Creating a Framework for Target-Setting Problems of Univariate Data

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Benjamin J. Metzger

Spring 2020

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Richard D. Jacques, Ph.D., Department of Engineering and Society

Abstract

Gambling institutions (lotteries, casinos, online sports betting, etc) generate revenue by offering users bets that have a negative expected return, but usually with a high degree of variability. The high variability creates the illusion for potential gamblers that certain bets have a positive expected return or that previous success is indicative of future success (e.g. "the table is hot") and so on. I seek to establish a framework for determining if there exists an opportunity for a user to ever have a positive expected return for certain types of gambles: optimal "target setting" in the context of univariate data, and I will use one particular game as an example. The game is called "Bustabit" and involves gambling bitcoins.

Introduction

How the Game Works

Users bet a principle amount in bits, where 1 bit is 1 millionth of a bitcoin. When the game begins, there is a multiplier which starts at 1.00 and increases towards infinity. At any point during this increase, users decide when they "cash out," and they get their principle amount back multiplied by whatever the multiplier is at the moment of them "cashing out." Here's the catch: at some point during the multiplier's rise, it will "bust." If the bust occurs before the user cashes out, they lose their principle amount. So, users have to balance (1) waiting for the multiplier to increase with (2) the increasing chance that the round busts and they lose their money

There is one small caveat – if, by some coincidence, you select to cash out at the exact same value as the bust, you still bust. I will discuss how to handle this when discussing the opportunity loss function.

Procedure and Relevance

I seek to identify an opportunity for making a positive expected return, a phenomenon rarely seen in the gambling world. In doing so, I hope to establish a framework for how to go about these problems, using Bustabit as the example. Users may create scripts to play Bustabit for them, and if the analysis yields an indication that there may be an opportunity for a positive expected return, I will create a script.

Analysis begins by determining whether or not observations are independent. Independence means that previous values of our variate will have no bearing on the next realization, and so phenomena such as "the past few multipliers have been small, this next one must be big" have no foundation whatsoever.

If the trials turn out to be dependent, I will model a time series of the multiplier (in fact, the creation of a time series is how we determine dependence). If the trials turn out to be independent, they can be modeled as coming from a common distribution family. In the former case, the optimal target "cash out" value will be a function of previous realizations. In the latter case, the optimal decision will be a constant value. It is possible that the optimal decision would have a negative expected return. In this case, the new optimal decision simply becomes "do not

play." Regardless, the procedure followed in this paper is intended to establish a framework for making optimal decisions in gambles that involve setting a target value based on univariate data.

Note on Data Collection

For analysis to begin, I needed to collect data. A "random sample" is what we always seek to start with, but I found this to not be feasible – there was no way to export historical data, so I was going to have to collect it manually. I discerned that there was no issue with this procedure given my two separate analytic processes. If the trials are truly independent, then a sample of all the observations happening sequentially is just as good as a random sample from historical data. If the trials are not independent, then I would need precisely data from a sequential set of trials. So, collecting the data all at once does not pose any issues to the analyses.

Notation

a – the decision variable. This represents the value at which a user cashes out. The feasible values are from 1.00 to infinity. Note that a user should never select 1.00, because if the target was to get a flat return, one should simply not play.

a^* – the value of a which constitutes an optimal decision. If the trials are dependent, this will be a function of the previous realizations. If the trials are independent, this will be a constant value.

W – the random variate representing the "bust" value for the game. Any realization of this variate is w.

G – the distribution function of W such that for any realization $w \in [1.00, \infty)$,

$$G(w) = P(W \leq w)$$

$\hat{G}$ – an estimation of the distribution function of W.

g – the density function of W. Note that

$$g(w) = \frac{d[G(w)]}{dw}, \qquad or\ G(w) = \int g(w)\, dw$$

N – The total number of realizations. $N = 675$. Any particular value is denoted n.

$l(a, w)$ – the opportunity loss function based on our decision, a, and the realization for the multiplier variate, w.

$L(a)$ – the expected opportunity loss of decision variable a. It is defined as

$$L(a) = E[l(a, W)]$$

In the context of continuous modeling,

$$L(a) = \int_{-\infty}^{\infty} l(a, w) * g(w)\, dw$$

$u(a, w)$ – the utility function based on our decision, a, and the realization for the multiplier variate, w.

$U(a)$ – the expected utility of decision variable a. It is defined as

$$U(a) = E[u(a, W)]$$

Furthermore,

$$U(a) = \int_{-\infty}^{\infty} u(a, w) * g(w)\, dw$$

k – the number of bits invested in the gamble.

Section 1 – Analysis Involving Dependence

1.1 Procedure

For this section, I attempted to model the data as a time series. This data does not constitute a traditional Univariate Time Series, which is made up of observations of a single variate recorded at equal time intervals. This is a traditional stipulation because it makes the math much easier if the time differentials are equidistant. In this situation, the index of a realization is a good proxy for the time variable. From here, analysis is done based on how realizations affect each other (hence "dependence") at different index differentials.

For us, the time between observations varies, and time lag is actually a function of the magnitude of the observations: the multiplier starts at 1 and rises towards infinity, and the larger the end result of w, the longer interval between "rounds" of the game. However, we do not really care about the time involved between observations, we only care in seeing how these observations affect observations at different *index* differentials. Thus, we too will only care about indices and their affects. So, we may utilize the methods of traditional time series. This procedure begins with modeling Stationarity and Seasonality, and then auto-regressive and moving average components are modeled, if applicable.

Once all this is accounted for with a model, we can use that model to forecast 1 future observation. If all has been done correctly and the models are accurate, the only remaining stochastic term will be a random noise term, which is assumed to be normally distributed. This stochastic relationship will constitute a distribution for w_{N+1} (a forecast 1 observation in the future) which can be utilized in an opportunity loss function to find a maximizer

1.2 Assessing Stationarity

For the time series to be stationary, it must exhibit constant location and scale. This basically means that both the mean and variance of the data are generally the same for any sub-set of the data. This can be assessed either visually (by plotting the index on the x-axis and realizations on the y-axis and determining whether or not means and variances seem constant) or mathematically (by actually calculating the mean and variance for different subsets of the data and comparing them).

I found some issues in doing this. Our data had many outliers that threw off the mean by a considerable amount for the intervals that included an outlier. These outliers may hurt he accuracy of any potential models, so I decided that the data needed to be transformed.

The goal of transforming the data is so that it behaves more like a normal distribution, which will fit the assumptions of more kinds of models. Box-Cox transformations are common, and they are what I will use. The formula (Box-Cox) is:

$$w_n(\lambda) = \begin{cases} \frac{w_n^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(w_n), & if\ \lambda = 0 \end{cases}$$

Where $w_n(\lambda)$ is the transformed version of w_n, and this transformation is applied to all values $\{w_n : n = 1, \dots N\}$. Calculation of the parameter λ is outside of the scope of this paper, and R provides functions that allow us to shortcut calculations of the parameter. In short, the **forecast** library in R provides functions for transformations. The function **BoxCox.lambda(data)** will give the optimal value for λ, and the function **BoxCox(data, lambda)** will return a transformed data set. The former function will be the second argument for the latter function, which returns to us a transformed data set more near a normal distribution

I produced a plot of the transformed data and found it to visually fit the criteria of constant mean and variance. I then calculated the means and variances for sets of the transformed data set, and found them to be sufficiently close to one another. So, I will proceed with the rest of this analysis using the transformed data. The stationarity criterion for the model is met.

For the sake of completion, I will briefly discuss how to get the parameters for a trend in R, despite such components not being valid for this particular game. A time series model of w that includes a trend would be:

$$w_n = w_0 + c_1 * n$$

In this relationship, w_0 is the intercept, n is the time variable (which will just be the index in our case), and c_1 is the coefficient for the time or index variable. R will automatically produce values for w_0 and c_1.

1.3 Assessing Seasonality

Seasonality refers to repeating, regular oscillations of data values. In the model, this would be captured in sine or cosine functions. An example of data that would exhibit this characteristic would be a series of the maximum temperature each day over the course of several years: we would expect to see regular oscillations of highs and lows corresponding to the seasons. These can be assessed visually, and I found there to be none.

For the sake of completion of the framework, I will briefly discuss how to incorporate seasonal components. The optimal frequency of the model can be assessed from the Periodogram: it will be equal to the frequency (on the horizontal axis) which maximizes the spectrum (on the vertical axis). Periodograms are created in R with the **spec.pgram** function.

Once you have found the optimal value for the frequency, ω, you should incorporate it into both a sine and cosine function in the following way:

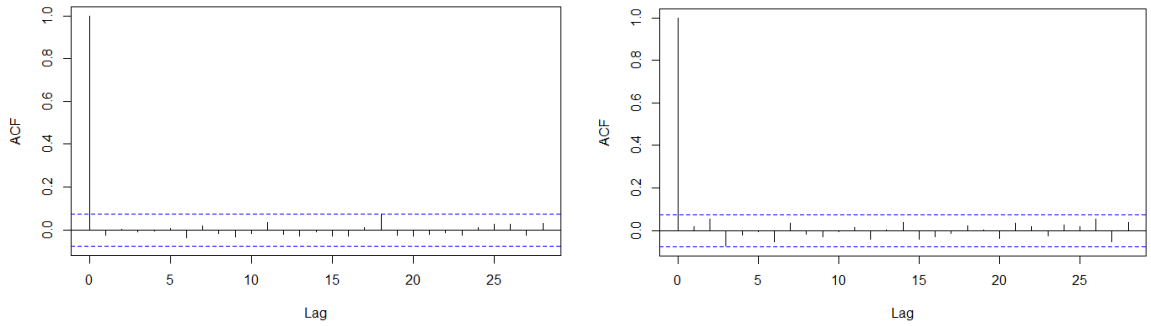
$$w_s = c_1 * \sin(2\pi * \omega * n) + c_2 * \cos(2\pi * \omega * n)$$

Then, to test the validity, create a regression in R inserting the above terms into a model. Running a summary of the model will indicate whether or not each term is significant: the values for c_1, c_2 are calculated in R, and if they are not statistically significant (i.e. different from 0 with 95% confidence), then either can be dropped from the model.

1.4 Determining auto-regression

The goal of time series analysis is to make a model which results in residuals (the difference between the model's value and the recorded value) that are normally distributed. Any sort of relationship between residuals would indicate that there is more modeling to be done. So, once trend and seasonality have been factored into the model, we then look at the resulting *residuals* to determine if auto-correlation exists, as opposed to looking at the initial (or transformed) data.

The validity of an auto-regression or a moving average model is determined from looking at the auto-correlation plot. This plot essentially demonstrates the relationship between observations' residuals as a function of the time (or in our case, the index) lag between them.



The left graph is the ACF of the transformed data's residuals, and the right is the ACF of the untransformed data's residuals. The y-axis is the auto-correlation value between data points, where our data points are now residuals, and the x-axis is the index lag between them. The dashed blue line represents our critical value – the threshold for determining whether or not a value is statistically significant. A realization will always perfectly correlate with itself, hence the value of 1 at index lag zero. The Engineering Statistics Handbook (Box-Jenkins) provides a table for what model we should use based on the shape of the auto-correlation plot. Combining what we see from the table and the plot above, we can gather that there is no relationship between some observation and past observations, regardless of whether or not we have transformed the data

1.5 Conclusion for Bustabit

We have robust evidence that the data are independent, and thus there is no validity in modeling a time series of this data. We will move on to other modeling methodologies. Note

that if we found there to be no auto-regressive or moving average components, but there was still a trend or seasonality involved, we would still have dependence and a time series model would be valid. It just so happens that in this case, we found there to be none of those elements, indicating that data are independent.

1.6 Continuing the framework

For the sake of completion, I will continue to discuss the modeling methodologies for this section regardless of the fact that they are not valid for Bustabit. Most of the analysis for the remainder of this session can be automated leveraging functions in R.

Auto-correlation will take the form of two main components: Auto-Regression and Moving Average, abbreviated as ARMA or the more complicated ARIMA. Modeling of these components, if valid, can be done by looking at the output of **auto.arima(data)** in R, which will give the ideal values of p, q, which represent the *order* of the models. A time series of a residual, ε_n that is $ARMA(p, q)$ will take the form:

$$\varepsilon_n = \phi_1\varepsilon_{n-1} + \phi_2\varepsilon_{n-2} + \cdots + \phi_p\varepsilon_{n-p} + \theta_1\sigma_{n-1} + \theta_2\sigma_{n-2} + \cdots + \theta_q\sigma_{n-q} + \sigma_n$$

where values for ϕ and θ are constants produced in R, ε_{n-1} is the most recent *modeled* residual, and σ_{n-1} is the most recent *actual* residual. As discussed previously, we only model ARMA components on the residuals once trend and seasonality have been calculated. Those modeled residuals are noted as ε. They will not be perfectly accurate, as it is just a model, and so these residuals will have residuals. These are denoted as σ. Note that these models only become valid for when $n > p$ and $n > q$, because indices of zero or lower are not valid for this context.

The model to this point would take the form:

$$w_n = w_t + w_s + \epsilon_n$$

Where w_s is the seasonal component (taking the form of sine or cosine functions), w_t is the trend of the model (found by performing a regression in R), and ε_n is the value of an estimated residual. These residuals have been modeled to relate to one another via an $ARMA(p, q)$ model. Pulling it all together, we get:

$$w_n = w_0 + c_1 * n + c_2 * \sin(2\pi * \omega * n) + c_3 * \cos(2\pi * \omega * n) + \phi_1\varepsilon_{n-1} + \phi_2\varepsilon_{n-2} + \cdots + \phi_p\varepsilon_{n-p} + \theta_1\sigma_{n-1} + \theta_2\sigma_{n-2} + \cdots + \theta_q\sigma_{n-q} + \sigma_n$$

The only stochastic term remaining in this model is σ_n. All other values are constants (calculated in R), an index/time value, previous estimated residuals, or previous actual residuals. If all modeling components are valid, σ_n should be a normally distributed random variable with a mean of zero and a variance of γ^2 (mathematically: $\sigma \sim N(0, \gamma^2)$). The variance for σ can be calculated once we have roughly 30 values.

Because we only have one the one stochastic term with a mean of zero, the mean of w_n will equal $w_0 + c_1 * n + c_2 * \sin(2\pi * \omega * n) + c_3 * \cos(2\pi * \omega * n) + \phi_1\varepsilon_{n-1} + \phi_2\varepsilon_{n-2} + \cdots + \phi_p\varepsilon_{n-p} + \theta_1\sigma_{n-1} + \theta_2\sigma_{n-2} + \cdots + \theta_q\sigma_{n-q}$ (all terms except for σ_n). The variance of w_n will equal the variance of σ_n, again because it is the only stochastic term.

1.7 Finding a^*

From the last section, we know the mean and variance of w_n, and we know that it is of the normal distribution family. We can also see that it relies on past values and coefficients (approximated from past values). We can think of w_n as being the *next* realization, where all other indices $< n$ are the *past* observations.

The optimal target value, a^*, can be thought of as the value that either (1) minimizes the loss or (2) maximizes the utility. Finding a^* is done in the same way as sections 2.6-2.9, just with a different distribution family: The Normal Distribution, with a mean and variance specified in the previous section. Information regarding the Normal Distribution and its approximation is taken from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 2, Pages 14-15).

Note that a^* will need to be calculated and updated for every single time the game is played, keeping track of past values and past residuals. How to keep track of this data and update it can be done any number of ways, and so a discussion on how to apply the above formulas into code has been omitted. Also note that if a transformation has been done on the data, you will need to apply the inverse of that transformation to your calculated a^*.

Section 2 – Analysis Assuming Independence

2.1 Procedure

This analysis proceeds in 6 steps. The first step is to create an "empirical distribution" resulting from the data sample. The next step is to hypothesize parametric models for the distribution based on the sample space of the data and the shape of the density functions. Then the parameters for each type of hypothesized model need to be estimated, and the best one will be selected based on a criterion discussed later. Once the distribution is modeled, the next step is to create an expected opportunity loss function of any particular decision, a, based on the most fitting parametric distribution. Finally, use the opportunity loss function to find an optimal target, or "cash out" value.

This distribution has a nuance that I will discuss before diving into the analysis: it is neither purely discrete nor continuous. The smallest increment, for both the cash out and bust values, is 0.01. Having a set increment is a characteristic of a discrete distribution. However, a continuous distribution can be thought of as an infinite number of discrete, numerical, sequential, equidistant outcomes, with the increment approaching zero. I decided that 0.01 was a small enough increment, and so I went ahead with the modeling as if it were continuous.

2.2 Creating an Empirical distribution

I followed the procedure for creating an empirical distribution from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 3, Pages 2-3), with some notation changes. I leveraged excel to handle these iterative calculations, exact details on this have been omitted. It proceeds like this, when given a sample of N realizations of continuous variate W:

$$\{w(n): n = 1, \dots, N\}$$

Step 1: Arrange realizations into ascending order:

$$w_{(1)} \leq w_{(2)} \leq \cdots \leq w_{(N)}$$

Note that $w(n)$ refers to the n-th observation, where $w_{(n)}$ refers to the n-th *smallest* observation.

Step 2: Create N non-exceedance probabilities, or plotting positions:

$$p_n = P\left(W \leq w_{(n)}\right), \quad n = 1, \dots, N$$

These will create an increasing sequence.

Step 3: Calculate the plotting positions from Step 2. I have done this using the *Meta-Gaussian* plotting positions, which are defined in Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 3, Page 3) as:

$$p_n = \left[\left(\frac{N - n + 1}{n}\right)^{t_N} + 1\right]^{-1}$$

$$t_N = 1.9574 * N^{-0.8039} + 1$$

Derivation of these plotting positions are outside the scope of this paper. Note that there are several different styles of creating plotting positions, which can be found in the Appendix. I chose to do this one during the first wave of analysis. After calculating MAD values (the criterion for fitness of a model, discussed later), I went back and changed plotting position styles. MAD values across the board increased, indicating that these plotting positions are the most accurate for modeling in this situation.

Step 4: Pulling it all together, the empirical distribution is defined as:

$$\{(w_{(n)}, p_n): n = 1, \dots N\}$$

The format of an empirical distribution is thus an ordered set of points, where $w_{(n)}$ represents some threshold for our variate, and p_n represents the probability that the variate does not exceed that threshold.

2.3 Hypothesizing Parametric Models

I selected parametric models to evaluate based on (1) the sample space of the variate (in our case, we clearly have a lower bound with no upper bound) and (2) the shape of the density functions. Regarding the latter, I created a histogram of the observations of W with very small binning. A histogram with many small, equidistant binning can be thought of as an empirical distribution function. So, I was able to compare my histogram of observations with the density functions found in Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 19-33). Meeting the two criteria of sample space and shape were four distributions: **exponential, Weibull, Inverted Weibull, and Log Logistic**. Information about each of these distributions was also taken from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 4-12) and can be found in the Appendix of this paper, including graphs of the density functions which I compared to my histogram of realizations. The histogram can also be found in the appendix.

The lower bound posed some issues in the analysis. Some steps involve the natural log of [an observation minus the lower bound]. The lower bound is 1.00, and several observations were also 1.00. Given that the natural log of zero does not exist, we must have some way of handling these cases. This issue arises because for a continuous distribution, an observation should never touch the lower bound; however, as discussed, this distribution is not purely continuous. I saw that there were two feasible options of handling this.

The first option was to shift the lower bound from 1.00 to 0.99. The new lower bound is not within the sample space of the multiplier, and so we effectively remove the “natural log of zero” issue. The trade-off is that we have changed a parameter of the distribution, which may hurt the accuracy.

The second option was to create a piecewise distribution of the multiplier. This allows us to separately handle the issue posed by our observations of 1.00. More specifically, the distribution function would take the form:

$$G(w) = \begin{cases} P_{1.00} & for\ w = 1.00 \\ (1 - P_{1.00})H(w) & for\ w > 1.00 \end{cases}$$

where $H(w)$ is a hypothesized distribution of W based on estimation of the remaining data points (i.e. for $w_n \neq 1.00$), and $P_{1.00} = P[W = 1.00]$, which is the number of realizations of $w_n = 1.00$ divided by the total number of realizations. Thus, for each hypothesized model, I would treat it twice – once for each method of handling the 1.00 issue. In the former case, $N = 675$, and in the latter case, $N = 656$ (there were 19 occurrences of $w = 1.00$).

2.4 Estimating Parameters

I will use the *Least Squares* Algorithm to estimate parameters, taken from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix B, Pages 1-4). This algorithm requires an empirical distribution to begin, which was the basis of the previous section. Note that this algorithm had to be applied twice for each distribution because we treated each one twice to handle the natural log of zero issue, as noted previously. Thus, this was done 8 times in total. As before, I leveraged excel to do these iterative calculations, and exact details have been omitted.

Step 1: Transform the empirical distribution from $\{(w_{(n)}, p_n): n = 1, \dots, N\}$ to $\{(v_n, u_n): n = 1, \dots, N\}$. Essentially, we transform an empirical distribution to another set of points. The transformations from w to v and from p to u are specified in the appendix for each hypothesized distribution.

Step 2: Calculate the estimates $\hat{a}, \hat{b}$ of the coefficients a, b, which are intermediates in our parameter estimation. They appear in the linearized quantile function, which again are specified for each relevant distribution in the Appendix. There are two cases for this that apply to the distributions I worked with:

Case 1: When the linearized quantile function takes the form $v = bu + a$:

$$\hat{b} = \frac{\sum_{n=1}^{N} v_n u_n - N\bar{v}\bar{u}}{\sum_{n=1}^{N} u_n^2 - N\bar{u}^2}, \qquad \hat{a} = \bar{v} - \hat{b}\bar{u}$$

Case 2: When the linearized quantile function takes the form $v = u + a$:

$$\hat{a} = \bar{v} - \bar{u}$$

Note: $\bar{v}$ and $\bar{u}$ are the arithmetic means of v and u, respectively.

Step 3: Transform $\hat{a}, \hat{b}$ into $\hat{\alpha}, \hat{\beta}$. These transformations are specified for each relevant distribution in the appendix.

2.5 Selecting the fittest model

Once $\hat{\alpha}, \hat{\beta}$ have been calculated for each hypothesized model, we can calculate $\hat{G}(w_{(n)})$ for $n = 1, \dots, N$. We will now have three relevant columns: the threshold points, $w_{(n)}$; the *empirical* probability that the multiplier, W, will not exceed this threshold point, denoted p_n; and the *estimated* probability that the multiplier, W, will not exceed the threshold point, denoted

$\hat{G}(w_{(n)})$. The methodology of this section is to compare the empirical probabilities to the estimated probabilities to see which estimation provides the best fit. Recall that there were 8 estimations: 4 distribution families, each of which had to be applied twice to match the two different options of handling the "natural log of zero" issue discussed before.

I decided to use the *maximum absolute difference* (MAD) as my criterion to select the best model. The MAD is defined as:

$$MAD = \max_{1 \le n \le N} \left| p_n - \hat{G}(w_{(n)}) \right|$$

In words, the MAD is the maximum of the absolute differences between the empirical distribution and a hypothesized parametric distribution, evaluated at all threshold points. The results yielded that the fittest model, by this criterion, was the Log-Logistic model, treated with a lower bound of 0.99. It had a MAD of 0.019, far lower than the next lowest. Thus, the Log-Logistic distribution family is what I used for the remainder of this section. The values for $\hat{\alpha}, \hat{\beta}$ came out to be 0.889911 and 0.990791, respectively.

2.6 Creating the Opportunity Loss Function

The methodology here is to create an opportunity loss function, which is a function of w and a, and then find the value for a that minimizes the expected opportunity loss, which is $L(a)$. That optimal value for a is denoted a^*. The loss function will be piecewise for the two different outcomes of the game.

The first case is that $w \le a$. In this situation, you are losing your total amount invested, which is k. As noted previously, if you try to cash out at $w = a$, you still "bust." So, the best you can ever do is select $a = w - 0.01$. Because opportunity loss represents the best a user could've done, the opportunity loss is $[w - 0.01] * k$.

The second case is that $a \le w$. In this situation, your opportunity loss arises from the fact that you could've "held out" longer and realized a greater multiplier (in other words, you could've had an a that was closer to w). Recall that the best possible outcome is $a = w - 0.01$, and so the opportunity loss is $[w - a - 0.01] * k$

Pulling this all together:

$$l(a, w) = \begin{cases} [w - 0.01] * k, & w \le a \\ [w - a - 0.01] * k, & a < w \end{cases}$$

The next step is to find the value of a that minimizes the expected opportunity loss. So, we must first derive the expected opportunity loss function. The formula is given for this in the notation section: $L(a) = \int_{-\infty}^{\infty} l(a, w) * g(w)\, dw$. In short, our expected opportunity loss for some selection, a, is equal to the opportunity loss at some realization w of W, multiplied by the probability that $W = w$, integrated over the whole sample space of the variate. This must be applied to each piece of the piecewise function.

Simplification Step 0:

$$L(a) = \int_{-\infty}^{a} [w - 0.01] * k * g(w)\, dw + \int_{a}^{\infty} [w - a - 0.01] * k * g(w)\, dw$$

Notes on the simplification between Steps 0 and 1. The constant value of k appears in both integrals, and so I will factor it out. Also, the lower bound of negative infinity is just representative of the lowest feasible bound. In our case, that value is actually 0.99, and so I made that substitution as well. Lastly, I expanded (or "distributed") the terms of each Integral so that I could evaluate integrals separately.

Simplification Step 1:

$$L(a) = k * \left[\int_{0.99}^{a} w * g(w)\, dw - \int_{0.99}^{a} 0.01 * g(w)\, dw + \int_{a}^{\infty} w * g(w)\, dw - \int_{a}^{\infty} a * g(w)\, dw - \int_{a}^{\infty} 0.01 * g(w)\, dw\right]$$

Notes on simplification between steps 1 and 2. If two integrals have the same *argument*, and *continuous bounds* (i.e. the upper bound of one integral is the lower bound of the other), then you can simplify the integrals to become an integral of the original argument, with bounds from the lower-lower bound to the upper-upper bound. We have two separate cases of this here – the first and third integrals, and the second and the fifth integrals. Executing this simplification:

Simplification Step 2:

$$L(a) = k * \left[\int_{0.99}^{\infty} w * g(w)\, dw - \int_{a}^{\infty} a * g(w)\, dw - \int_{0.99}^{\infty} 0.01 * g(w)\, dw\right]$$

Notes on simplification between steps 2 and 3. The first integral is the very definition of the expected value of w. Regarding the latter 2 integrals, a and 0.01 are both constants with respect to what we are integrating (i.e. w), and so they may be factored out of the integrals.

Simplification Step 3:

$$L(a) = k * \left[E[W] - a \int_{a}^{\infty} g(w)\, dw - 0.01 \int_{0.99}^{\infty} g(w)\, dw\right]$$

Notes on the simplification between steps 3 and 4. Regarding both the remaining integrals, by definition, $G(w) = \int g(w)\, dw$, and since we have bounds, they simplify to the following:

Simplification Step 4:

$$L(a) = k * \left[E[W] - a * [G(\infty) - G(a)] - 0.01 * [G(\infty) - G(0.99)]\right]$$

Notes on the simplification between steps 4 and 5. $G(w)$ is shorthand for $P(W \leq w)$. So, $G(\infty) = P(W \leq \infty)$, which can be reasoned to equal 1. Additionally, $G(0.99) = P(W \leq 0.99)$, and because 0.99 is not in the sample space of W, this can be reasoned to equal zero. Substituting these values in:

Simplification Step 5:

$$L(a) = k * \left[E[W] - a * [1 - G(a)] - 0.01 * [1]\right]$$

Simplification Step 6: (distribution of values from Step 5).

$$L(a) = k * [E[W] - a + a * G(a) - 0.01]$$

The equation above represents the expected opportunity loss of any selected target point, a^*.

2.7 Finding a^*

Now that we have the function for the expected opportunity loss for any selected value of a, we need to find the value that will *minimize the loss*, and this will be denoted as a^*. I followed the usual procedure for finding local extrema: (1) take the derivative with respect to some variable, (2) set the resulting function equal to zero, and then (3) solve for the value of our variable that satisfies the equation. Taking the derivative of $L(a)$, which is Simplification Step 6, we leverage both the product rule of derivation and the relationship between $g(w)$ and $G(w)$ in that $g(w) = \frac{d[G(w)]}{dw}$

Optimization Step 1:

$$\frac{d[L(a^*)]}{da} = 0 = k * [-1 + G(a^*) + a^* * g(a^*)]$$

Optimization Step 2: Divide both sides of the equation by k, leaving us with:

$$0 = G(a^*) + a^* * g(a^*) - 1$$

Our goal is to find the value(s) of a^* which satisfy this equation. Note that this is extremely difficult, if not impossible, to do algebraically: when substituting in for our functions $g(w)$ and $G(w)$ (recall that they are of the Log-Logistic Family), the resulting relationship becomes:

Equation 10

$$0 = \left[1 + \left(\frac{a^* - 0.99}{0.889911}\right)^{-0.990791}\right]^{-1} + a^* \frac{0.990791}{0.889911}\left(\frac{a^* - 0.99}{0.889911}\right)^{-1.990791}\left[1 + \left(\frac{a^* - 0.99}{0.889911}\right)^{-0.990791}\right]^{-2} - 1$$

However, it works in our favor that extreme precision is unnecessary – as mentioned before, the feasible values of a are truly in increments of 0.01. So, what I decided to do was make an approximation. Again, leveraging excel and omitting most details, I made a column starting at 1 and incrementing by 0.01. In another column, I applied equation 10. I then found the value in the equation 10 column eclipsing zero.

2.8 Results

My findings were telling. The value for a in which Equation 10 eclipses zero was at $a^* = 8.905$ (Equation 10 was positive and $a = 8.90$, and negative and $a = 8.91$). Recall that equation 10 represents the *derivative* of the expected opportunity loss function. We want a value for a that *minimizes* the expected opportunity loss, and for this to be the case, the derivative would need to go from negative to positive at a^*. However, I found the opposite to be true, indicating that 8.91 is actually a local *maximizer* of opportunity loss, not a minimizer.

This means that there is no *local* minimizer for the opportunity loss, but that result alone isn't enough information to determine if we should play the game or not. What if, by chance, the expected opportunity loss function starts as negative, but increases, crosses from negative to positive, and then maxes out at around 8.91? Our derivative procedure would fail to catch this initial negative interval. Thus, we need to evaluate the expected opportunity loss function, $L(a)$, at all values of a to determine if the condition $[L(a) < 0]$ ever occurs. Recall that a negative loss would indicate a gain.

Complications arise because our derived equation for the expected opportunity loss (Simplification Step 6) has a value which does not exist: $E[W]$. This is because the $E[W]$ for our distribution (Log-Logistic) contains gamma functions, which are undefined if the argument is less than or equal to zero. For this to be avoided, β must be strictly greater than 1 (refer to Derivation 1 in the Appendix), but our value for β is 0.990791. Further discussion on gamma functions is outside the scope of this project.

So, to fully determine if the game is worth playing or not, we will need to do further analysis.

2.9 Further Analysis – The Utility Function

We will shift gears to the utility function. The goal here is the same as always – to determine if there exist any decisions, a, for which there is a positive expected return. The previous sections involved hunting for opportunities for a negative expected opportunity loss; however, an equivalent approach would be to look for opportunities for a positive expected utility. We will go through some similar steps as of that with the opportunity loss. The utility function will also be piecewise to match the two possible cases.

The first case reflects a "bust" for the user. This will occur whenever $w \leq a$. In this case, the user loses their whole investment amount, k. So, the return is negative k.

The second case is for when the user successfully "cashes out." In this situation, $w > a$. The utility in this situation is $k * [a - 1]$. The value $[a - 1]$ is used instead of a because the return on the multiplier is total winnings minus the invested amount, which is $ka - k$, which simplifies to $k[a - 1]$.

Pulling this all together:

$$u(a, w) = \begin{cases} -k, & w \leq a \\ k[a - 1], & a < w \end{cases}$$

The evaluation for the expected utility will involve many of the same formulae and relationships as the evaluation of the expected opportunity loss from before, and so descriptions of the simplification have been omitted. Using the expression of $U(a)$ from the notation section above:

$$U(a) = \int_{-\infty}^{\infty} u(a,w) * g(w)\, dw$$

$$U(a) = \int_{-0.99}^{\infty} u(a,w) * g(w)\, dw$$

$$U(a) = \int_{-0.99}^{a} u(a,w) * g(w)\, dw + \int_{a}^{\infty} u(a,w) * g(w)\, dw$$

$$U(a) = \int_{-0.99}^{a} -k * g(w)\, dw + \int_{a}^{\infty} k * [a-1] * g(w)\, dw$$

$$U(a) = -k * [G(a) - G(0.99)] + k[a-1] * [G(\infty) - G(a)]$$

$$U(a) = -k * [G(a) - 0] + k[a-1] * [1 - G(a)]$$

$$U(a) = -k * G(a) + ka - ka * G(a) - k + k * G(a)$$

$$U(a) = ka - ka * G(a) - k$$

To make the analysis easier, I will use a value of $k = 1$. The investment is an arbitrary amount set by the user, and it is in every term of the expected utility function at a power of 1, so it may be factored out. You may think of the investment amount as a simple multiplier of the utility.

Derived Expected Utility Function:

$$U(a) = a - a * G(a) - 1$$

Taking the derivative of the Expected Utility Function should produce the same result as the Expected Opportunity Loss Function. I will verify that here, again omitting descriptions because this section is similar to the expected opportunity loss analysis.

$$\frac{d[U(a)]}{da} = 0 = 1 - G(a) - a * g(a)$$

If we were to multiply both sides of the above equation by negative 1, we would get the same result as Optimization Step 2. Thus, we know that the procedures are equivalent. Substituting in for $G(a)$ in the Derived Expected Utility Function:

$$U(a) = a - a * \left[1 + \left(\frac{a - 0.99}{0.889911}\right)^{-0.990791}\right]^{-1} - 1$$

The difference here is that the expected utility function exists, whereas the expected opportunity loss did not. I employed excel to calculate the utility for values of our decision

variable, a, starting at 1.00 and incrementing by 0.01 towards infinity. I then applied the above equation in a separate column.

Conclusion

The results from the previous section were that for all values of a, the expected utility of the gamble was always less than zero. This is an indication that there does not exist a value for our decision variable that makes the gamble in the best interest of the user. If the trials are truly all independent (which I have reason to believe from the first section), and you seek to maximize your utility, you should simply not play the game. If there were positive values for $U(a)$, the player should use the maximum, which can be found by applying the same procedure as the *Finding* a^* section, with the only difference being that the Utility function would be used instead of the opportunity loss function.

Appendix

The Normal Distribution

All of the following information is retrieved from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 2, Pages 15).

Density Function

$$g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Where, for our purposes,

$$z = \frac{w - E(w_n)}{\gamma}$$

And γ is the standard deviation of the $\sigma_n$variable (refer to section 1.6)

Distribution Function: Must be approximated in the following way:

If $0 \leq z < \infty$

$$\hat{G}(z) = 1 - \frac{1}{2}(1 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4)^{-4}$$

$$a_1 = 0.196854, \qquad a_3 = 0.000344$$

$$a_2 = 0.115194, \qquad a_4 = 0.019527$$

If $-\infty < z \leq 0$, then $\hat{G}(z) = 1 - \hat{G}(-z)$

Quantile Function: Must be approximated in the following way:

If $0.5 \leq p < 1$, then

$$\hat{z}_p = t - \frac{a_0 + a_1 t}{1 + b_1 t + b_2 t^2}$$

$$t = \sqrt{-2\ln(1-p)}$$

$$a_0 = 2.30753, \qquad b_1 = 0.99229$$

$$a_1 = 0.27061, \qquad b_2 = 0.04481$$

If $0 < p \leq 0.5$, then $\hat{z}_p = -\hat{z}_{1-p}$

Other Plotting Positions

Other styles of plotting positions exist in addition to the plotting positions I used in section 2.2. Advantages exist between the different methodologies, and as discussed earlier, I found the ones I used to be the most appropriate based on the MAD values. The others, taken from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 3, Page 3) are:

The Standard Plotting Positions:

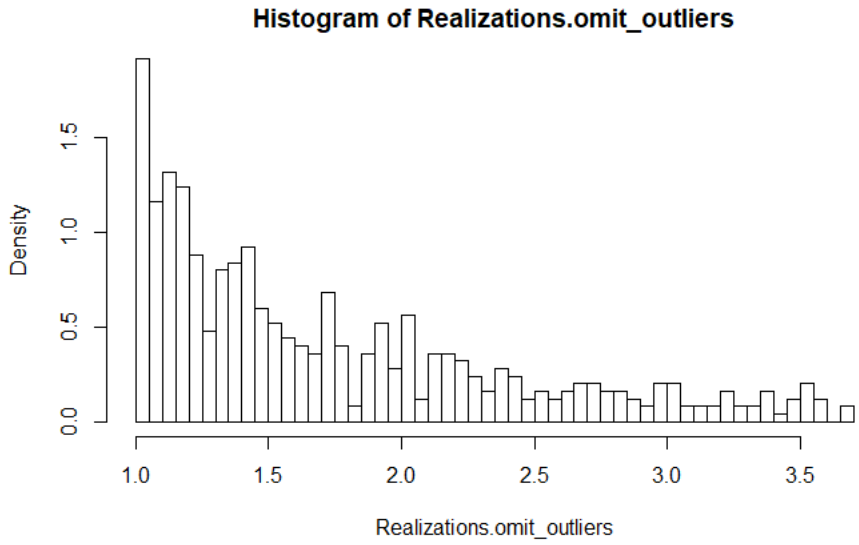
$$p_n = \frac{n}{N}$$

The Weibull Plotting Positions:

$$p_n = \frac{n}{N+1}$$

Histogram of Data

The histogram of data: Note that I omitted the right-tailed outliers of the data to make the visualization better – the shape of the distribution becomes more obvious as you "zoom in."



The Exponential Distribution

All of the following information is retrieved from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 8, 23).

Density Function

$$g(w) = \frac{1}{\alpha}\exp\left(-\frac{w-\eta}{\alpha}\right)$$

Distribution Function $p = G(w)$

$$G(w) = 1 - \exp\left(-\frac{w-\eta}{\alpha}\right)$$

Linearized Quantile Function $v = u + a$

$$v = \ln(w - \eta)\,, \qquad u = \ln[-ln[1-p]]$$

$$\alpha = \exp(a)$$

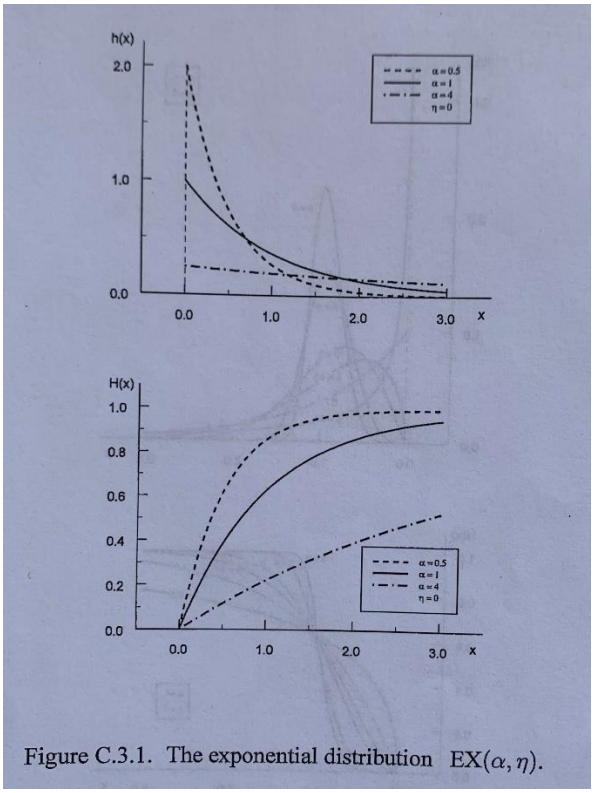


Figure C.3.1. The exponential distribution $EX(\alpha, \eta)$.

The Weibull Distribution

All of the following information is retrieved from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 9, 24).

Density Function

$$g(w) = \frac{\beta}{\alpha}\left(\frac{w-\eta}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{w-\eta}{\alpha}\right)^{\beta}\right]$$

Distribution Function $p = G(w)$

$$G(w) = 1 - \exp\left[-\left(\frac{w-\eta}{\alpha}\right)^{\beta}\right]$$

Linearized Quantile Function $v = u + a$

$$v = \ln(w - \eta), \qquad u = \ln[-ln[1-p]]$$

$$\beta = \frac{1}{b} \qquad \alpha = \exp(a)$$

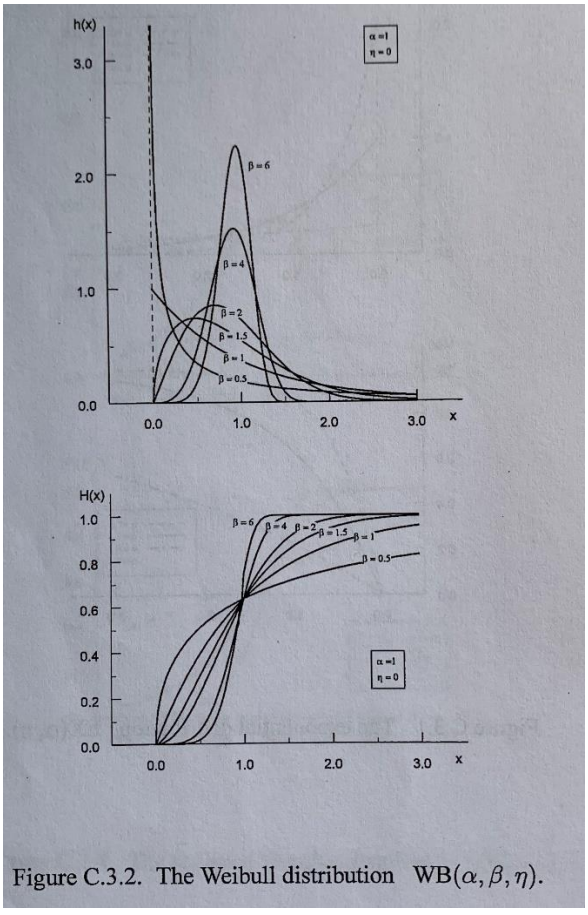


Figure C.3.2. The Weibull distribution $WB(\alpha, \beta, \eta)$.

The Inverted Weibull Distribution

All of the following information is retrieved from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 10, 25).

Density Function

$$g(w) = \frac{\beta}{\alpha}\left(\frac{\alpha}{w-\eta}\right)^{\beta+1} \exp\left[-\left(\frac{\alpha}{w-\eta}\right)^{\beta}\right]$$

Distribution Function $p = G(w)$

$$G(w) = \exp\left[-\left(\frac{\alpha}{w-\eta}\right)^{\beta}\right]$$

Linearized Quantile Function $v = u + a$

$$v = \ln(w - \eta), \qquad u = -\ln(-ln(p))$$

$$\beta = \frac{1}{b} \qquad\qquad \alpha = \exp(a)$$

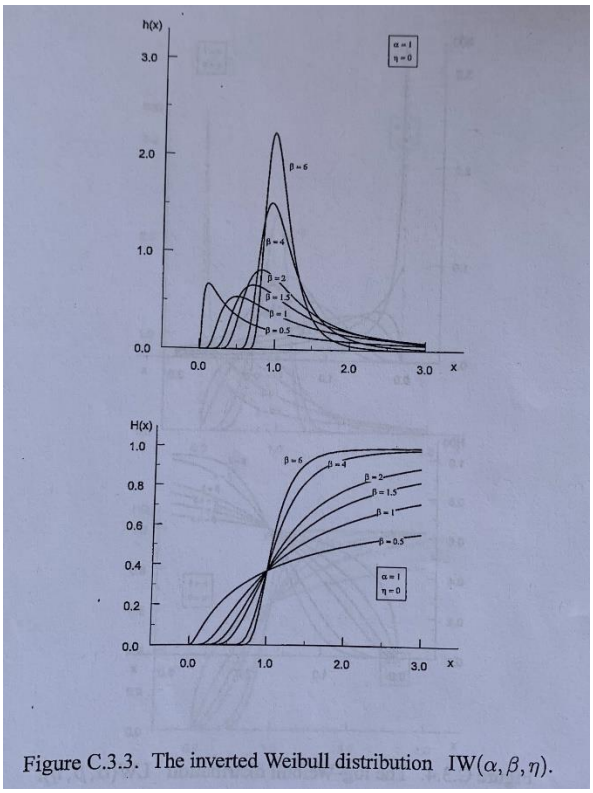


Figure C.3.3. The inverted Weibull distribution $\mathrm{IW}(\alpha, \beta, \eta)$.

The Log-Logistic Distribution

All of the following information is retrieved from Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Appendix C, Pages 12, 27).

Density Function

$$g(w) = \frac{\beta}{\alpha}\left(\frac{w-\eta}{\alpha}\right)^{-\beta-1}\left[1+\left(\frac{w-\eta}{\alpha}\right)^{-\beta}\right]^{-2}$$

Distribution Function $p = G(w)$

$$G(w) = \left[1+\left(\frac{w-\eta}{\alpha}\right)^{-\beta}\right]^{-1}$$

Linearized Quantile Function $v = u + a$

$$v = \ln(w-\eta), \qquad u = \ln\left(\frac{p}{1-p}\right)$$

$$\beta = \frac{1}{b} \qquad\qquad \alpha = \exp(a)$$

$$E[W] = \alpha * \Gamma\left(1+\frac{1}{\beta}\right) * \Gamma\left(1-\frac{1}{\beta}\right) + \eta$$

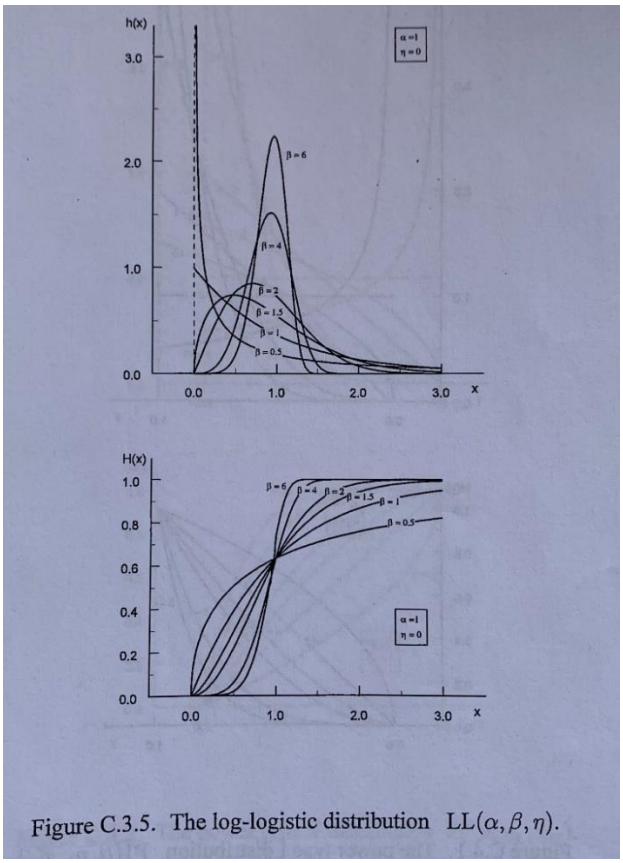


Figure C.3.5. The log-logistic distribution $\mathrm{LL}(\alpha, \beta, \eta)$.

Derivation 1

Demonstrating why, for the Log-Logistic Distribution, $E[W]$ only exists when $\beta > 1$

The gamma function as defined in Section 2.7 of Probabilistic Forecasts and Optimal Decisions (Krzysztofowicz, Chapter 2, Page 26):

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt\,, \qquad x > 0$$

$$E[W] = \alpha * \Gamma\left(1 + \frac{1}{\beta}\right) * \Gamma\left(1 - \frac{1}{\beta}\right) + \eta$$

This $E[W]$ is only applicable for the log-logistic distribution, and it contains two gamma functions. We know that $\beta > 0$ by definition, and so the first gamma function will have a positive argument for any value of β. The second gamma function needs a positive argument as well. Mathematically:

$$1 - \frac{1}{\beta} > 0$$

Adding $\frac{1}{\beta}$ to both sides of the equality:

$$1 > \frac{1}{\beta}$$

Multiplying both sides of the inequality by β:

$$\beta > 1$$

Thus, $E[W]$ exists if and only if $\beta > 1$.

References

"Box Cox Transformation." *Statistics How To*, 20 May 2018, www.statisticshowto.com/box-cox-transformation/.

"Box-Jenkins Model Identification." *Engineering Statistics Handbook*, NIST Sematech, www.itl.nist.gov/div898/handbook/pmc/section4/pmc446.htm.

Krzysztofowicz, Roman. *Probabilistic Forecasts and Optimal Decisions*. University of Virginia, 2019.