mu_G\mu_B-\mu_G^2-\mu_B^2}{2m^2\sigma^2} - \ln m$, then $k_0 \leq \mu_B$.

2. If $\ln C \geq \frac{\mu_G^2+\mu_B^2-2\mu_G\mu_B}{2\sigma^2} - \ln m$, then $k_0 \geq \mu_G$.
3. If $\frac{2\mu_G\mu_B-\mu_G^2-\mu_B^2}{2m^2\sigma^2} - \ln m < \ln C < \frac{\mu_G^2+\mu_B^2-2\mu_G\mu_B}{2\sigma^2} - \ln m$, then $\mu_B < k_0 < \mu_G$.

Proof:

1. If $\ln C \leq \frac{2\mu_G\mu_B-\mu_G^2-\mu_B^2}{2m^2\sigma^2} - \ln m$, then

$$\begin{aligned}
\mu_G^2+\mu_B^2-2\mu_G\mu_B &\leq -2m^2\sigma^2\ln Cm \\
m^2\mu_B^2+\mu_B^2-2\mu_G\mu_B &\leq m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm \\
-\frac{\left((\mu_G-\mu_B)^2-(m^2\mu_B-\mu_G)^2\right)}{m^2-1} &\leq m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm \\
(\mu_G-\mu_B)^2 &\geq (m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm) \\
\mu_B m^2-\mu_B &\geq m^2\mu_B-\mu_G \\
&\quad + \sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm)} \\
\mu_B &\geq \frac{1}{m^2-1}((m^2\mu_B-\mu_G) \\
&\quad + \sqrt{(\mu_G-m^2\mu_B)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln Cm)}\ )
\end{aligned}$$

Given

$$k_0 = \frac{(m^2\mu_B-\mu_G)+\sqrt{(\mu_G-m^2\mu_B)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln Cm)}}{m^2-1}$$

then $k_0 \leq \mu_B$.

2. If $\ln C \geq \frac{\mu_G^2+\mu_B^2-2\mu_G\mu_B}{2\sigma^2} - \ln m$, then

$$\begin{aligned}
2m^2\mu_G\mu_B - m^2\left(\mu_G^2+\mu_B^2\right) &\geq -2m^2\sigma^2\ln Cm \\
m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm &\leq -\left(\mu_G^2(m^2+1)-2m^2\mu_G\mu_B\right) \\
\left(-\left(\mu_G^2(m^2+1)-2m^2\mu_G\mu_B\right)\right) &\leq -(m^2-1)\left(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm\right) \\
m^2\left(\mu_G-\mu_B\right) &\leq \sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm)} \\
\mu_G m^2-\mu_G &\leq m^2\mu_B-\mu_G \\
&\quad + \sqrt{(m^2\mu_B-\mu_G)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2m^2\sigma^2\ln Cm)} \\
\mu_G &\leq \frac{1}{m^2-1}((m^2\mu_B-\mu_G) \\
&\quad + \sqrt{(\mu_G-m^2\mu_B)^2-(m^2-1)(m^2\mu_B^2-\mu_G^2-2\sigma^2m^2\ln Cm)}\ )
\end{aligned}$$

Given

$$k_0 = \frac{(m^2\mu_B - \mu_G) + \sqrt{(\mu_G - m^2\mu_B)^2 - (m^2-1)(m^2\mu_B^2 - \mu_G^2 - 2\sigma^2 m^2 \ln Cm)}}{m^2-1}$$

then $k_0 \geq \mu_G$.

3. If

$$\frac{2\mu_G\mu_B - \mu_G^2 - \mu_B^2}{2m^2\sigma^2} - \ln m < \ln C < \frac{\mu_G^2 + \mu_B^2 - 2\mu_G\mu_B}{2\sigma^2} - \ln m$$

then

$$\frac{(m^2\mu_B - \mu_G) + \sqrt{(\mu_G - m^2\mu_B)^2 - (m^2-1)(m^2\mu_B - \mu_G^2 - 2\sigma^2 m^2 \ln Cm)}}{m^2-1} > \mu_B$$

and

$$\frac{(m^2\mu_B - \mu_G) + \sqrt{(\mu_G - m^2\mu_B)^2 - (m^2-1)(m^2\mu_B - \mu_G^2 - 2\sigma^2 m^2 \ln Cm)}}{m^2-1} < \mu_G$$

Therefore, when

$$k_0 = \frac{(m^2\mu_B - \mu_G) + \sqrt{(\mu_G - m^2\mu_B)^2 - (m^2-1)(m^2\mu_B - \mu_G^2 - 2\sigma^2 m^2 \ln Cm)}}{m^2-1}$$

then $\mu_B < k_0 < \mu_G$. ■

A.3 Case 3: Γ with $\beta_G = \beta_B$

Let $k^* = k_0$ be the optimal threshold given in Theorem 5. Then

1. If $\ln(CD) \leq \ln(\mu_B)\,(\alpha_G - \alpha_B)$, then $k_0 \leq \mu_B$.
2. If $\ln(CD) \geq \ln(\mu_G)\,(\alpha_G - \alpha_B)$, then $k_0 \geq \mu_G$.
3. If $\ln(\mu_B)\,(\alpha_G - \alpha_B) < \ln(CD) < \ln(\mu_G)\,(\alpha_G - \alpha_B)$, then $\mu_B < k_0 < \mu_G$.

Proof:

1. If $\ln(CD) \leq \ln(\mu_B)\,(\alpha_G - \alpha_B)$, then $(CD)^{\frac{1}{\alpha_G - \alpha_B}} \leq \mu_B$.
   Given $k_0 = (CD)^{\frac{1}{\alpha_G - \alpha_B}}$, then $k_0 \leq \mu_B$.
2. If $\ln(CD) \geq \ln(\mu_G)\,(\alpha_G - \alpha_B)$, then $(CD)^{\frac{1}{\alpha_G - \alpha_B}} \geq \mu_G$.
   Given $k_0 = (CD)^{\frac{1}{\alpha_G - \alpha_B}}$, then $k_0 \geq \mu_G$.
3. If $\ln(\mu_B)\,(\alpha_G - \alpha_B) < \ln(CD) < \ln(\mu_G)\,(\alpha_G - \alpha_B)$,
   then $(CD)^{\frac{1}{\alpha_G - \alpha_B}} > \mu_B$ and $(CD)^{\frac{1}{\alpha_G - \alpha_B}} < \mu_G$.
   Therefore, given $k_0 = (CD)^{\frac{1}{\alpha_G - \alpha_B}}$, then $\mu_B < k_0 < \mu_G$. ■

B Extra Simulations

B.1 Case 1: Normal with $\sigma_G = \sigma_B$

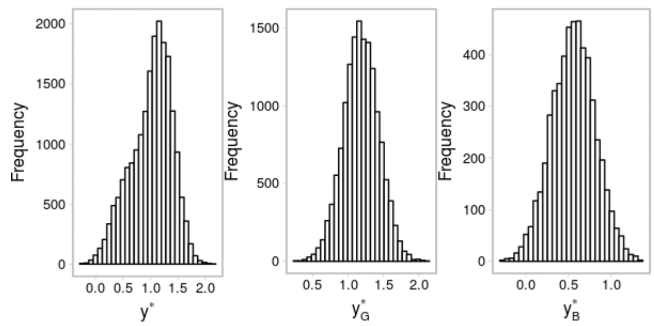

Figure 3: Histograms of $Y_G^* \sim N(\mu_G = 1.1867, \sigma_G^2 = 0.125)$ and $Y_B^* \sim N(\mu_B = 0.5628, \sigma_B^2 = 0.125)$

In the first Normal Distribution case where $\sigma_G = \sigma_B = \sigma$, we have arbitrarily chosen $\sigma = 0.25$. With this relatively small standard deviation, we can see that the distributions of the good and bad populations only overlap in their tails, so the optimal threshold should be able to separate the distributions from each other and we can expect a high TP and low FP. All of our data lie in the interval $[y^*_{min}, y^*_{max}] = [-0.410, 2.077]$. Refer to C.1 for the calculation.

B.2 Case 2: Normal with $\sigma_G \neq \sigma_B$

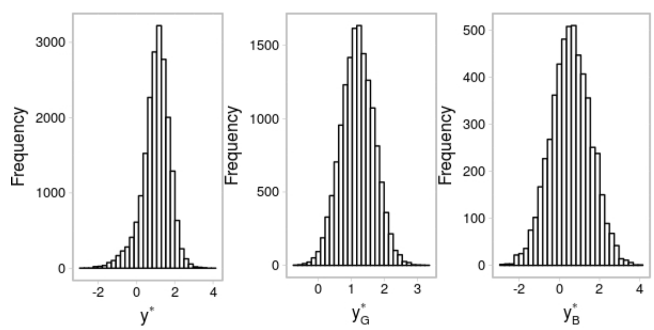

Figure 4: Histograms of $Y_G^* \sim N(\mu_G = 1.1867, \sigma_G^2 = 0.25)$ and $Y_B^* \sim N(\mu_B = 0.5628, \sigma_B = 1.0^2)$

In the second Normal Distribution case where $\sigma_G \neq \sigma_B$, we have arbitrarily chosen $m = 0.5$ and $\sigma = 1.0$, where $\sigma_G = m\sigma = 0.5$ and $\sigma_B = \sigma = 1.0$. Since nearly the entire Good population lies within the upper tail of the Bad population, we do not expect that the optimal threshold to be able to perfectly separate the distributions. There should be high TP and moderately high FP proportions. And all of the data lie in the interval $[y^*_{min}, y^*_{max}] = [-3.340, 4.628]$. Refer to C.2 for the calculation.

B.3 Case 3: χ^2 or Γ with $\beta_G = \beta_B$

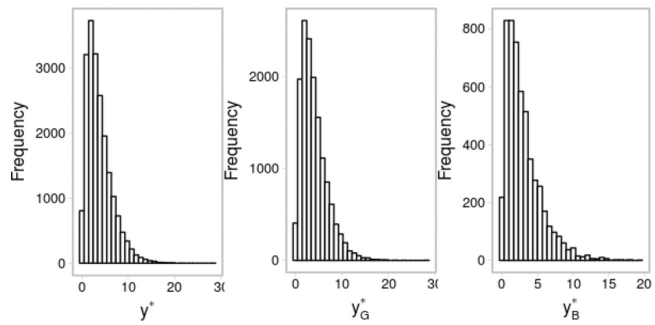

Figure 5: Histograms of $Y^*_G \sim \chi^2(\nu_G = 4)$ and $Y^*_B \sim \chi^2(\nu_B = 3)$

When working with the Γ Distribution case where $\beta_G = \beta_B = 2$, we have a χ^2 Distribution [2]. We have arbitrarily chosen two means, $\nu_G = \mu_G = 4$ and $\nu_B = \mu_B = 3$, such that our assumption that $\mu_G > \mu_B$ is true. With these distributions, we can see that there is great deal of overlap and the threshold will likely be unable to distinguish the two distributions. There should be high TP and FP proportions. All of our data lie in the interval $[y^*_{min}, y^*_{max}] = [0.001, 21.984]$. Refer to C.3 for the calculation.

C Extra Calculations

C.1 Case 1: Normal with $\sigma_G = \sigma_B$

First, we consider the case of two Normal distributions with equal variance, so our optimal threshold is calculated by Thm. 1.

$$\begin{aligned} k_0 &= \frac{\sigma^2 \ln C}{\mu_G - \mu_B} + \frac{\mu_G + \mu_B}{2} \\ &= \frac{0.25^2 \ln 0.361}{1.1867 - 0.5628} + \frac{1.1867 + 0.5628}{2} \\ &= 0.773 \end{aligned}$$

Since $k_0 \in [y^*_{min}, y^*_{max}]$, by Thm. 2, $k_0 = k^*$ and is our optimal threshold.

$$\begin{aligned} R(k^*) &= P_G \cdot TP_{k^*}(r_G + c_G) - P_B \cdot FP_{k^*}(r_B + c_B) + P_B r_B - P_G c_G \\ &= 0.735 \cdot 0.951(1400 + 560) - 0.265 \cdot 0.201(280 + 1680) + 0.265(280) - 0.735(520) \\ &= 928.64 \end{aligned}$$

Evaluating $R(k^*)$, we can see that the profit for this example is \$928.64. Combining our simulated data and calculated threshold, we can see that the calculated threshold value does maximize the profit function.

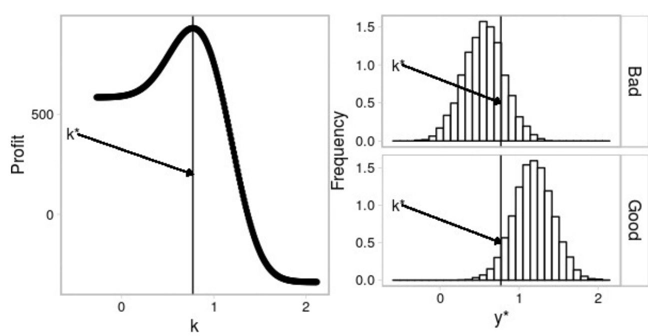

Figure 6: Simulated Data with Optimal Threshold k^* for Case 1, where $\sigma_G = \sigma_B = 0.25$

C.2 Case 2: Normal with $\sigma_G \neq \sigma_B$

Second, we consider the case of two Normal distributions with unequal variance, so our optimal threshold is calculated by Thm. 3.

$$
\begin{aligned}
k_0 &= \frac{(m^2\mu_B - \mu_G) + \sqrt{(\mu_G - m^2\mu_B)^2 - (m^2-1)(m^2\mu_B^2 - \mu_G^2 - 2\sigma^2 m^2 \ln Cm)}}{m^2-1} \\
&= \frac{1}{0.5^2-1}((0.5^2(0.5628) - 1.1867) \\
&+ \sqrt{(1.1867 - 0.5^2(0.5628))^2 - (0.5^2-1)(0.5^2(0.5628^2) - 1.1867^2 - 2^2(0.5^2)\ln(0.361(0.5)))}\ ) \\
&= 0.248
\end{aligned}
$$

Again, since $k_0 \in [y^*_{min}, y^*_{max}]$, by Thm. 4, $k_0 = k^*$ and is our optimal threshold.

$$
\begin{aligned}
R(k^*) &= 0.735 \cdot 0.970(1400 + 560) - 0.265 \cdot 0.624(280 + 1680) + 0.265(280) - 0.735(520) \\
&= 735.78
\end{aligned}
$$

Evaluating $R(k^*)$, we can see that the profit for this example is \$735.78. Combining our simulated data and calculated threshold, we can see that the calculated threshold value does maximize the profit function.

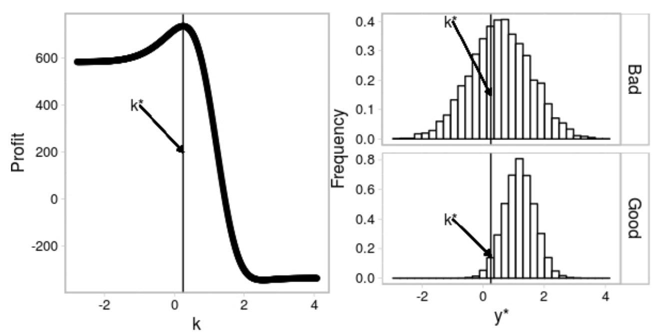

Figure 7: Simulated Data with Optimal Threshold k^* for Case 2, where $\sigma_G = 0.5$ and $\sigma_B = 1.0$

C.3 Case 3: χ^2 or Γ with $\beta_G = \beta_B$

Third, we consider the case of two χ^2 distributions, so our optimal threshold is calculated by Thm. 5.

$$
D = \frac{\Gamma(\frac{\nu_G}{2})}{\Gamma(\frac{\nu_B}{2})} = \frac{\Gamma(\frac{4}{2})}{\Gamma(\frac{3}{2})} = 1.1284
$$

$$k_0 = 2(CD)^{\frac{2}{\nu_G - \nu_B}}$$
$$= 2(.361(1.1284))^{\frac{2}{4-3}}$$
$$= 0.331$$

Again, since $k_0 \in [y^*_{min}, y^*_{max}]$, by Thm. 6, $k_0 = k^*$ and is our optimal threshold.

$$R(k^*) = 0.735 \cdot 0.988(1400 + 560) - 0.265 \cdot 0.954(280 + 1680) + 0.265(280) - 0.735(520)$$
$$= 589.96$$

Evaluating $R(k^*)$, we can see that the profit for this example is \$589.96. Combining our simulated data and calculated threshold, we can see that the calculated threshold value does maximize the profit function.

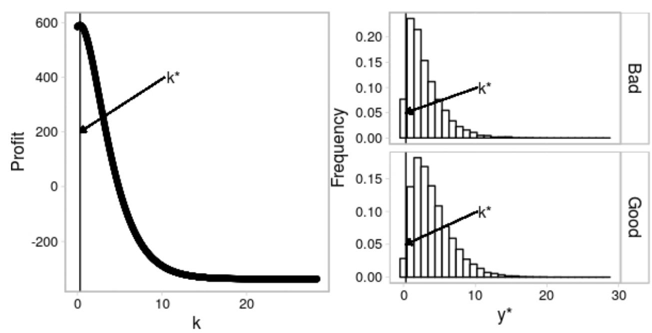

Figure 8: Simulated Data with Optimal Threshold k^* for Case 3, where $\alpha_G = 2$, $\alpha_B = 1.5$, and $\beta_G = \beta_B = 2$