Comparative Study of Probabilistic Models for Dietary Intake Data

Ayona Chatterjee *, Santosh Gummidipundi, Henry Lankin, and Christine Marachi

Department of Statistics and Biostatistics, California State University East Bay, CA 94542

Abstract

An important area of food safety risk assessment involves monitoring intake of pesticides through the diet. Studying dietary intake of pesticide involves modeling intake of various foods. Dietary data obtained on consumption of certain food products may have a large number of zeros. This is observed while monitoring consumption of products consumed infrequently such as peaches and strawberries which often feature in the dirty dozen list. Distribution of intakes for such products have a peak at zero which need to be accounted for while modelling such data. Most dietary intake data such as the NHANES provide consumption values for only two days. Also consumption of certain foods may be correlated. In this paper we compare two models to account for the issues above using a Bayesian framework; a propensity model and a latent Gaussian model. The propensity-model is a two-stage model which first assigns each individual a certain probability of consumption for a product and then we model the non-zero consumption. We also develop a latent Gaussian model for the data with the additional assumption that consumptions between foods may be correlated. We compare predicted values from our Bayesian models with those observed. We also discuss extending our models for predicting long-term consumption patterns to study chronic risk.

Key Words: Dietary Data, Modeling zero intakes, Bayesian analysis, Risk prediction.

1. Introduction

Exposure assessment of risk from pesticide includes studying various pathways through which a human is exposed to pesticide. The primary pathways include the diet, water and air. Dietary exposure assessment includes studying intakes of various pesticides, contaminants and nutrients through food. Based on the type of intake, a low or a high amount of intake is of interest to us. For example with most pesticides, high intakes are of concern but with nutrients such as vitamin A both low and high intakes have negative impact on health and are important to assess. Exposure assessment also involves studying both acute and chronic risks. For dietary intake data, acute risk is associated with single intake of a large dose and the chronic risk is associated with long term intake of small levels of of pesticide or

*Corresponding author: Telephone:(001) 510-885-4133, Email: ayona.chatterjee@csueastbay.edu

contaminant.

To study risk from ingestion of a pesticide or multiple pesticides involves combining information from "consumption" and "concentration" data set (Kroes et al. [2002]). For this study the data which provides information about consumption of produce by an individual is called the "consumption" data set. The data set which gives us information for levels of various pesticides on these produce is called the "concentration" data set. Various models from empirical to probabilistic have been suggested to combine data from both data sets to generate the distribution of the intake of the pesticide or contaminant of interest. Probabilistic models give more robust estimates for estimating proportion of the population at risk from unsafe levels of consumption and Lunchick [2001] lists advantages and disadvantages of using deterministic and probabilistic modelling.

Most dietary consumption data are positively skewed with a large proportions of zeros. Consumption data sets provide information about an individuals intake of food and drinks over a period of 2-7 days. The zero intake may represent a true zero implying that the person never consumes that produce or in many cases just be because the individual did not consume that particular produce on the days of the survey. Since we are using the data to model both acute and chronic risk these zero-intakes should be taken into consideration while developing a model for such dietary data or one may omit important features of the data set.

In this paper we focus on modeling "consumption" data and present two model which can incorporate the zero-intakes and predict daily and long-term intakes well. We present our models using a Bayesian framework to study consumption of strawberries from the National Health and Nutrition Examination Survey (NHANES) data set.

Several models such as in Johnson et al. [1992] and Berk and Lachenbruch [2002] have been developed to account for high proportions of zeros in the data. Zero-modified distributions which is usually a combination of a discrete distribution with the degenerate distribution with all probability concentrated on the origin is one possible approach. Another approach commonly used in econometrics is the Tobit model which was introduced by Tobin [1958]. We develop two possible model to study consumption data sets with high proportions of zeros and compare them. We call Model I the propensity model based on the individuals' frequency of consumption of a produce and Model II is a Latent gaussian model.

2. Materials and Method

The data set used for this study is part of the NHANES data set. The NHANES is a major program of the National Center for Health Statistics which is part of the Centers for Disease Control and Prevention (CDC)(NHANES [2017]). The program is designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physi-

cal examinations. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5000 persons each year. These persons are located in counties across the country, 15 of which are visited each year. The data collected from this program is used for a variety of purposes such as assessing effectiveness of existing nutritional programs, finding prevalence of diseases, epidemiological studies and health sciences research to name a few. Since the NHANES has information about a huge number of items consumed by individuals, we decide to only consider the ones that are in the "dirty dozen".

The "dirty dozen" is a list of most contaminated produce observed in a given year (Dirty Dozen [2016]). The Environmental Working Group (EWG) analyzes test results of more than 35,200 samples of fruits and vegetables taken by the U.S. Department of Agriculture and the Food and Drug Administration. To compare the fruits and vegetables, the group came up with a composite score for each type of produce based on six measures of contamination. Some of the measures include the percent of the sample tested with detectable pesticides and the average number of pesticides found on a single sample. Top two most contaminated produce were strawberries and apples. These have been in the top five for the last five years and consistently make it to the "dirty dozen" list. In this study we produce results based on strawberries. We combine NHANES data from 2009 – 2014. We have 14428 individuals' intake for strawberry reported over two days along with the time of year when the study was conducted. Gender, age and BMI is available for each individual in the study. Along with the actual amounts of strawberries consumed, we also know the daily frequency of consumption. Thus an individual with a frequency of 0 will imply that the person did not consume any strawberry on that day but a frequency of two will imply that the person ate strawberries twice on the given day.

2.1 Data Summaries

Figure 1 shows the histogram for the frequency of consumption days for strawberry for all individuals and for both day. The spike at zero represents the large number of non-consumption days. On the right of Figure 1 is the same as the histogram but we zoom in to see the non-zero frequency days. Most times people have strawberry once a day, we have few individuals who consumed strawberries 4 or more times in two days. Here a 0 frequency may not imply that the person does not eat strawberries ever but maybe just not on the days the study was recorded. Figure 2 is the histogram of the actual consumption values of strawberries in grams. Again the peak at zero hides the rest of the feature. As before when we zoom in to the non-zero part of the histogram we notice the right skew of the data. We have a few very large intakes. The tail is of importance as large consumptions will lead to large pesticide exposure.

Table 1: Summary statistics of non-zero strawberry consumption in grams by gender and season

Group	Mean non-zero consumption	Median non-zero consumption	$\%$ Zero intakes
Male-Season 1	72.49	55.33	97.7
Male-Season 2	86.87	63.27	96.1
Female - Season 1	74.65	54.00	95.7
Female - Season 2	77.41	60.00	93.8

Next we explore the difference in strawberry consumptions between gender and season. We treat months from September through February as season 1 and March through August as season 2. Thus season 1 represents the winter months and season 2 the summer months. Since strawberries are more in season during season 2, we would expect to see a large intake during that time. This is indeed reflected in Table 1, both men and women have a larger mean and median intake of strawberries during the summer months. The proportion of non-consumption days are also less in season 2 than season 1 for both men and women.

Over all for the Strawberry consumption data we have 95.7% of the consumption values recorded as zero over the two-days. We want to use the information provided by the frequency of consumption along with the amount consumed while developing a probabilistic model for the strawberry consumption values. Since we also want to include possible gender, age and season effect; we develop hierarchical Bayesian models for the data set.

2.2 Propensity model

To model the zeros and non-zero intakes, we first model if the individual consumes the produce on a particular day; if yes then we define a distribution to determine how much the individual eats. The frequency of consumption is assumed to be from a Poisson distribution with mean π_{ij}. The distribution of π_{ij} depends on gender and season and we call it the propensity for consumption. Since the data set has the information about the frequency of consumption, we assume $E(\pi_{ij})$ is the observed frequency of consumption. The concept of consumption used here is similar to that as given in Carriquiry [2003] except here we model it as a Poisson variable rather than a Binomial.

To model consumption on days with non-zero propensity, we denote the notional response for an individual i on a given day j by α_{ij} which depends on the observed value for each individuals' gender, season the observation was taken, the age and BMI of the individual. For simplicity we assume the relationship is linear. We combine the gender and two seasons in to a single factor which can take values

from 1 to 4 where 1 is for males in season 1, 2 is for males in season 2, 3 is for females in season 1 and 4 is for females in season 2. The regression relationship is similar to the one developed by Myles et al. [2003] and is given in equation 1.

$$\alpha_{ij} = \beta_1 * gender - season[i] + \beta_2 * age[i] + \beta_3 * bmi[i] + \epsilon_i + \varepsilon_{ij} \quad (1)$$

where $\epsilon_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma_w^2)$.

We define the actual observed daily strawberry consumption for an individual i on day j as $a_{ij} = \alpha_{ij}$ if $\pi_{ij} > 0$ or else if $\pi_{ij} = 0$ we set the observed value to be zero. Here $i = 1, \ldots, 14428$ and $j = 1, 2$.

To handle the sparseness of the data set and the dependence of intake on gender and season we work on a Bayesian hierarchical model. The model presented in the paper assumes that each individual has a within individual variability which is given by σ_w^2 for the intakes between days and then we have the variability between individuals given by σ_b^2. If we have more than two days of data available for each individual, we can let the within-individual variability differ from person to person, and let σ_{wi}^2 dependent on the individuals demographic such as age and or gender.

2.3 Prior distributions for the model parameters

In attempting to develop a realistic model for the dependence of strawberry intakes on individual's gender, age, bmi and season, we have introduced a large number of unknown parameters. Models such as Paulo et al. [2006] and Boon et al. [2004] propose Bayesian models for dietary data. In order to fit the model, we follow the increasingly popular practice of giving these parameters a probability distribution intended to reflect knowledge of their likely values before examining the data. Thus we adopt the Bayesian approach to parametric inference, and specify a prior distribution for the parameters Gelman et al. [2004]. Under this approach, information on the parameters from the prior distribution and the data is combined in a posterior probability distribution.

Our choice of prior distribution reflects the hierarchical or multi-level structure of the data set, in which there is variation in the response over the two days within each individual, variation between individuals and variation between these gender-season groups. The recorded intakes are modelled conditionally on certain parameters, which are themselves modelled in terms of parameters, known as hyperparameters, corresponding to a higher level of the hierarchy Gelman et al. [2004].

Normal distributions are not appropriate for variances, and with one exception we follow the common practice of giving the reciprocal of each variance a Gamma prior distribution. Using $G(\alpha, \lambda)$ to denote a Gamma distribution with expectation α/λ and variance α/λ^2, we assign σ_1^{-2}. We fit the Bayesian mixture model using WinBUGS Spiegelhalter et al. [2004], which is a freely available program for Bayesian model fitting using Markov chain Monte Carlo simulations Gilks et al.

[1996]. The model is run for 30000 iterations with the first 5000 used as burn ins. Results from this model are compared to the Latent Gaussian model and presented in the next section (Cowles and Carlin [1996]).

2.4 A Latent Gaussian Model

In contrast with the Propensity model, the latent Gaussian model allows us to model both the occurrence and the amount of the measured quantity to be described by a single random variable. The latent Gaussian model as described by Allcroft and Glasbey [2003] assumes that zero observations are actually censored observations, smaller than a known threshold. Since for our data set we know the exact amount of intake by each individual, we treat the zeros as true zeros and set our threshold to be zero. The model assumes that there exists a transformation such that the non-zero part of the data fits the tail of a Normal distribution above the threshold.

The data here have a hierarchical structure. As before we define a Normal distribution to describe each individual's notional strawberry intakes. However we have a single transformation to all the non-zero intakes as opposed to having a transformation for each individual's non-zero intakes to fit the right tail of a Normal distribution. This is only an approximate method. Each individual's intakes are assumed to be from a Normal distribution which is left-censored at zero. As before we denote the notional response for an individual i on a given day j by α_{ij} where α_{ij} are from a Normal distribution with mean μ_i and within-individual variance σ_w^2.

Here we refer to μ_i as the expected notional response for individual i and depends on gender-season, age and bmi as given in equation 1. The actual strawberry intake a_{ij} for an individual i on day j is α_{ij} if $\alpha_{ij} > 0$ or it is set to zero. The likelihood for this model can be written as

$$L(a_{ij}|\mu_i, \sigma_w) = \{\prod_{a_{ij}>0} \sigma_w^{-1}\phi(\frac{a_{ij} - mu_i}{\sigma_w})\}\{\prod_{a_{ij}=0} \Phi(\frac{-\mu_i}{\sigma_w})\} \tag{2}$$

Here ϕ and Φ denote the probability density and cumulative distribution function of the standard normal distribution respectively.

We develop a hierarchical Bayesian latent Gaussian model for the responses. The gender-season effects are given Normal prior distributions, around 0 and have a variance of 100. Since most daily intakes are zero, we might expect μ_i to be negative. The between-individual precision σ_b^{-2} and the within individual precision σ_w^{-2} are both given a Gamma distribution $Ga(0.01, 0.01)$. We specify the censored observations in our model using the $I(lower, upper)$ function in WinBUGS. Since for our latent Gaussian model we assume the zeros to be censored we replace the 0's by NA in our data file. We also specify the lower and upper values between which the censored and uncensored observations lie. When censoring is specified the censoring node contributes a term to the full conditional distribution of its parents. Thus for censored observations the interval is $I(-\infty, 0)$ and for uncensored

observations it is $I(-\infty, 10000)$. WinBUGS allows only one limit in the interval to vary between individuals. Here we fix the lower limit to $-\infty$ and the upper limit is 0 or 10000 for censored and uncensored observations respectively. WinBUGS does not allow the varying limit to be infinity and hence we fix the upper limit for non-zero observations to be 10,000.

We used WinBUGS to obtain posterior parameter distributions of our model and also predicted daily and longer term intakes. The model was run for 30,000 simulations and among these the first 5000 simulations were discarded as burn-ins.

3. Results

The posterior estimates for some of the model parameters are given in Table 2. From table 2 we can see that BMI has the largest effect on strawberry intake. For both genders, season 2 has a larger posterior mean that season 1. We also studies history plots and MC error for convergence.

Table 2: Posterior means expectations for model parameters from Propensity Model.

Parameter	Mean	SE	MC Error
Male + Season 1	0.05	0.0038	$6.59X10^{-5}$
Male + Season 2	0.08	0.0049	$8.0X10^{-5}$
Female + Season 1	0.09	0.0052	$9.72X10^{-5}$
Female + Season 2	0.14	0.0059	$1.10X10^{-4}$
Age effect	0.18	0.0622	0.0050
BMI effect	1.20	0.1009	0.0081

The posterior distribution for the propensity parameter based on the gender and season is given in Figure 3. Recall the propensity parameter has a Poisson distribution and we can observe from the posterior that the frequency of zero consumption has the largest probability for all four groups.

The results from the Latent gaussian model are for the square root transformed data. The posterior expectations for the factors effects in equation 1 are in Table 3. The smaller the posterior mean value of the parameter effect, the larger is the posterior probability of getting an intake less than zero.

3.1 Sensitivity to the choice of prior distributions

Since some arbitrary decisions about the prior distributions are inevitable, we examine the effects on the posterior estimates of changing the parameter values of the prior distributions. The prior expectations for the β parameters were increased and decreased by 50% of the previously stated value. The prior expectation for

Table 3: Posterior means expectations for model parameters from Latent Gaussian Model.

Parameter	Mean	SE	MC Error
Male + Season 1	-13.28	2.86	0.106
Male + Season 2	-6.50	2.87	0.096
Female + Season 1	-2.21	2.82	0.083
Female + Season 2	5.639	2.77	0.076
Age effect	-0.20	0.14.	0.01
BMI effect	10.77	0.33	0.031

σ_b^2 and σ_w^2 was also increased and decreased by 50%. These changes in the prior distributions did not cause any substantial changes in the posterior expectations of the parameters. There was less than 5% change in the posterior expected means for the Normal distribution for the non-zero intakes. For the variances, an increase in the prior distribution parameters saw the value of σ_b^2 and σ_b^2 decrease by less than 10%.

3.2 Predictions from the model

The model's ability to predict intakes over one or more days is more important for our study than inferences about model parameters. In Figure 4 we compare the cumulative distribution function(cdf) of the data with the predictive cdf from both the Propensity model and the Latent Gaussian model in the original scale. There seems to be an overall good agreement between the observed and predicted values from both models. Both models seem to under-estimate the proportion of zero intakes. The predictions are generated for each gender-season group along with a randomly selected age and bmi for an individual.

From the predicted daily intakes we can find the predictive probability of zero intakes in gender-season group. Table 4 has these probabilities from the Propensity model and the Latent Gaussian model along with the observed proportions. Both models perform similarly and under estimate the proportion of zero intakes.

Parameter	Percentage of Zero Intakes		
	Data	Propensity Model	Latent Gaussian
Male + Season 1	97.7	96.0	93.7
Male + Season 2	96.1	93.3	93.2
Female + Season 1	95.7	92.7	92.1
Female + Season 2	93.8	91.4	91.6
Total	95.7	93.5	92.7

Table 4: Posterior predicted percentages of zero strawberry intake compared with observed value from data.

4. Discussion

Monitoring dietary intake plays an important role in studying pesticide exposure. Along with studying acute it is important to assess the chronic risk too. The 2-3 days dietary data is often used to predict long term consumption patterns. Including the information about the frequency of consumption for a produce along with the amount of consumption can provide more accurate and novel way to estimate chronic risk. Data for consumption are sparse and hence hierarchical Bayesian modeling is appropriate. By sampling from the posterior distribution of daily intakes and combining with concentration levels, we can obtain the distribution of a particular residue on a particular product. Probabilistic exposure assessment are known to yield more robust estimates than empirical sampling.

Future work involves looking at transformation to normality for the propensity model and transformation to the non-zero values for the latent gaussian model to improve fit. The results presented here have been only fitted to strawberries and we are currently applying the model to other produce such as apples and peaches. The Propensity model and the Latent Gaussian model can be extended to study simultaneous consumption of produce that may be correlated using a multivariate approach such as in Cornick et al. [1994] and Chatterjee et al. [2008].

References

D. J. Allcroft and C. A. Glasbey. Analysis of crop lodging using a latent variable model. *Journal of Agricultural Science*, 140:383–393, 2003.

K. N. Berk and P. A. Lachenbruch. Repeated measures with zeroes. *Statistical Methods in Medical Research*, 11:303–316, 2002.

E. P. Boon, S. Lignell, J. D. van Klaveren, and I. M. E. Tjoe Nij. Estimation of acute dietary exposure to pesticides using the probabilistic approach and the point estimate methodology. *RIKILT Institute of Food Safety*, Project No. 805.71.833.01, 2004.

A. L. Carriquiry. Estimation of usual intake distribution of nutrients and foods. *American Society of Nutritional Sciences*, 133:601S–608S, 2003.

A. Chatterjee, G. Horgan, and C. Theobald. Exposure assessment for pesticide intake from multiple foood products: A bayesian latent-variable approach. *Risk Analysis*, 28:1727–1736, 2008.

J. Cornick, T. L. Cox, and B. W. Gould. Fluid milk purchases: A multivariate tobit analysis. *American Journal of Agricultural Economics*, 76:74–82, 1994.

M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91: 883–904, 1996.

Dirty Dozen . Shopper's Guide to Pesticides in Products, 2016. `http://www.ewg.org/foodnews/dirtydozenlist.php.WbbEObKGPX4` [Accessed: Sept 2017].

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. New York: Chapman & Hall, second edition, 2004.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 1996.

N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate discrete distributions*. New York: John Wiley, second edition, 1992.

R. Kroes, D. Müllet, J. Lambe, M. R. H. Löwik, J. van Klaveren, J. Kleiner, R. Massey, S. Mayer, I. Urieta, P. Verger, and A. Visconti. Assessment of intake from the diet. *Food and Chemical Toxicology*, 40:327–385, 2002.

C. Lunchick. Probabilistic exposure assessment of operator and residential non-dietary exposure. *Annals of Occupational Hygiene*, 45:S29–S42, 2001.

J. P. Myles, G. M. Price, N. Hunter, M. Day, and S. W. Duffy. A potentially useful distribution model for dietary intake data. *Public Health Nutrition*, 6:513–519, 2003.

NHANES. National Health and Nutritional Examination Survey , 2017. https://www.cdc.gov/nchs/nhanes/index.htm [Accessed: Sept 2017].

M. J. Paulo, H. van der Voet, J. Wood, G. Marion, and J. van Klaveren. *Food and Chemical Toxicology*, 44:994–1005, 2006.

D. J. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS, version 1.4.1, 2004. BUGS 1996–2004: Medical Research Council (MRC) UK, WinBUGS. http://www.mrc-bsu.ac.uk.bugs/winbugs/contents.html.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

Figure 1: Frequency of strawberry consumption for all individuals over all days, with the non-zero frequencies zoomed in

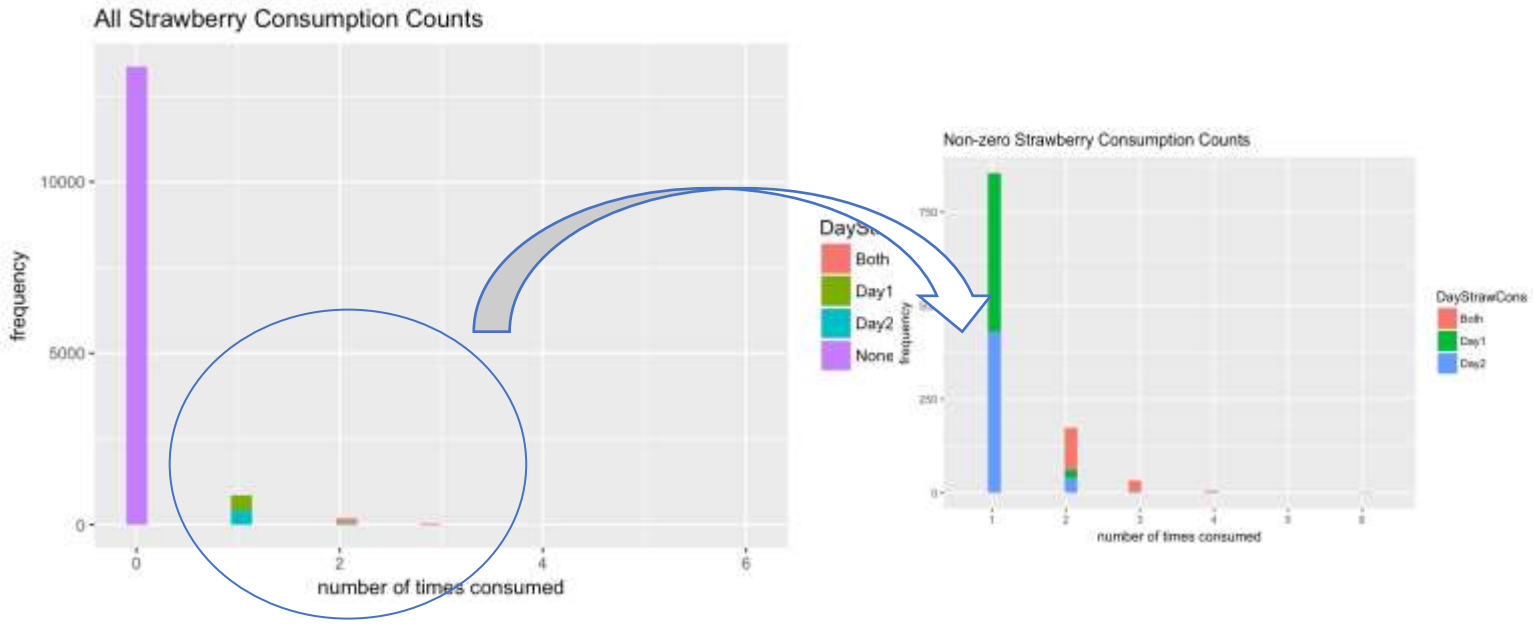

Figure 2: Histogram of daily strawberry consumption in grams for all individuals over all days, with the non-zero intakes zoomed in

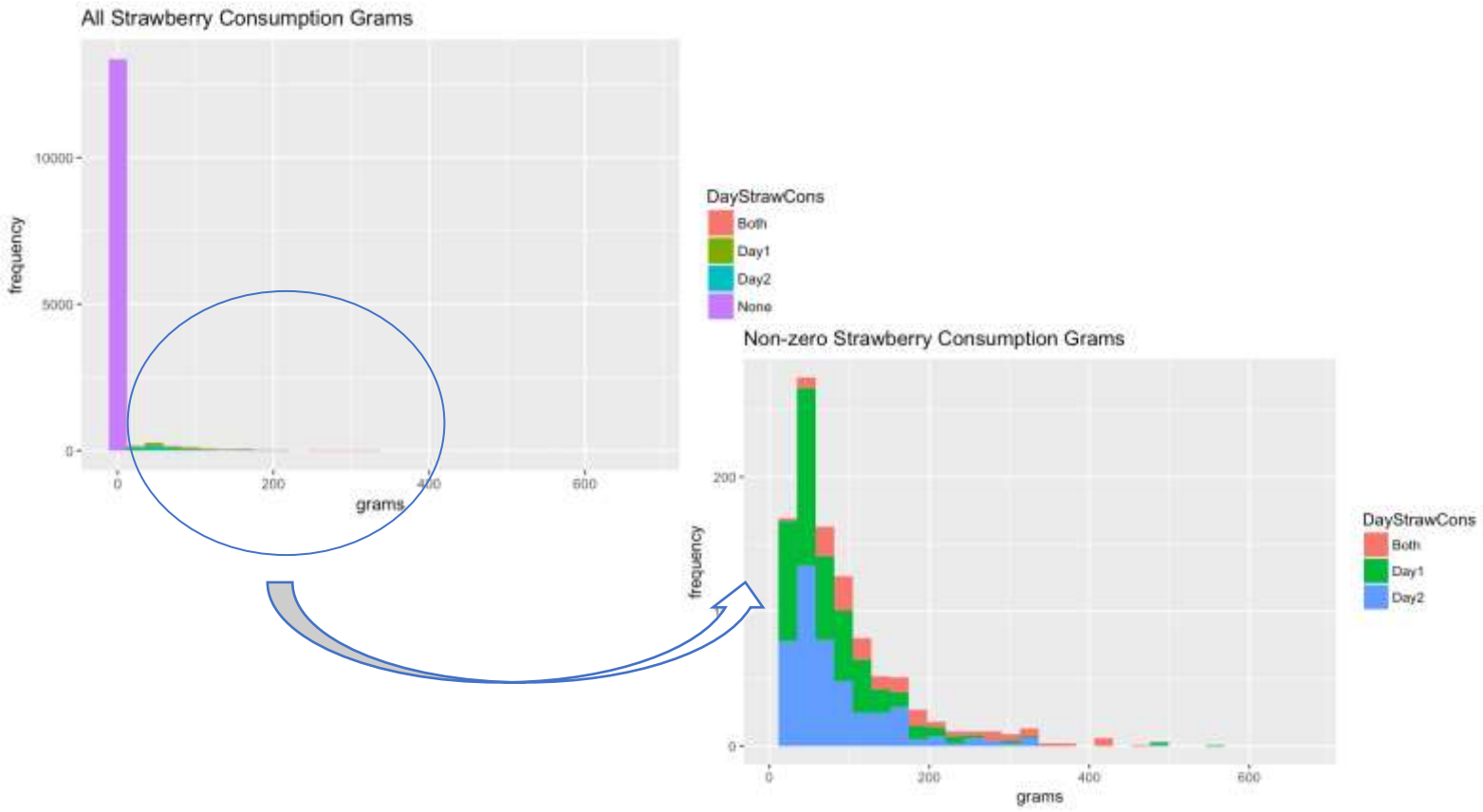

Figure 3: Comparison of empirical cdf of consumption values from the data with those generated using the posterior predictive distribution from the Propensity and Latent Gaussian Models