Optimization of Inventory Allocation

Introduction

In 2016, e-commerce sales totaled an estimated $394.9 billion, accounting for 8.1 percent of total annual sales. This total was a 15 percent increase from 2015. Advances in technology and adoption of the internet has forced the retail industry to make dramatic shifts toward e-commerce. While this presents a tremendous opportunity for business growth, the cost associated with inefficiencies in supply chains makes optimally allocating inventory to fulfillment centers integral to retailers' success. Customer retention and loyalty requires the business to deliver products quickly and efficiently, so inventory allocation to fulfillment centers must account for two primary factors:

- item cost of delivering units from fulfillment centers to delivery addresses
- length of time required to fulfill orders to delivery addresses from fulfillment centers

While the first of these costs can be associated with a particular dollar value, the cost related to the second consideration is driven by customer disatisfaction with delayed product arrivals and is more difficult to quantify. Additionally, there are a host of other factors to be considered when optimizing inventory allocation. These include the cost of transferring units between fulfillment centers to respond to shifts in demand, as well as decisions about whether to fulfill online orders from stores instead of dedicated fulfillment centers. In the following sections, we propose a method to determine the proper allocation of inventory to fulfillment centers. Our method assumes that no transfers are made between fulfillment centers and that e-commerce orders are fulfilled only from dedicated fulfillment centers and no orders will be fulfilled by stores, however, potential future work could be explored by relaxing these assumptions.

[Typical Solution] How they work:

There are several products on the market for supply chain optimization. Many of those products are outfitted with the functionality for determining efficient allocation of inventory to distribution centers. We can decompose the common approach to optimizing inventory allocation into two primary tasks: predicting future inventory demand and then given this forecast, develop a plan to purchase and distribute products so as to meet the demands at each fulfillment center in the most cost efficient way possible.

Demand predictions are made by fitting statistical models using past demand as well as predicted future events to forecast future demand. Often, these forecasts are done using data that has been aggregated to a fairly high level, such as by week, subpopulation of customers, category of merchandise, etc, though new technical ability to store and access large amounts of data quickly allows demand forecasting to be done at a far more granular level, such as at the daily level, per item per customer. Safety stock is the extra stock kept on hand to mitigate rish of inventory depletion due to unpredictability in demand.

The *service level* refers to the probability of the period-specific demand quantity exceeding the inventory quantity on hand:

$$Prob\left(\text{period demand} \leq \text{available inventory stock}\right).$$

Setting safety stock levels appropriately is critical to minimizing costs. Typically, the amount of safety stock is chosen so that the probability of not meeting customer demand is low, based on the predicted sales. Loss is incurred due to incorrect forecasts by when

- demand is overestimated, leading to excess inventory which will be wasted or discounted, or when
- demand is underestimated, resulting in missed sales opportunities and unsatisfactory customer experience.

Once they specify the service level and forecast sales, retailers determine the safety stock levels and the corresponding product *buy b* - the total amount of product to purchase for order fulfillment. Determining how this inventory is distributed is then a deterministic task - there is little to no uncertainty associated with which fulfillment center is most optimal for shipping a given order. The center that will fulfill each online order is determined using a set of rules which specify the logic necessary for minimizing

- the cost of shipping orders from fulfillment centers to delivery locations and
- the length of time required for products to be delivered to the customer.

Using the set of rules that determines which fulfillment center will fulfill orders from households in which regions, the task of allocating inventory is simply a matter of executing some straighforward bookkeeping. We do not seek to optimize this set of rules. Instead, we aim to address the difficulty in choosing an optimal allocation presented by the uncertainty in the forecasted demand for each fulfillment center territory.

Typical Solutions: Why they suck

Despite the numerous existing applications offering solutions to the allocation problem, inventory allocation inefficiencies are still quite prevelant across retail supply chains. The difficulty in accounting for demand forecast uncertainty is the root of the underperformance of many tools on the market. Many of these tools overly simplify this uncertainty in various ways. Many rely on rigid assumptions about the distribution of future product demand for a given fulfillment center. For example, many universally accepted methods assume that safety stock is proportional to the standard deviation of the demand, which is assumed to follow a Normal distribution. It is also common in the existing market for software solutions to assume that the demand at a given fulfillment center is independent from week to week, i.e. that the demand in the current week contains no useful information for predicting demand for the week to follow. In most practical situations, this assumption is invalid due to the innate dependency between time and demand. The classic example is the trendsetter: she purchases an item and after wearing it, the crowd follows suit and demand for a given product increases until it reaches a certain point at which the market is saturated. Invalid assumptions about the variability in demand inhibit the ability to adequately quantify the uncertainty with predicted demand. Characterizing this variability faithfully is integral to specifying the initial product buy and allocation to each fulfillment center.

Even when all the assumptions of a forecasting model can be verified to be reasonable, there is still uncertainty associated with both future predicted demand as well as the model we use to make these predictions. Most existing inventory allocation tools determine the best initial allocation of inventory to fulfillment centers by assuming that this prediction error doesn't exist. Unfortunately, the allocation quantities inherit the uncertainty in the demand forecast used to determine these quantities. Failure to account for the uncertainty can lead to suboptimal allocation of inventory, leading to loss incurred due to fulfillment centers with overstock or inability to fulfill customer orders where the demanded exceed projections.

In the section to follow, we will provide a simple illustration of the uncertainty associated with an inventory allocation specification. We propose adjusting for this uncertainty by employing a Bayesian approach to statistical modeling which allows retailers to consider all of the *what-ifs* when forecasting sales and allocating the corresponding appropriate inventory quantities.

Solution: skinny in 3 weeks!

While optimizing retail supply chain and inventory allocation is not a new concern for retailers, to our knowledge, most of the existing products providing solutions to these problems aren't leveraging Bayesian methods. Bayesian techniques are a subset of statistical techniques that naturally permit accounting for all of the uncertainty in a forecast that will be used to determine an optimal business action. These methods also naturally allow one to incorporate prior knowledge into predictions and update these beliefs as more data is observed. Consider the following example: suppose a retailer is launching an email promotion for a

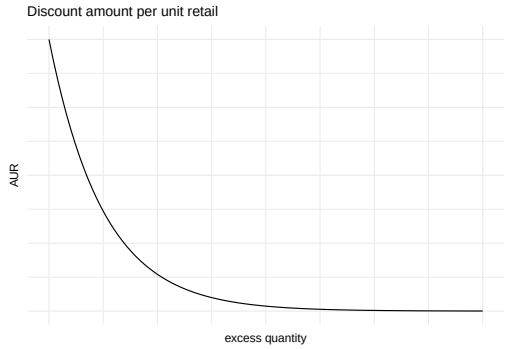


Figure 1: A typical curve for describing the relationship between the price of excess units and the overstock quantity - the amount by which the buy exceeds demand.

particular product, and they believe that the campaign will result in a 40% conversion rate - that is, for every 100 customers who receive the promotional ad, 40 customers will purchase the promoted product through the email link. They launch the campaign, and upon completion,

- 100,000 customers recieved a promotional email, and
- 10,000 customers made a purchase.

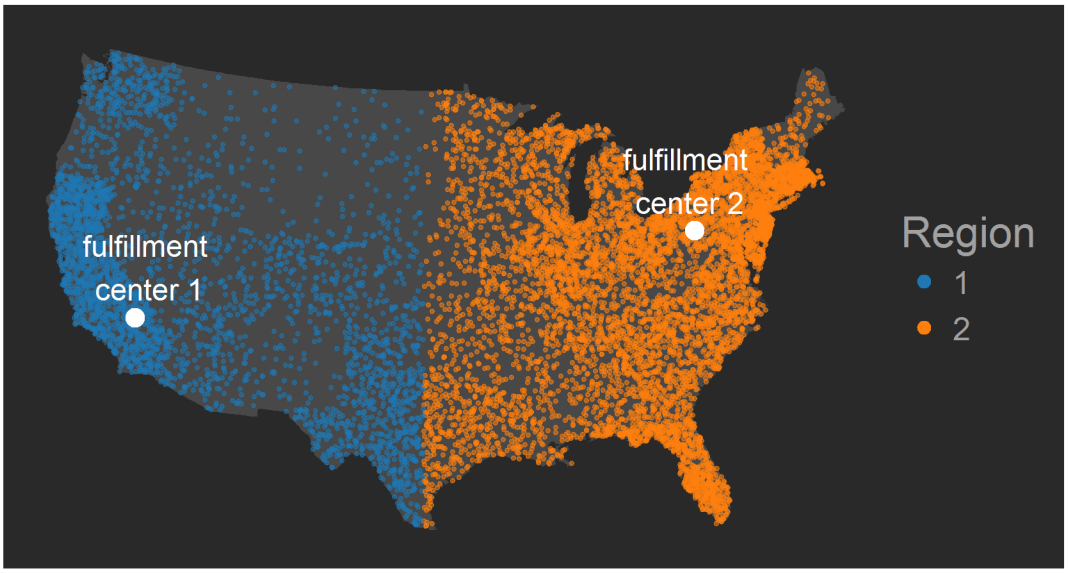
The retailer plans to launch a similar campaign for an analogous product. If you were asked to predict the coversion rate for the new campaign, would you predict the same 40% conversion rate? Perhaps the Bayesian methods us to incorporate prior belief about products and the market as well as update this belief as we observe actual sales.

The total buy must be determined before choosing how to best divide the total inventory among the available fulfillment centers. Denote the *true demand* during the pre-specified time period d^*. Retailers mark down prices to accelerate the rate of sales of inventory that hasn't been sold after a prespecified amount of time has passed. When the demand does not meet the buy, i.e. $d^* < b$, there are $b - d^*$ excess units to be sold. Excess units are sold at a price resulting in a lower AUR than during the planned sales period.

We assume that the AUR for excess units can be expressed using some *markdown* function m. A typical choice for a function describing the relationship is displayed in Figure~. For a product with an AUR of $\$r$, the total revenue is given by

$$\text{revenue} = \begin{cases} br & \text{when demand exceeds the buy.} \\ \underbrace{[rd^*]}_{\substack{full\ price \\ revenue}} + \underbrace{[m\,(b, d^*)]}_{\substack{discounted \\ revenue}} & \text{when the buy exceeds demand.} \end{cases}$$

Revenue is maximized when the buy is equal to the demand, so forecasting the total number of units sold across the entire market is critical.



The total buy is determined using a forecast for the total demand over the entire market, but in practice, the predicted total demand is constructed by adding up the predicted demand over a set of geographic areas which *partition* the market: each household belongs to exactly one of the specified regions, and together, the regions cover the entire market area. For example, we might predict sales within political boundaries (*e.g.,* county or state) or within designated marketing areas. Delivery costs and predicted demand in each region drive the allocation of inventory among fulfillment centers, from which product will be shipped to households. Order fulfillment in a given region is carried out by the nearest fulfillment center in order to minimize costs associated with fulfilling the order. The cost of fulfillment center f fulfilling an order to household h may be expressed as the sum of three individual costs: the fixed cost of fulfilling a single order, the cost of shipping an order one unit distance, and the cost incurred each day required for the customer to receive the order.

$$c(f,h) = \text{fixed cost} + \text{cost of distance traveled} + \text{cost of delivery time.}$$

Consider a simple scenario: a retailer has two fulfillment centers available for distributing orders to locations across the continental US. The map in Figure~ shows the locations of these fulfillment centers and a sample of possible locations to which the retailer will need to ship product.

Suppose that the retailer was able to perfectly predict the total demand over the sales period, and specified the total buy is equal to the total demand. ... ...

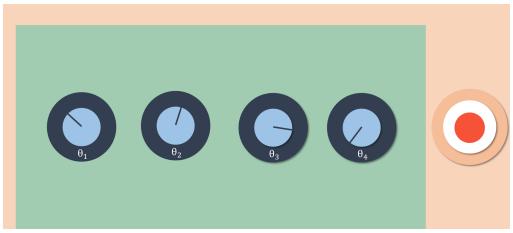
Given an AUR value, the discounting applied to excess inventory, an initial allocation α_{10} and α_{20}, the rules for determining the region to which each fulfillment center is assigned, the costs associated with fulfilling an order, and the demand in each region, we can calculate the monetary value ν of an allocation rule:

$$\begin{aligned} \text{value of allocation} = \nu(\alpha_{10}, \alpha_{20}, y_1^*, y_2^*) &= \text{total revenue} \\ &\quad - \text{cost of fulfilling orders in Region 1 from center 1} \\ &\quad - \text{cost of fulfilling orders in Region 2 from center 2} \end{aligned} \tag{1}$$

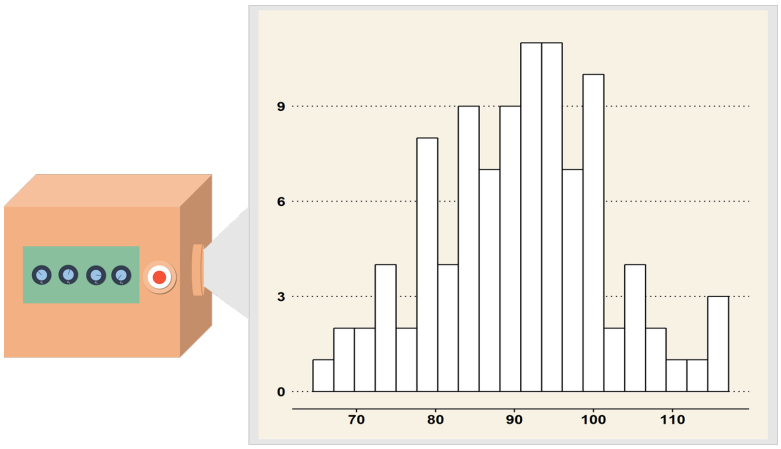
An initial allocation $\left(\alpha_{10}^1, \alpha_{20}^1\right)$ having value ν_1 is better than an alternative initial allocation $\left(\alpha_{10}^2, \alpha_{20}^2\right)$ having value ν_2 if $\nu_1 > \nu_2$.

Why do we want to average over parameter values?

Unfortunately, calculating ν_1 and ν_2 requires knowing the value of future sales in each region, which is unaccessible. Instead, one substitutes predicted future demand for the actual future demand to perform the calculation. Since the value of an allocation will depend on predicted demand, it inherits the uncertainty that we have in predicted demand. We need a way to account for this uncertainty when identifying the allocation with the best value ν. Bayesian methods are a class of statistical techniques that provide a natural way of accounting for uncertainty in demand prediction. To understand how this approach are different from traditional methods and what advantages they provide in this situation, we first need to understand the different sources of uncertainty in predicted demand (and thus, in our estimated value of an allocation.)



Imagine that I have a magic box, and on the box is a magic button and a set of dials. When the button is pressed, the box generates values of total sales in each of the two regions. The dials allow you to specify the "state of the market" in each region during the time that the product will be sold; they control the specific relationship between influential factors, or *independent variables* and total sales. For example, the box might have four dials which control how each of the seasons affect sales in each region. In certain southern regions, we might set these dials so that winter has a strong, positive impact on sales if, for instance, there is an influx of tourism during the winter months that leads to a bump in sales. In northern regions, winter weather may dampen retail activity, so we may set the dial to generate sales given that we expect, say, a 30% decrease during winter months. Imagine that on the box, there are also dials for specifying the effect of all the other relevant independent variables such as population demographics and marketing spend, as well as dials for specifying the relationship between past sales and future sales.



In the context of statistical modeling, the dials control the values of the *model parameters*, $\theta_1, \theta_2, \ldots, \theta_p$. We denote the collection of parameters by $\boldsymbol{\theta}$. The box, however, does not output *deterministic* values; without changing any of the dial settings, if I press the magic button multiple times, the box will produce different output. Repeatedly pressing the button generates a set of values according to the *distribution* of sales; values with higher frequency are more likely to occur given the conditions specified by the dials. This process of generating potential values for sales according to how likely the value is to be observed is called *sampling* from the distribution of sales. Figure~**??** shows what the result of sampling 100 values from the distribution of sales might look like.

The variability in the box's output is referred to as the *intrinsic variance* in sales - the variability in units sold even knowing the model parameters that govern how the box produces the values. While this may seem concerning, we need not worry: standard statistical methods give us the tools for accounting for this variability. If we had the proper crystal ball that could provide the correct settings for the dials reflecting the true nature of the relationship between each of the independent variables and sales, then we can account for the uncertainty in predicted demand in our valuation of an initial allocation in the following way: suppose we have a set of initial allocations $\left(\alpha_{10}^1, \alpha_{20}^1\right), \left(\alpha_{10}^2, \alpha_{20}^2\right), \ldots, \left(\alpha_{10}^A, \alpha_{20}^A\right)$ from which we want to choose the allocation with the highest value ν. In a large sample from the distribution of sales, for each number output, we can calculate an associated value for each initial allocation. For each initial allocation we consider, this effectively converts the distribution of sales to a distribution of allocation value. To assign a single monetary value to each allocation, we simply take the mean of this distribution. An algorithm for this sequence of calculation may be written as follows:

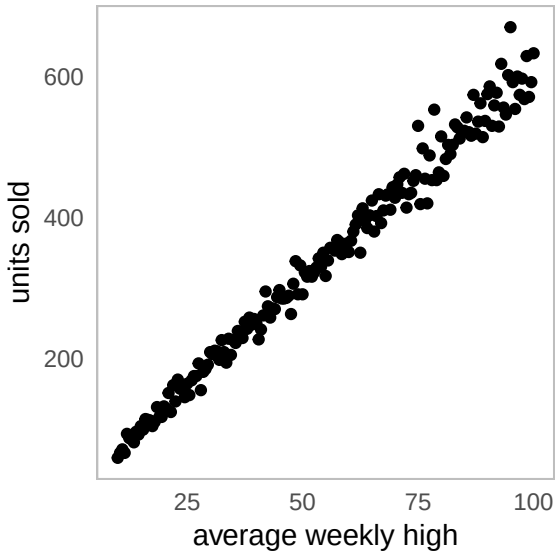
1. n <- 0
2. Press the magic button to obtain predicted demand in Region 1 and Region 2, y_1^* and y_2^*.
3. Given y_1^* and y_2^*, calculate $\nu_a^{(n)} = \nu\left(\alpha_{10}^a, \alpha_{20}^a, y_1^*, y_2^*\right)$ for each allocation in the set $(\alpha_{10}^a, \alpha_{20}^a)$, $a = 1, \ldots, A$.
4. n <- n + 1
5. if n = N, stop

By averaging over the N samples, we assign allocation a value

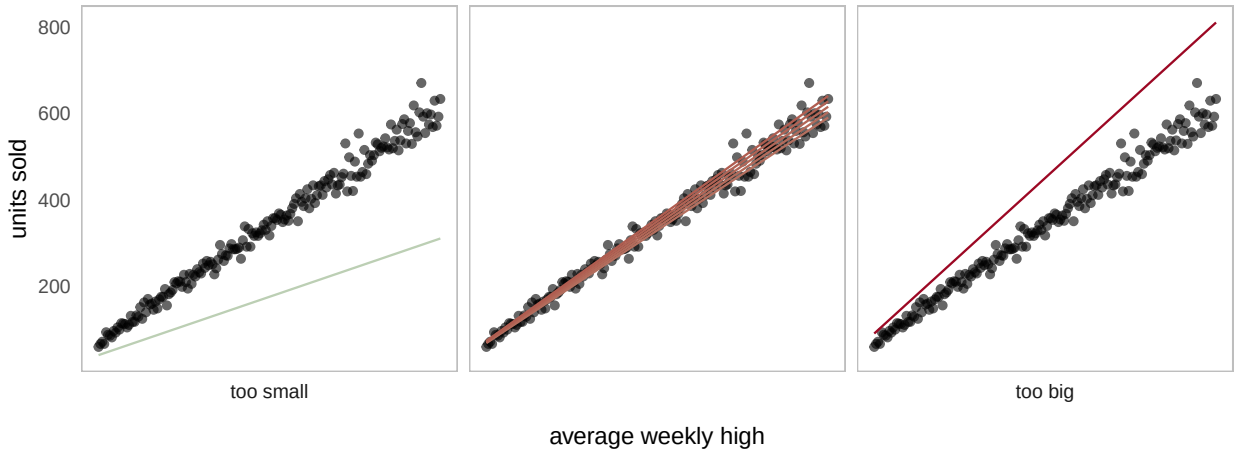
$$\bar{\nu}_a = \frac{1}{N} \sum_{n=1}^{N} \nu_a^{(n)}.$$

Generating a sample of allocation values in this way is known as *Monte Carlo simulation*, and it allows us to account for the variability in sales leftover after accounting for the predictor variables. This is not, however, the only variability that we need to consider. Like the future demand, in practice, model parameters are unknown, so we don't actually know exactly where to set the dials on the black box before outputting

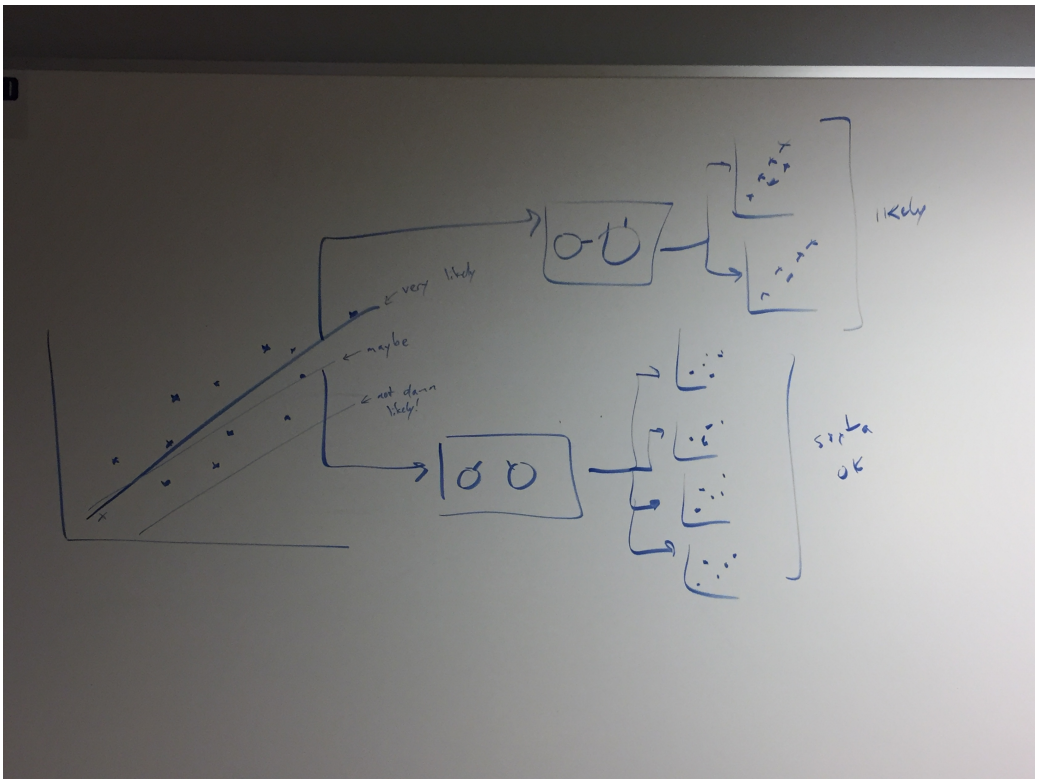
the sample of futures sales. The simplest remedy for this is to estimate the model parameters using past information.



For example, suppose a coffee retailer has records weekly weather and the number of iced coffee units sold per week. Figure ~ displays the relationship between historical weekly sales and the weekly average high temperature. By fitting a statistical model to the historical data, we can infer where to place the dial controlling the parameter for weekly high temperature before using the box to output a sample from the future distribution of coffee sales.



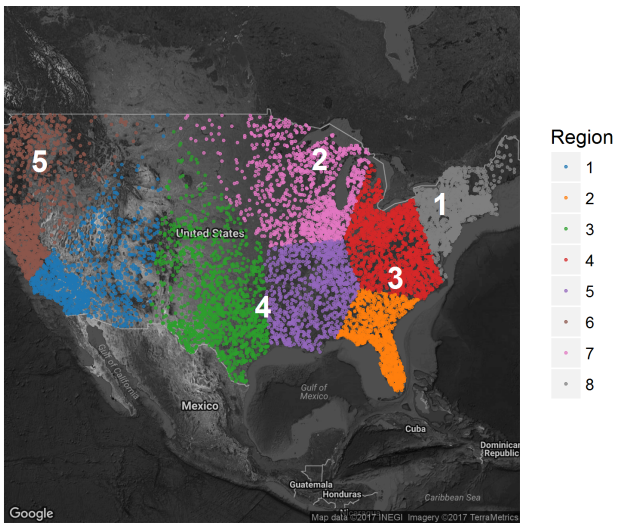
[TODO : Insert muted melon dials set to appropriate parameter value below each facet block, cut off the facet labels]



Figure~ shows how adjusting the parameter values changes the form of the relationship between sales and temperature. Some curves (and hence their corresponding parameter values) appear to be more supported by the data than others. The green curve fits the historical data very well, so it is very likely that those parameter values are close to the parameters that were used to generate the data. If we were to set the dials to these values and output several samples of sales, to which we could compare the historical data and reasonably conclude that each of the curves were produced by the same machine. The blue curve does not fit the data quite as well, but it still appears to approximate the relationship between sales and temperature reasonably well, so it is somewhat likely that those parameter values could be the correct ones. The red curve is the least satisfactory fit to the historical data. For each of the blue and green curves, we could repeat the same investigation as with the green curve. For each group of samples, how likely is it that the historical data and the new samples came from the same machine?

[TODO: add picture here to show how to use the posterior weights to construct my sample from $\pi\left(Y^{*}|Y,X,X^{*}\right)$.]

We could produce many more curves that fit the data well and represent reasonable candidate parameter settings, so how do I choose just one specification of the dials to obtain a sample of sales? Rather than select just one configuration of the dials and obtaining our sample from one setting of the machine, we can account for our uncertainty in the parameter settings by constructing our sample from a mixture of the outputs from each setting. But rather than giving each dial configuration the same weight in the mixture, we give larger weight to machines with parameters that are more likely to have produced the historical data. Output obtained from the machine with dials set to the parameters of the green curve would have high weight, while output from the machine with the blue configuration would have slightly less weight. Though not likely, it's still possible that the output from the machine set to the red parameter values could have produced the historical data, so we allow that output to contribute to the sample, but place lower weight on these samples. Sampling future sales so that we consider each parameter value in proportion to how likely is it that that value is the correct one - given the historical data is the cornerstone of Bayesian forecasting. By taking into consideration every "what-if" case, we can account for prediction uncertainty that standard statistical approaches cannot.



After

- Output -> Outcome -> minimize loss -> increase ROI