A Bayesian Model for Online Activity Sample Sizes

Thomas S. Richardson* University of Washington

Yu Liu* Amazon.com Inc

James McQueen Amazon.com Inc

Doug Hains Amazon.com Inc

Abstract

In many contexts it is useful to predict the number of individuals in some population who will initiate a particular activity during a given period. For example, the number of users who will install a software update, the number of customers who will use a new feature on a website or who will participate in an A/B test. In practical settings, there is heterogeneity amongst individuals with regard to the distribution of time until they will initiate. For these reasons it is inappropriate to assume that the number of new individuals observed on successive days will be identically distributed. Given observations on the number of unique users participating in an initial period, we present a simple but novel Bayesian method for predicting the number of additional individuals who will participate during a subsequent period. We illustrate the performance of the method in predicting sample size in online experimentation.

1 INTRODUCTION

There are many situations in which it is necessary to predict the number of individuals who will engage in an activity for the first time during some specified period. For example, a credit card company may want to predict the number of customers who respond to an offer for a new card; online e-commerce companies may want to predict the number of registered users who opt-in to new features or upgrade their membership. In the domain of online experimentation or A/B testing it is extremely useful to predict the number of individual users who will take part in the experiment as this can be used to estimate how long an experiment needs to run in order to get a sufficient sample size in order to achieve a desired power level. Such experiments are important as they provide a means by which "end-users can help guide the development of features" (Kohavi et al., 2007).

The problem that we consider in this paper is as follows: given a fixed population and observations on the number of unique individuals who initiate activity during an initial period of time, for example, a week, predict the number of additional individuals who will initiate activity during a subsequent period of time, for example one or three weeks.

Though formally the data can be seen as forming a time-series, the problem is not amenable to conventional time-series models. Though time-series model like ARIMA or Dynamic Regression have been used to predict the overall level of online traffic (Anderson, 2016; Björklund and Hasselblad, 2021), such models typically require long series and also assume stationarity of innovations after de-trending. Consequently, such methods are not suitable in many applications in which there is a wish to make predictions shortly after a new service has been launched. The data may also be viewed as a (non-stationary) counting process for which Hawkes processes are sometimes used (Hawkes, 1971a,b). However, since the decision of each individual regarding when (or if) they will participate is taken independently, the process does not typically exhibit self-excitatory behavior of the type described by Hawkes processes.

Another challenge is that, in general, there will be heterogeneity within the population: some customers will initiate much more quickly than others; for example daily users of a mobile app will initiate much sooner than those who use it only for specific or occasional purposes. The users who are observed to initiate activity at the start of the observation period will be weighted more heavily towards frequent users than is the case subsequently. This unknown heterogeneity must be taken into account when making predictions for subsequent periods.

*These authors contributed equally to this work

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

The primary contribution of this paper is the use of a hierarchical Bayesian Beta-Geometric model with censoring to model the heterogeneity of individuals in the population with respect to their propensity to initiate activity on any given day. Estimates for the number of additional individuals arriving in subsequent periods can then be obtained by simulating from the posterior predictive distribution. A convenient by-product of the simplicity of our model is that it is possible to sample from the posterior via straight Monte-Carlo simulation, thereby avoiding issues that may arise with Markov Chain approaches, such as assessing convergence and mixing times, which could present difficulties in large scale or fully-automated applications.

Relating the approach described here to prior work, there is a long history of using hierarchical beta-binomial models in Bayesian statistical inference for count data; see for example, Leonard (1972); Novick et al. (1973); Dempster et al. (1983). However, we are not aware of prior work that has applied these models to inference for the number of individuals initiating behavior in a given time period.

The rate at which customers participate in an online experiments, often called A/B tests, will have a great influence on the effect sizes that can be detected statistically within a given time period. Recent works outlining current best practice for such online experiments include Kohavi et al. (2007) and Bakshy et al. (2014). However, again to the best of our knowledge, previous work has not considered the specific question of predicting how many additional unique customers will be accrued in a subsequent time interval, given the number of unique customers who participated during an initial period. Such predictions are obviously of great relevance to decisions regarding the organization and scheduling of A/B experiments (Kohavi and Longbotham, 2017; Casella and Berger, 2021).

In section 2, we describe our proposed accrual model and the assumptions underlying it. In section 3, we describe how to make inferences about the number of additional individuals arriving in the second period by computing the posterior predictive distribution. In section 4 we illustrate the posterior distribution, predictive distributions and demonstrate the performance of our model on a large meta-analysis of A/B experiments.

2 ACCRUAL MODEL

Our goal here is to build a simple accrual model which uses the number of unique users or customers who first participate at day t, $t = 1, 2, \ldots, d$ to make a prediction regarding the total number of new (unique) customers who will participate in the second period of d^* days. We take $d = 7$ (first week) and d^* a multiple of 7 (multiple weeks) in our data analyses.

2.1 Notations and Model Assumptions

Consider a set of customers indexed by $i \in \{1, \ldots, n\}$ and "Day t" means the t-th day on which it was possible to participate.

Our model makes the simplifying assumption that each customer i participates for the first time on any given day with a probability π_i. We will also assume that if on day $t = 1, 2, \ldots$ a customer has never participated previously, then their probability of participating on the next day remains π_i. Consequently, the day of the experiment on which customer i first participates will follow a Geometric distribution with probability π_i.

2.2 Simple Accrual model: π discrete

In the model that we propose below the underlying heterogeneity among customers in their behavior is described by modeling π_i as being drawn from a distribution with support on $[0, 1]$. However, to provide intuition, before describing this model, consider first the simpler setting in which π_i is discrete, taking 100 different uniformly spaced values:

$$\pi_i \in \{0.01, \ldots, 0.98, 0.99, 1.00\}.$$

π_i can be viewed as a characteristic of a customer. Thus it make sense to group together customers with the same value of π. Let N_π denote the number of customers in this group. Thus, in the simple setting described, there will be 100 groups of customers. It follows that the expected number of customers of type π who first participate on Day 1 will be $\pi \times N_\pi$, the expected number who first participate on Day 2 will be $(1-\pi)\pi N_\pi$, and the expected number participating on Day k will be $(1-\pi)^{k-1}\pi N_\pi$.

Consequently, in expectation, the set of customers who participate on Day 1 will consist of the following, ordered according to the value of π:

$$(0.01 \cdot N_{0.01},\ \ 0.02 \cdot N_{0.02}, \ldots, 0.99 \cdot N_{0.99},\ \ 1 \cdot N_1).$$

Similarly, the set of customers who first participate on Day 2 will consist, in expectation, of the following:

$$(0.99 \cdot 0.01 \cdot N_{0.01},\ \ 0.98 \cdot 0.02 \cdot N_{0.02}, \ldots,\\ 0.01 \cdot 0.99 \cdot N_{0.99},\ \ 0 \cdot 1 \cdot N_1).$$

Notice that relative to Day 1, a smaller proportion of people with $\pi = 0.99$ participate *for the first time* on Day 2, since 99% of such people already participated on Day 1. By repeating this procedure, we may simulate the number and type of customers who first participate on Day t.

2.3 Accrual model: Geometric likelihood with censoring

We now consider a more realistic setting in which π is continuous and build a model for the distribution of the participation probabilities π. Similar to section 2.2, given the model assumptions, the day on which individual i first participates will follow a Geometric distribution with probability π_i:

$$\text{day on which } i \text{ initiates} \,|\, \pi_i \quad \sim \quad Geo(\pi_i). \qquad (1)$$

For each individual $i \in \{1, 2, \ldots, n\}$ who participates in the first period of length d we will define X_i to be the day on which they first participated; if an individual i never participates in the first period, then we define $X_i = 0$. Thus $X_i \in \{0, 1, \ldots, d\}$, where, for example, $d = 7$ if the first period, for which participation data is available, is a week. We formulate the model hierarchically, with a Beta distribution over the unknown participation probabilities. The model is specified formally as follows:

$$\begin{aligned} p(X_i = x_i \mid \pi_i) \quad &= \quad \begin{cases} (1-\pi_i)^{x_i - 1}\pi_i, & x_i \in \{1, \ldots, d\}; \\ (1-\pi_i)^d, & x_i = 0; \end{cases} \\ \pi_i \mid \alpha, \beta \quad &\sim \quad \text{Beta}(\alpha, \beta); \\ p(\alpha, \beta) \quad &\propto \quad (\alpha + \beta)^{-5/2}, \quad \alpha > 0, \beta > 0. \end{aligned} \qquad (2)$$

Note that the first term here can also be rewritten as:

$$p(X_i = x_i \mid \pi_i) = \pi_i^{\mathbb{I}(x_i > 0)} (1-\pi_i)^{(x_i - 1)\mathbb{I}(x_i > 0) + d\,\mathbb{I}(x_i = 0)};$$

Here $\propto$ indicates *proportional to*, while $\mathbb{I}(\cdot)$ is the indicator function that takes the value 1 if the condition is true and 0 otherwise. $p(X_i = x_i \mid \pi_i)$ follows a censored geometric distribution with success probability π_i, where values larger than d are censored and recorded as 0. We use the conjugate prior for π_i, which is Beta(α, β). The hyper prior $p(\alpha, \beta)$ is a default prior recommended by Gelman et al. (2004)[§5.3].

A graphical depiction of the proposed model is in Figure 1. π_i is the individual-specific probability of customer i participating on a given day. As described above, π_i can be viewed as describing a given customer's propensity to initiate activity. Similarly, the hyper-parameters α, β which determine the distribution from which the π_i values are drawn, can be viewed as describing the character of different activities. For example, customers may be more likely to participate in an experiment changing the layout of the home page of a website than they might be to participate in an experiment changing the customer support contact page.

From the two assumptions given in Section 2.1, it follows that the parameters $(\pi_1, \ldots, \pi_n)$ are i.i.d. conditional on the hyper parameters α, β. In other words, under the model each individual's probability π_i is an i.i.d. draw from a Beta(α, β) distribution.

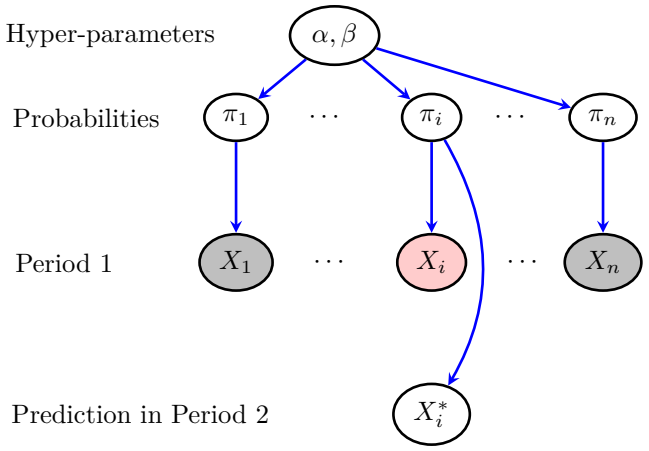


Figure 1: Graphical Model: The hierarchical Beta-Geometric model. (α, β) are hyper-parameters; for an individual i, π_i is the individual-specific probability of first participating on a given day, having not participated previously; X_i indicates *either* the day on which the customer i first participates in the first period *or* whether the individual did not participate in the first period ($X_i = 0$). Shaded nodes are observed; nodes shaded red indicate individuals who didn't participate in the first period. X_i^* is a prediction of when an individual who did not participate in the first period will first participate in the second period, or if they will again not participate ($X_i^* = 0$).

To understand and motivate the hyper prior, we first recall the following facts about Beta distributions:

If $\pi^* \sim \text{Beta}(\alpha, \beta)$ then:

(a) $E[\pi^*] = \frac{\alpha}{\alpha+\beta} \in [0, 1]$.

(b) The variance is given by:

$$\begin{aligned} V[\pi^*] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{E[\pi^*](1 - E[\pi^*])}{\alpha+\beta+1} \leq \frac{1}{4(\alpha+\beta+1)} \end{aligned} \qquad (3)$$

where we have used the fact that $E[\pi^*](1 - E[\pi^*]) \leq 1/4$. Consequently, for large α, β, $\text{sd}(\pi^*) \approx (\alpha + \beta)^{-0.5}/2$.

Consequently, the hyper-prior (2) corresponds to putting independent (improper) uniform priors on $E[\pi^*] = \alpha/(\alpha + \beta)$ and $(\alpha + \beta)^{-0.5} \approx \text{sd}(\pi^*)$. Note that the hyperprior (2) used here is *improper*; it does not integrate to 1. However, it will lead to a proper posterior provided that for at least one i, $0 < x_i < d$; see Gelman et al. (2004)[Ex. 5.7].

3 INFERENCE FOR THE ACCRUAL MODEL

In this section, we discuss how to make predictions using the proposed accrual model.

3.1 Posterior predictive probability of continued non-participation

We first compute the elements in the formula of the posterior predictive distribution for future point X_i^* in Figure 1. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the vector recording the day of the first participation or 0 if there was no visit during the first period for n individuals, and let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ be the individual-specific probabilities.

From Bayes rule, following the independence assumptions, we can obtain the *joint posterior over parameters and hyper-parameters* (α, β):

$$\begin{aligned} &p(\boldsymbol{\pi}, \alpha, \beta \,|\, \boldsymbol{x}) \\ &\quad \propto p(\alpha,\beta) p(\boldsymbol{\pi} \,|\, \alpha, \beta) p(\boldsymbol{x} \,|\, \boldsymbol{\pi}, \alpha, \beta) \\ &\quad \propto p(\alpha,\beta) \prod_{j=1}^{n} p(\pi_j \,|\, \alpha, \beta) \prod_{k=1}^{n} p(x_k \,|\, \pi_k) \\ &\quad \propto p(\alpha,\beta) \prod_{j=1}^{n} \Bigg(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_j^{\alpha-1+\mathbb{I}(x_j>0)} \\ &\qquad\qquad \times (1-\pi_j)^{\beta-1+(x_j-1)\mathbb{I}(x_j>0)+d\,\mathbb{I}(x_j=0)} \Bigg). \end{aligned} \tag{4}$$

Similarly we obtain the *conditional posterior distribution over* π_i *given hyper-parameters* (α, β) *and the data* x_i:

$$\begin{aligned} &p(\pi_i \,|\, \alpha, \beta, x_i) \\ &\quad \propto p(\pi_i \,|\, \alpha, \beta) p(x_i \,|\, \pi_i) \\ &\quad = \mathrm{Beta}(\alpha + \mathbb{I}(x_i>0), \beta + (x_i-1)\mathbb{I}(x_i>0) + d\,\mathbb{I}(x_i=0)) \\ &\quad = \frac{\Gamma(\alpha+\beta+x_i+d\,\mathbb{I}(x_i=0))}{\Gamma(\alpha+\mathbb{I}(x_i>0))\Gamma(\beta+(x_i-1)\,\mathbb{I}(x_i>0)+d\,\mathbb{I}(x_i=0))} \\ &\qquad \times \pi_i^{\alpha-1+\mathbb{I}(x_i>0)} (1-\pi_i)^{\beta-1+(x_i-1)\,\mathbb{I}(x_i>0)+d\,\mathbb{I}(x_i=0)}, \end{aligned} \tag{5}$$

where we have used the fact that $\mathbb{I}(x_i > 0) + (x_i - 1)\,\mathbb{I}(x_i > 0) = x_i \mathbb{I}(x_i > 0) = x_i$. Here (5) follows from the form of the joint posterior (4), which shows $\pi_i \perp\!\!\!\perp \boldsymbol{\pi}_{-i}, \boldsymbol{x}_{-i} \mid x_i, \alpha, \beta$, where we use $\boldsymbol{\pi}_{-i} \equiv (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \pi_n)$ to indicate the 'other' elements of $\boldsymbol{\pi}$ and likewise for $\boldsymbol{x}_{-i}$; the conditional independence here may also be obtained by applying the graphical d-separation criterion to the DAG shown in Figure 1.

Thus from (5) we obtain:

$$\begin{aligned} &p(\boldsymbol{\pi} \mid \alpha, \beta, \boldsymbol{x}) \\ &= \prod_{j=1}^{n} p(\pi_i \mid \alpha, \beta, x_i) \\ &= \prod_{j=1}^{n} \frac{\Gamma(\alpha+\beta+x_j+d\,\mathbb{I}(x_j=0))}{\Gamma(\alpha+\mathbb{I}(x_j>0))\Gamma(\beta+(x_j-1)\,\mathbb{I}(x_j>0)+d\,\mathbb{I}(x_j=0))} \\ &\qquad \times \pi_j^{\alpha-1+\mathbb{I}(x_j>0)} (1-\pi_j)^{\beta-1+(x_j-1)\,\mathbb{I}(x_j>0)+d\,\mathbb{I}(x_j=0)}. \end{aligned} \tag{6}$$

Similar to the development in Gelman et al. (2004)[§5.3, Eq.(5.5) Eq.(5.8)], the *posterior over the hyper-parameters* α, β is then:

$$\begin{aligned} &p(\alpha, \beta \,|\, \boldsymbol{x}) \\ &= \frac{p(\alpha, \beta, \boldsymbol{\pi} \,|\, \boldsymbol{x})}{p(\boldsymbol{\pi} \,|\, \alpha, \beta, \boldsymbol{x})} \\ &\propto p(\alpha,\beta) \prod_{j=1}^{n} \Bigg(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &\times \frac{\Gamma(\alpha+\mathbb{I}(x_j>0))\Gamma(\beta+(x_j-1)\,\mathbb{I}(x_j>0)+d\,\mathbb{I}(x_j=0))}{\Gamma(\alpha+\beta+x_j+d\,\mathbb{I}(x_j=0)} \Bigg). \end{aligned} \tag{7}$$

For a given individual i, who did not participate in the first period, so $X_i = 0$, we will define X_i^* to be the first day on which the individual participates in the second period of length d^*, or, as before, we define $X_i^* = 0$ if they also never participate in the second period. Given the hyper-parameters (α, β), it follows from (5) that:

$$\begin{aligned} &p(X_i^* = x_i^* \mid \alpha, \beta, X_i = 0) \\ &= \int p(X_i^* = x_i^* \mid \pi_i, \alpha, \beta, X_i = 0) p(\pi_i \mid \alpha, \beta, X_i = 0) d\pi_i \\ &= \int p(X_i^* = x_i^* \mid \pi_i) p(\pi_i \mid \alpha, \beta, X_i = 0) d\pi_i \\ &= \frac{\Gamma(\alpha+\beta+d)}{\Gamma(\alpha)\Gamma(\beta+d)} \\ &\times \frac{\Gamma(\alpha+\mathbb{I}(x_i^*>0))\Gamma(\beta+d+(x_i^*-1)\mathbb{I}(x_i^*>0)+d^*\mathbb{I}(x_i^*=0))}{\Gamma(\alpha+\beta+d+x^*+d^*\mathbb{I}(x_i^*=0))}. \end{aligned} \tag{8}$$

Thus, conditional on hyper-parameters (α, β), the *probability that an individual who has not participated in the first period will again not participate in the second period* is:

$$\begin{aligned} &p(X_i^* = 0 \mid \alpha, \beta, X_i = 0) \\ &\qquad = \frac{\Gamma(\alpha)\Gamma(\beta+d+d^*)}{\Gamma(\alpha+\beta+d+d^*)} \frac{\Gamma(\alpha+\beta+d)}{\Gamma(\alpha)\Gamma(\beta+d)} \\ &\qquad \equiv q_0(\alpha, \beta, \boldsymbol{x}). \end{aligned} \tag{9}$$

We express the probability $q_0(\alpha, \beta, \boldsymbol{x})$ of continued non-participation as a function of the full first period data $\boldsymbol{x}$ because this determines the set of individuals i who did not participate in the first period, for whom $X_i = 0$.

Note that under the model this probability will be the same for all people who did not participate in the first week.

3.2 Predicting the number of new individuals in the second period

Let n_0 be the number of individuals who did not participate in the first period. Similarly, let $\{i_1, \ldots, i_{n_0}\}$ be the subset of individuals who did not participate in the first period.

Given α, β the corresponding variables $\{X^*_{i_1}, \ldots X^*_{i_{n_0}}\}$ are i.i.d., and have probability $q_0(\alpha, \beta, \boldsymbol{x})$ of taking the value 0. It follows that given α, β the *number* of individuals n_{00} who did not participate in the first period and who again did not participate in the second period will follow a binomial distribution, with the following parameters:[1]

$$n_{00} \mid \alpha, \beta, \boldsymbol{x} \sim \text{Binomial}(n_0, q_0(\alpha, \beta, \boldsymbol{x})) \tag{10}$$

Hence a simple Monte-Carlo scheme for simulating from the posterior for n_{00} is as follows:

1. Simulate α, β from the posterior distribution (7);
2. For each value of (α, β) draw a number from the Binomial distribution given by (10).

In step one we use the (straight) Monte-Carlo Ratio-of-Uniforms algorithm (Wakefield et al., 1991; Kinderman and Monahan, 1977) to sample α, β. A point estimator of n_{00} may be obtained by taking the median of the different draws from step 2. We use the posterior median as it is more robust than the posterior mean.

Thus the number of individuals who did not participate in the first period, but did participate in the second period is:

$$\textit{no. of individuals participating for first time in Period 2} = n_0 - n_{00}. \tag{11}$$

Hence we are able to efficiently obtain samples from the posterior for the number of additional individuals participating during the second period.

3.3 Inference when total population size is unknown

In the development so far, we have supposed that the number of censored observations, n_0 corresponding to individuals who did not show up in the first period, is known. This will hold in settings where the population of users is known, for example, where subscriptions or accounts are required for participation.. However, in some situations this may not be the case.

Fortunately, this turns out not to be a practical issue: provided that the number of censored observations is large relative to the number of uncensored observations, the posterior distribution for the number of additional individuals who will participate for the first time in the second period is not sensitive to the exact number of censored observations; see Section 4. In more detail, one may use a plug-in estimate of n_0 given by a multiple λ of the uncensored observations:

$$\hat{n}_0 = \lambda \sum_{i=1}^{n} \mathbb{I}(x^*_i > 0). \tag{12}$$

If necessary, λ can be tuned via cross-validation (James et al., 2013) on past data.

4 Experimental Results

In order to evaluate the performance of the model, we obtained participation data from experiments performed at Amazon.com, Inc. during the last year in the United States. We collected data from 1961 experiments that ran for at least 2 weeks, and 976 that ran for at least 4 weeks. For each experiment we used the number of new customers first participating in the experiment each day during the first week to predict the total sample size if the experiment were to continue running for k weeks in total, where $k = 2, 3, 4, 5$. (Axes in these plots are not annotated owing to business confidentiality.)

4.1 Illustrative Example

We first demonstrate the method using data from a single experiment that ran for at least 2 weeks. Figure 2a shows the participation data from the first week of the experiment. Here the X-axis represents the day index $t = 1, 2, \ldots, 7$, while the Y-axis is the number of customers who first participate in the experiment on day t.

Following the development in Section 3 for this experiment we first obtained samples from the joint posterior distribution over the hyper-parameters α and β, and then drew samples from the joint posterior predictive distribution for the number of new customers who first participate in each successive week. We define S_t to be the number of new customers who first participate in the experiment on day t. Thus the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ in Section 3.1 will consist of S_1 ones, S_2 twos $\ldots$, S_7 sevens. Since data on the total population size was not available, the number of individuals

[1] Note that the posterior distribution for n_{00} is *not* a Beta-Binomial distribution. n_{00} is Binomial given (α, β), but the posterior over (α, β) is *not* a Beta distribution.

who did not participate in the first week was estimated via (12) with $\lambda = 10$; note that this may be expressed as $\hat{n}_0 = \lambda \sum_{i=1}^{7} S_i$ and is an estimate of the number of zeros in the vector $\boldsymbol{x}$. Here $\lambda = 10$ was selected by tuning from a set of four past experiments. Since n can be large in some experiments, the use of the sufficient statistics $(S_1, \ldots, S_7)$ can greatly speed up the computation relative to simply using $\boldsymbol{x}$.

For the experimental data shown in Figure 2a, Figure 2b shows a scatterplot of 1000 samples of the hyper-parameters (α, β) drawn from posterior distribution (7) together with an estimated density plot. We used an implementation of the (straight) Monte-Carlo Ratio-of-Uniforms algorithm (Wakefield et al., 1991; Kinderman and Monahan, 1977; Northrop, 2021, 2020) to obtain these samples. An implementation using sufficient statistics, based on code from the Bang package (Northrop, 2020), can be found in github (Richardson and Liu, 2022). The actual scales of the X and Y axes in Figure 2b are small indicating that, under the model, the posterior distribution for α and β is concentrated.

For each pair of (α, β) values, we then use Equation (9) with $d^* = \{7, 14, 21, 28\}$ to obtain samples from the joint posterior predictive distribution for the probability that an individual who did not participate in the experiment in the first week will again not participate in the subsequent 1, 2, 3 and 4 weeks if the experiment were to continue running. Plugging these probabilities and $\hat{n}_0$ into Equation (10) results in predictions for the number of customers who did not participate in the initial period and who also do not participate in the second period of d^* days, denoted $\hat{n}_{00}(d^*)$ where $d^* = \{7, 14, 21, 28\}$. Thus the predicted number of customers first participating in the experiment in the k-th week is $\hat{n}_{00}((k-2)*7) - \hat{n}_{00}((k-1)*7)$ where $k \in \{2, 3, 4, 5\}$ and $\hat{n}_{00}(0) = \hat{n}_0$. We repeat this procedure for all pairs of α and β values. This results in samples from the posterior distribution over the function giving the number of new customers in each subsequent successive week. The 1000 predicted curves corresponding to the posterior draws are shown in Figure 2c. The predicted curves show little variability, which is perhaps not surprising, given that the posterior over α and β, shown in Figure 2b, is highly concentrated,

In Figure 3a we show how the predicted number of customers first participating in the experiment at week 2 is relatively insensitive to changes in λ. Specifically, Figure 3a shows boxplots giving 1000 posterior draws for each value of $\lambda \in \{1, 2, 3 \ldots, 30\}$. As can be seen, when λ ranges from 8 to 30, the median of the boxplot for the posterior predictive distribution is approximately constant. The red dotted line in this plot is the ground truth. Though the posterior median underestimates the ground truth, the accuracy is adequate for practical purposes.

4.2 Comparison of Bayesian Model to Other Approaches

We compared our approach to the following three baseline models: 1. Time-Series Models; 2. Log-Linear Models; 3. Random Forest Models. Results from two further baselines are given in the Supplementary Materials.

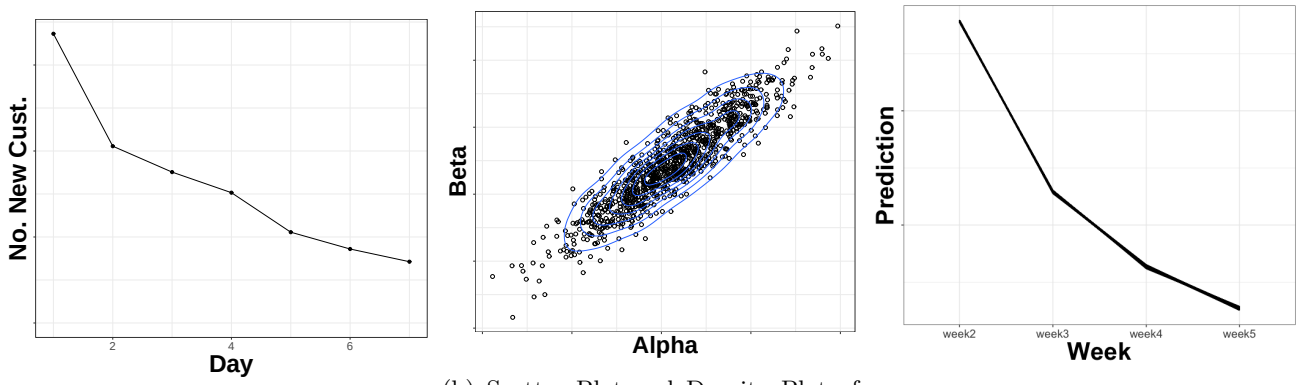
Time-series (ARIMA) models are widely used to predict online traffic trends (Anderson, 2016; Björklund and Hasselblad, 2021). However, there are several obstacles to applying such models in our context: (1) online A/B tests often take place during a relatively short experiment window and thus there are an insufficient number of observed time points; (2) The participation data are non-stationary requiring the specification of a model for the trend, but the trend here is of primary interest! (3) Though in some other settings it is possible to model the trend as a function of relevant baseline covariates, these may not be available. We fitted ARIMA models to the initial 7 days of data using `auto.arima` in the `forecast` R package (Hyndman et al., 2021). Random walk ARIMA(0, 1, 0) was fitted but generated very poor predictions. We also tried to de-trend using linear regression on the number of days t, but this gives poor predictions due to the extrapolation. Since the time-series approaches were qualitatively much less competitive than the others, we do not consider these models further here.

Instead, we fitted a simple log-linear model:

$$\log(T_d + 1) = \beta_0 + \beta_1 d + \epsilon_d, \tag{13}$$

where T_d is the number of customers who first participate in the experiment on day d and ϵ_d are independent errors with $E[\epsilon_d \,|\, d] = 0$. Estimates for β_0 and β_1 may be obtained by linear regression of the log of the number of new participants on day d against d, for $d = 1, \ldots, 7$. The predicted number of new customers at week k can be extrapolated by plugging in $d = 7 * (k-1) + 1, \ldots, 7 * k$ and then summing the predictions.

A Random Forest Model (Breiman, 2001) is a flexible and general machine learning model. It uses bagging and ensemble techniques to obtain estimates from a set of regression trees. Here we assume that there is a training set of previous participation data, both from the initial period *and* the subsequent periods for a large set of experiments. In more detail, the i-th row and j-th column of the feature matrix $\mathbf{S}$ is $S_{i,j}$: the number of customers first participating in the experiment i at

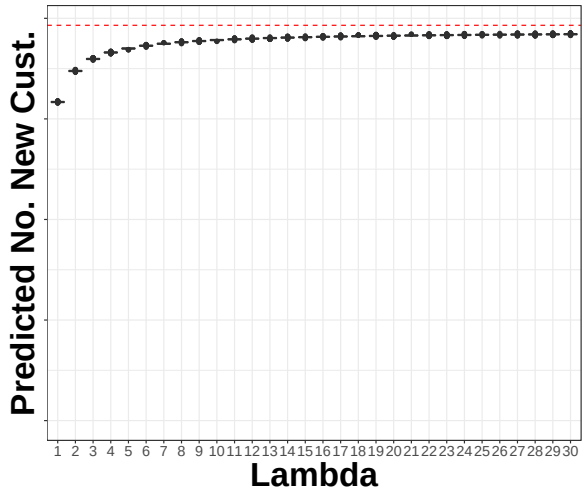


(a) Observed Data for an Experiment

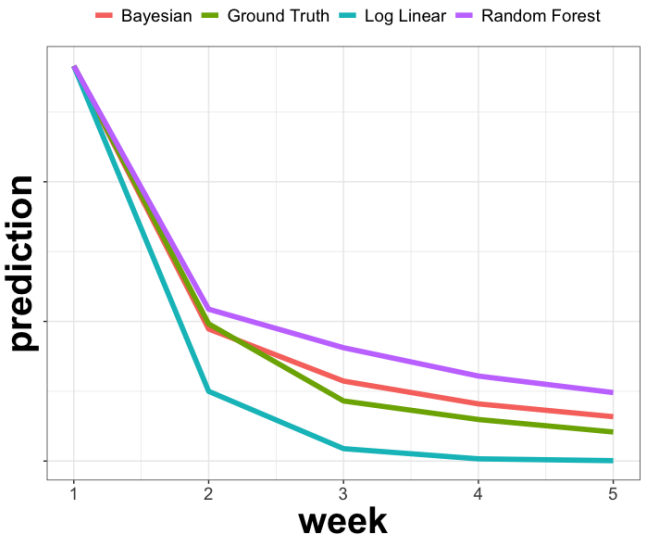
(b) Scatter Plot and Density Plot of Posterior (α, β)

(c) Predicted Curves from 1000 Simulations.

Figure 2: Illustrative Example: For the sake of simplicity, we label customers participating in the experiment first time as "new customers". (a) the X-axis represents the number of days since the start of the experiment; the Y-axis represents the number of new customers who first participate in the experiment during each day in the initial period. (b) Scatter plot of 1000 pairs of (α, β) values sampled from Equation (7) given the data in Figure 2a with $\lambda = 10$. Contour lines show the density of (α, β) and nested contours indicate regions of higher local density. (c) The Y-axis represents the predicted number of new customers participating in the experiment at week k, where $k \in \{1, 2, 3, 4\}$. The plot displays 1000 posterior draws for the prediction functions corresponding to the (α, β) samples in Figure 2a.

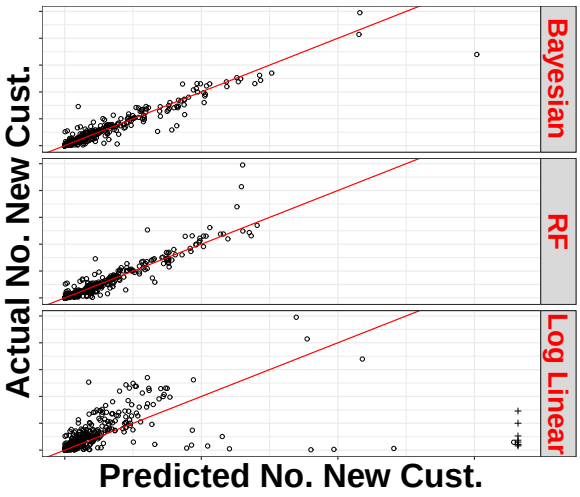


(a) Box Plots of the Predicted Number of New Customers Participating in the Experiment at Week 2 Against Different λ Values

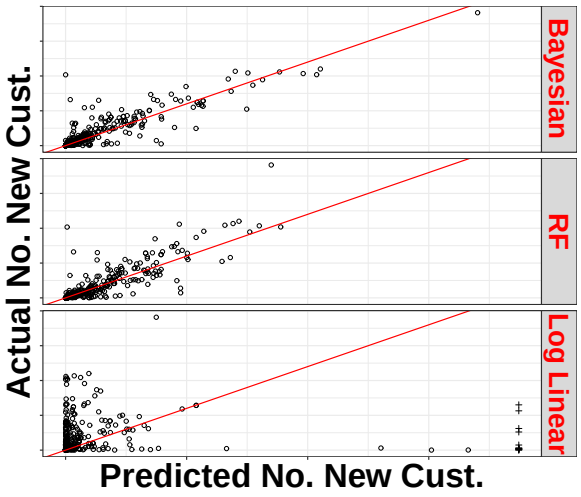


(b) Comparison of Predictions Using Different Methods

Figure 3: Results For Illustrative Example: (a) Box plots of the predicted number of new customers participating in the experiment in Figure 2a at week 2 using different lambda values, $\lambda \in \{1, 2, \ldots, 30\}$. The X-axis is the value of lambda. Left bottom corner of the plot is $(0, 0)$. Each boxplot displays the distribution of 1000 draws from the posterior predictive distribution for the number of new customers at week 2 obtained when using a specific value of lambda. The Red dotted line is the ground truth observed at week 2. (b) Comparison of the predicted number of new customers participating in the experiment shown in Figure 2a at the k-th week using different methods. The X-axis represents the week index and the Y-axis represents the number of customers first participating in the experiment at week k. The true number for week 1 is observed; the numbers for weeks 2, 3, 4, 5 are predicted using different methods. The red curve shows the predictions using the proposed Bayesian model. Predictions from the Log Linear and Random Forest models are given by the blue and purple curves respectively. The ground truth is given by the green curve.



(a) Comparison at Week 2



(b) Comparison at Week 4

Figure 4: Meta-Analysis Results: The vertical axes show the true number of new customers participating in an experiment at the k-th week; the horizontal axes give the predicted count of new customers participating based on the first week of data. The 45° line, corresponding to perfect prediction, is shown in red. Scatter plots contain predictions at week k for experiments running for at least k weeks. The left panel shows results for 981 experiments that ran for 2 weeks; the right panel shows 488 experiments that ran for 4 weeks. The plots show the performance of the proposed predictor (top), random forest regression (middle) and the Log Linear predictor (bottom) at (a) 2 weeks and (b) 4 weeks. For visualization purposes, outliers in the Log Linear predictions are Winsorized at the maximum value on the horizontal axis and indicated by +.

day j where $j \in \{1, \ldots, 7\}$. Let the response variable be $F_{i,k}$, the number of customers first participating in the experiment i at week k, where $k \in \{2, 3, 4, 5\}$. We fit 4 different random forest models on the training set: $F_k \sim RF(\mathbf{S})$, where $k = 2, 3, 4, 5$ and make prediction for each model on the test set. We use the random forest package in R (Liaw and Wiener, 2018) with 500 trees and default number of variables $(2 - 3)$ randomly sampled as candidates at each split as parameters.

4.3 Predictive Performance

Figure 3b compares the predicted curves using the proposed Bayesian model, log linear regression, random forest regression and the ground truth for the example experiment in Figure 2a. We use the median among 1000 simulations as the predictor for the proposed Bayesian model. For the random forest, the test data is the feature matrix of the target experiment while the training data are the feature matrix and responses for experiments (approximately 300) that took place before the target experiment. The selection of training data here mimics the way that such a model would be implemented in practice. As illustrated in Figure 3b, the Bayesian predictor has the best performance in this illustrative example.

To assess performance on the full meta-analysis, we sort the 1961 experiments that ran for at least 2 weeks by their start date and, in order to accommodate the Random Forest, we split the first half of this set as the training dataset and the second half as the testing dataset. As before, we used $\lambda = 10$ in the Bayesian model. We repeat the same procedure for 976 experiments running for at least 4 weeks and predict the number of customers first participating at week 4. Figure 4 shows the true number of new customers participating in the experiment at the kth week vs the predicted values using different methods. The $y = x$ line is shown in red in order to aid visual inspection of the performance of the various models. The plots suggest that (1) the Bayesian method has similar performance to the random forest model while the log linear predictor is unstable; (2) The Bayesian model tends to underestimate the ground truth.

We use the Root Mean Square Error (RMSE) and Mean absolute percentage error (MAPE) to quantify the predictive performance.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Predicted_i - Actual_i)^2}{n}}.$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{Predicted_i - Actual_i}{Actual_i} \right|.$$

Table 1 compares the RMSE and MAPE for the random forest regression, log linear regression and the proposed

Bayesian estimator. These results confirm that the Bayesian model slightly outperforms the log-linear and random forest model in terms of MAPE and has similar performance to the random forest model in terms of RMSE. Both models outperform the Log Linear model. Note that whereas the Random Forest model has access to training data, this is not used by the log-linear or Bayesian approaches (other than the four experiments that were used for selecting λ).

Table 1: Meta-Analysis Predictive Performance: RMSE and MAPE comparison between the Log Linear, Random Forest regression and Bayesian predictors.

	Week 2		Week 4	
	RMSE	MAPE	RMSE	MAPE
Log-Lin	1.84e+11	5.72e+5%	9.06e+17	1.12e+13%
RF	7.05e+05	44.06%	5.49e+05	165.22%
Bayes	7.11e+05	32.59%	5.01e+05	84.57%

In Appendix A we include additional experiments comparing the performance of the proposed method with a Neural Net model and a Censored Weibull model. These results confirm that the proposed method is competitive with the other methods.

Note that when n_0 is known the proposed Bayesian method does not require any past experiments as training data and thus may be applied in contexts where past experiments are either not available or have a different distribution compared to the target experiment; even when n_0 is unknown, only a small number of experiments are typically required to tune λ. In contrast, traditional machine learning methods, require training data from the same distribution. The latter assumption may not hold if experiments are conducted by different teams or at different times. Indeed, this may explain some of the results given in the Supplement, where the proposed Bayesian method outperforms the Random Forest in Table 2. Compared with the Log Linear and Censored Weibull methods, the proposed method is derived from a simple, interpretable model of customer behavior.

5 Conclusion

We have proposed a Bayesian approach to predict online activity in a fixed population. We demonstrated the utility of this method by applying it to predict sample size in online A/B testing. We evaluated the performance of our predictions by comparing them to the ground truth and other baselines in a large collection of online experiments. Our results show the practical utility of the proposed method.

The proposed Bayesian method does not require past experiments to use as training data when n_0 is known. It also does not require stationarity assumptions, or long observation periods. The proposed method also takes into consideration the heterogeneity within the population and thus captures differences in customer behavior when using online services. Our method is simple but effective, and is being used in production to predict sample sizes for online experiment in Amazon.

There are several further refinements that we intend to explore. As indicated by Figure 3a, we found that in practice the proposed model typically slightly underestimates the ground truth. One possible explanation is that there may be growth in the pool of potential users that is occurring while the experiments is running. This suggests that it may be possible to improve predictions by incorporating additional (dynamic) information regarding population size.

References

Linda Anderson. Library website visits and enrollment trends. *Evidence Based Library and Information Practice*, 11(1):4, 2016.

Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on the World Wide Web*, pages 283–292, 2014.

Martin Björklund and Felix Hasselblad. The effect of online advertising in a digital world: Predicting website visits with dynamic regression, 2021.

Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

George Casella and Roger L. Berger. *Statistical inference*. Cengage Learning, 2021.

Arthur P. Dempster, Murray R. Selwyn, and Barbara J. Weeks. Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association*, 78(382):221–227, 1983.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2004.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 04 1971a.

Alan G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971b.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2021. URL `https://cran.r-project.org/web/packages/forecast/forecast.pdf`. R package version 8.15.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

Richard Kay. Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3):227–237, 1977.

Albert J. Kinderman and John F. Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):257–260, 1977.

Ron Kohavi and Roger Longbotham. Online controlled experiments and A/B testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Ron Kohavi, Randal M. Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 959–967, 2007.

Thomas Leonard. Bayesian methods for binomial data. *Biometrika*, 59:581–589, 1972.

Andy Liaw and Matthew Wiener. *R package random Forest*, 2018. URL `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`. R package version 4.6-14.

C. H. Bryan Liu, Ângelo Cardoso, Paul Couturier, and Emma J McCoy. Datasets for online controlled experiments. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021; arXiv:2111.10198*, 2021.

Paul J. Northrop. *R package bang: Bayesian Analysis, No Gibbs*, 2020. URL `https://cran.r-project.org/web/packages/bang/index.html`. R package version 1.0.1.

Paul J. Northrop. *R package rust: Ratio-of-Uniforms Simulation with Transformation*, 2021. URL `https://cran.r-project.org/web/packages/rust/index.html`. R package version 1.3.12.

Melvin R. Novick, Charles Lewis, and Paul H. Jackson. The estimation of proportions in m groups. *Psychometrika*, 38(1):19–46, 1973.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, 2019.

Thomas S. Richardson and Yu Liu. *Hierarchical Bayesian Analysis with Bang*, 2022. URL `https://github.com/amazon-research/hierarchical-bayesian-analysis-with-bang`.

Jonathan C. Wakefield, Alan E. Gelfand, and Adrian F. M. Smith. Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, 1(2):129–133, 1991.

Supplementary Material: A Bayesian Model for Online Activity Sample Sizes

A ADDITIONAL EXPERIMENTS

Prompted by a suggestion from the reviewers, we added two additional baseline models for comparison, a Neural Network model and a Censored Weibull model. In addition, we compared the performance of all these methods on a recent public dataset containing online experiments (Liu et al., 2021); among the 78 experiments in this set, 10 (respectively, 8) experiments contain complete daily sample sizes from day 1 to 14 (respectively, 28). We show the results of predictions on these experiments in Table 2; given the small size of this public dataset, the Random Forest and Neural Net models were trained on the proprietary data.

In more detail, the Censored Weibull model (Kay, 1977) was fitted via maximum likelihood with data from week 1 (plus n_0 estimated by (12)). The Neural Network (Paszke et al., 2019) used Relu as activation function and 3 hidden layers with 16, 32 and 16 nodes respectively; 10% training data was used as a validation dataset to tune other parameters.

The results from Table 2 confirm that the proposed Bayesian model is competitive against all the other methods considered. We also note here that the Neural Net method requires greater computational resources than the other methods considered and hence is less scalable.

Lastly, following another suggestion from the reviewers, in the last two rows of the Table we compare results for the posterior mean and median of the Bayesian method. In our experiment they are statistically indistinguishable.

Table 2: Additional Results On Predictive Performance: RMSE and MAPE for the Log Linear, RF (random forest), NN (neural network), Censored Weibull and the proposed Bayesian method on public and proprietary data. **Bold** (underline) are best (second best). For Log Linear, Censored Weibull and Bayesian, we use the number of new samples for each day in week 1 to train and forecast the number of new samples in week 2 and week 4 (Weibull and Bayesian used $\lambda = 10$). Due to the small size of the public dataset, for RF and NN, proprietary data were used to train the models.

Method	Public Data				Proprietary data used in paper (additional results for NN and Weibull)			
	Week 2		Week 4		Week 2		Week 4	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
RF + **training data**	6.48e+5	22.78%	1.79e+6	50.43%	7.05e+05	44.06%	5.49e+05	165.22%
NN + **training data**	1.10e+6	47.58%	1.92e+6	115.86%	**5.64e+05**	4.65e+3%	5.81e+05	490.41%
Log-linear	**1.12e+5**	19.06%	6.86e+5	67.93%	1.84e+11	5.72e+5%	9.06e+17	1.12e+13%
Censored Weibull	4.36e+5	32.42%	**3.09e+5**	70.97%	1.85e+06	49.17%	1.41e+06	139.68%
Bayesian (mean)	1.59e+5	12.79%	5.09e+5	15.24%	7.11e+05	32.57%	5.01e+05	84.55%
Bayesian (median)	1.59e+5	**12.78%**	5.09e+5	**15.25%**	7.11e+05	**32.59%**	**5.01e+05**	**84.57%**