# Robust Modelling of Ordinal Survey Data Using Probabilistic Programming

Aleksi Lahtinen[1], James Rhys Edwards[2,3], Marc Calmbach[2], Isabella Tautscher[2], and Leo Lahti[1]

[1] Department of Computing, University of Turku, Turku, Finland
[2] SINUS-Institute, Berlin, Germany
[3] Institute for Information Law, University of Amsterdam, Amsterdam, Netherlands

**Abstract**

Surveys play a central role in much of the research conducted in the humanities and social sciences. A common data type encountered in surveys is the ordinal variable, which differs from nominal categorical variables. Several regression methods are available for analysing ordinal data, with the cumulative logistic model being one of the most widely used. However, ordinal survey data often present challenges, particularly in studies with small sample sizes, where some response categories and levels of explanatory variables can have low response rates. In such cases, classical statistical methods can produce unreliable or incomplete estimates. Here, we investigate the use of probabilistic programming, grounded in Bayesian analysis, as a more robust alternative for estimating category probabilities of ordinal variables and other model parameters. These models are better equipped to handle uncertainty and provide more reliable estimates, even in the presence of sparse data. We validate the approach with simulated data where the ground truth is known, and demonstrate the advantages of this approach by comparing it to its classical frequentist counterpart in the context of cultural participation and access survey.

**Keywords:** probabilistic programming, ordinal data, survey data, ordinal logistic regression

## 1 Introduction

Surveys are a common method for collecting information across various fields in the humanities and social sciences, including music studies [2; 8; 10], as well as in disciplines like psychology [14] and public health [6]. They are typically used to understand opinions, beliefs, or behaviours within a specific target population, such as the voting population of a country or attendees at a particular event. Surveys use probability sampling (or alternative sampling methods, such as quota sampling) to obtain estimates for the whole population while only surveying part of the population.

Survey data frequently include ordinal variables. Unlike continuous numerical data, they consist of ranked categories without assuming equal spacing between them [1]. Several statistical methods have been developed to analyse ordinal data, with classical frequentist approaches remaining the most commonly used [9; 11; 12; 16]. Many of these methods can also be implemented within a probabilistic programming framework [4]. In our previous work, we have demonstrated the application of the probabilistic framework in computational analyses of cultural production [19; 27; 28]. However, despite recommendations in the literature [4; 21], probabilistic approaches to analysing ordinal survey data remain relatively underused. This may be due, in part, to the limited familiarity with probabilistic models among application specialists and uncertainty about

Aleksi Lahtinen, James Rhys Edwards, Marc Calmbach, Isabella Tautscher, and Leo Lahti. "Robust Modelling of Ordinal Survey Data Using Probabilistic Programming." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 591–608. https://doi.org/10.63744/eCwMjQ976nWf.

whether probabilistic models can bring gains. Here, we demonstrate the validity of the probabilistic approach in a simulation study where the ground truth is known, and demonstrate its potential benefits by comparisons with the more conventional approaches in the context of original real-world survey data on cultural participation and access.

Robust methods for real survey data are needed because these data sets are often incomplete. Surveys often suffer from missing data and low response rates. Collecting large numbers of responses can be expensive and resource-intensive. Subsequently, the sample sizes often remain relatively small limiting the reliability of statistical analyses. This problem can be even more emphasized for more refined subgroups within the overall sample collection. We explore the possibilities of probabilistic models in providing more robust estimates of the response rates and their uncertainty in such cases.

Another way to address limited sample sizes is to pool information across multiple sources. Survey data are often collected from various locations, such as schools, cities, or countries, creating a hierarchical structure where much of the variation is shared across these units [17; 20; 26]. Hierarchical modelling allows such data to be analysed within a unified framework, increasing the effective sample size while accounting for location-specific differences. However, it can be challenging when data within groups are limited or the number of groups is small. Therefore, we also compare the performance of classical and probabilistic methods in the hierarchical modelling of ordinal survey data.

Taken together, our results aim to validate probabilistic programming as a robust alternative for analysing ordinal survey data and demonstrate its potential for broad application in computational humanities and social sciences

## 2 Materials and methods

Probabilistic programming provides a flexible alternative to classical statistical analysis by enabling the construction of robust, Bayesian models. These models represent uncertainty through priors, initial beliefs about parameters that are updated with data to yield posterior estimates. This approach performs particularly well with small sample sizes, where traditional methods often fail to produce reliable estimates. Recent advances and tools such as Stan [5] have made probabilistic programming increasingly accessible across disciplines. Because ordinal variables require regression models that account for the ordered nature of categories, applying standard methods can lead to biased or misleading results [21]. In this study, we compare classical and probabilistic approaches to fitting such models and assess the potential advantages of the latter.

### 2.1 Data

In this study, we use cultural access and participation survey data collected for the OpenMusE project [24].[1] The project aims to make the European music industry more competitive, fair, sustainable and transparent. The survey dataset offers us a relevant venue for investigating the possible benefits of the probabilistic modelling alternatives when analysing ordinal survey data.[2]

[1] OpenMusE is funded by the European Union under Grant No. 101095295. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission's Citizens, Equality, Rights and Values Programme. Neither the European Union nor the granting authority can be held responsible for them.

[2] The data used in this analysis are interim, as the survey was still ongoing at the time of writing. However, comparison to secondary data on cultural access and participation (e.g., Eurobarometer 56.0, 79.2, and 88.1) confirms that the subgroup proportions are broadly reflective (if not yet representative) of population trends. The interim data are thus well suited for this methodological analysis. The final, representative survey data will be further analysed by the authors in a forthcoming paper.

The survey was administered across five European countries: Germany, Poland, Spain, Italy, and France. It includes a broad range of questions related to respondents' cultural participation and attitudes toward culture. For instance, participants were asked how frequently they attend various cultural events. Of particular interest to our study are the numerous ordinal-scale questions included in the survey, which serve as the primary focus of our analysis.

Before analysing the data, we divided the sample into three subgroups (Table 1). The first group, music audience, includes respondents who attended three to five or more live music events in the past twelve months. The second, music professionals, comprises individuals who earned income from music-related activities during the same period. The third, musicians, includes those who played an instrument or sang within the last year. These subgroups were defined to examine differences among groups with distinct levels of engagement with music and to demonstrate the methods across samples of varying sizes. The music professionals group is considerably smaller than the other two. Respondents identifying with another gender identity or choosing not to disclose their gender are also few, as are those aged 70 and over, particularly within the music professionals subgroup. In addition to comparing the two modelling approaches, we assess potential differences across the three groups and the overall sample.

| Variable | Category | Total responses | Music audience responses | Music professionals responses | Musician responses |
|---|---|---|---|---|---|
| Gender | Female | 1934 | 585 | 224 | 579 |
| | Male | 1719 | 601 | 234 | 485 |
| | Another gender identity | 6 | 4 | 1 | 2 |
| | Don't want to say | 3 | 1 | 0 | 2 |
| Age group (years) | 16-29 | 801 | 400 | 215 | 369 |
| | 30-39 | 471 | 194 | 90 | 170 |
| | 40-49 | 586 | 190 | 67 | 164 |
| | 50-59 | 732 | 203 | 60 | 172 |
| | 60-69 | 717 | 135 | 20 | 115 |
| | 70 and over | 355 | 69 | 7 | 51 |
| Country | Germany | 754 | 225 | 107 | 205 |
| | Poland | 565 | 118 | 53 | 155 |
| | Italy | 894 | 344 | 120 | 304 |
| | Spain | 654 | 221 | 101 | 182 |
| | France | 795 | 213 | 78 | 222 |

**Table 1:** Demographic breakdown of the cultural access and participation survey data. "Music audience responses" refers to respondents who attended live music events at least 3–5 times in the past twelve months. "Musicians responses" includes those who played a musical instrument or sang during that period. "Music professionals responses" comprises respondents who earned income from any music-related activity in the same time frame.

### Ordinal variables analysed

As previously noted, the survey includes several ordinal variables suitable for our demonstrations. To keep the scope of this paper manageable, we selected two such variables that explore respondents' agreement with statements about music's role in both personal and broader social, cultural, and political contexts. The specific statements analysed are:

- "Music is/was an important part of life in my family."

- "Music connects me with people from different cultural backgrounds."

Both statements were answered using the same five-point Likert scale (Table 2). We chose these variables for our analysis because they share a common response scale that is widely used in surveys. Additionally, these questions let us explore whether subgroups differ in how they answered these statements.

Some demographic categories, such as respondents aged 70 and over in the music professionals subgroup, have a low number of observations (Table 1). This sparsity can pose challenges for classical modelling approaches, particularly when combined with item nonresponse in the ordinal variables (Table 2). Additionally, because the survey data were collected across five different countries, it may exhibit systematic variation that should be accounted for during the analysis. This could be done based on hierarchical models, which aim to distinguish between the shared and dataset-specific variation in order to obtain more robust estimates of the entire data collection. Classical models often struggle with such hierarchical data, especially when the number of clusters (e.g., countries) is small. We will investigate the potential benefits of the probabilistic modelling approach in these situations, but also in general modelling of ordinal variables.

The cultural access and participation survey includes numerous other ordinal variables, many using a five-point Likert scale, while others employ different response formats. The analysis of these other variables is not included here in order to keep the scope of the paper manageable.

| Statement | Subgroup | Answer options | | | | | No response |
|---|---|---|---|---|---|---|---|
| | | Strongly disagree | Somewhat disagree | Neutral | Somewhat agree | Strongly agree | |
| Music & family life | Total | 132 | 173 | 473 | 568 | 425 | 1891 |
| | Music audience | 21 | 47 | 119 | 180 | 199 | 625 |
| | Music professionals | 10 | 31 | 43 | 63 | 70 | 242 |
| | Musicians | 22 | 44 | 118 | 174 | 170 | 540 |
| Music connects cultures | Total | 137 | 143 | 523 | 611 | 357 | 1891 |
| | Music audience | 17 | 46 | 138 | 201 | 164 | 625 |
| | Music professionals | 12 | 15 | 44 | 75 | 71 | 242 |
| | Musicians | 20 | 44 | 141 | 194 | 129 | 540 |

**Table 2:** Response counts for the different categories of the ordinal variables analysed in the total sample and the different subgroups. There is noticeable amount of item nonresponse to both statements.

### 2.2 Cumulative logistic model

As we mentioned earlier, surveys commonly include ordinal variables, which often appear in the form of Likert-scale questions. For example, a question like "How satisfied are you with your life?" might offer responses such as "Very dissatisfied," "Dissatisfied," and so on. Ordinal variables represent a type of categorical data where the categories possess a natural order or ranking [21]. However, the intervals between categories are not necessarily equal, for instance, the difference in sentiment between "Disagree" and "Strongly disagree" may not be the same as that between "Agree" and "Strongly agree."

When using regression modelling to analyse ordinal variables, it is essential to account for the ordinal structure of the data. Applying models designed for metric (continuous) outcomes can lead

to inflated false positive rates, biased effect sizes, and other issues [21]. Such shortcomings can be particularly problematic if the results are intended to inform data-driven policy-making. One widely used approach for ordinal data is the cumulative logistic regression model (also known as the proportional odds model) [4]. This model does not predict the ordinal outcome directly. Instead, it introduces an unobserved continuous latent variable. This variable is modelled through the cumulative probabilities associated with the observed categories:

$$P(Y \leq j) = P(\tilde{Y} \leq \alpha_j \mid X) = F\left(\alpha_j - \sum_{i=1}^{p} \beta_i x_i\right), \quad j = 1, \ldots, J-1,$$

where $F$ is the cumulative distribution function (CDF) of the error term [1; 4]. An observation falls into category $j$ or below if the latent variable does not exceed the threshold $\alpha_j$. When $j = J$, the cumulative probability is $P(Y \leq J) = 1$. A common choice for $F$ is the logistic distribution, which leads to the use of the logit link function. This gives the linear predictor the following form:

$$F^{-1}(P(Y \leq j)) = \text{logit}(P(Y \leq \alpha_j)) = \alpha_j - \sum_{i=1}^{p} \beta_i x_i, \quad j = 1 \ldots, J-1. \tag{1}$$

Unlike in standard linear regression, the cumulative model includes $J - 1$ intercept terms, or threshold parameters. Often, the focus is on category-specific probabilities, which are obtained as differences between cumulative probabilities:

$$P(Y = j) = P(Y \leq j) - P(Y \leq j-1), \quad j = 1, \ldots J.$$

In our models, we include two categorical explanatory variables: gender (with male as the reference group) and age group (with 16–29 as the reference group). These variables were chosen because their differences are often of interest in the humanities (and tend to impact cultural participation: see, for instance, Eurobarometer 79.2). The ordinal response variables are measured on a five-point Likert scale ($J = 5$), yielding four threshold parameters. The regression coefficients $\beta_i$ are interpreted as log-odds, representing the effect of each explanatory variable on the cumulative probabilities.

For the threshold parameters, we specified distinct normal priors to allow flexibility and avoid constraining threshold spacing: $\alpha_1 \sim N(-1.4, 1)$, $\alpha_2 \sim N(-0.4, 1)$, $\alpha_3 \sim N(0.4, 1)$, and $\alpha_4 \sim N(1.4, 1)$. These priors imply equal likelihoods for all outcome categories when explanatory variables are zero. Regression coefficients were assigned $\beta_i \sim N(0, 1)$ priors, assuming small but plausible effects.

The cumulative logistic model relies on the proportional odds assumption, which states that the effect of the explanatory variables remains constant across all categories $j$. Violating this assumption can lead to model misspecification. Several classical methods exist for testing the assumption, including the Brant, Score, and Wald tests [22]. Within the probabilistic framework, one can evaluate the assumption using leave-one-out cross-validation (LOO-CV) [29]. This approach involves comparing the cumulative model to a more flexible alternative that allows category-specific effects. If the alternative model yields a better fit, it indicates that the proportional odds assumption may not hold [4].

The cumulative logistic model appropriately accounts for the ordered nature of ordinal survey data by introducing a latent continuous variable and estimating cumulative probabilities.

### 2.3 Hierarchical cumulative logistic model

Survey data is often collected across multiple levels, for example, from different schools, cities, or counties [17; 20; 26]. Such data is referred to as hierarchical or multilevel, and this structure

should be accounted for in the modelling process. Ignoring the hierarchical structure can obscure meaningful differences between groups, potentially leading to inaccurate or biased results [23]. Therefore, we take the hierarchical modelling approach in this paper to properly capture variation at each level of the data, allowing for more accurate parameter estimates and a clearer understanding of both group-level and individual-level effects. Also, the hierarchical structure is often the reality with survey datasets, which also motivates taking this approach.

Hierarchical models take into account the multilevel structure of the data. They are designed to learn from both individual groups and the overall population, sharing information across clusters based on the observed variation between them [15]. This process, known as partial pooling, allows for more stable and accurate estimates, especially when group sizes are imbalanced. Hierarchical models handle sampling imbalances effectively, ensuring that groups with larger sample sizes do not disproportionately influence the inference [23]. Another key benefit is that they provide explicit estimates of between-group variation, such as differences in effects across clusters. In general, when the structure of the data supports it, using hierarchical models can lead to more precise and more robust inferences.

We focus here on two-level hierarchical models, although hierarchical survey data can, and often are, structured at several levels. For example, countries can represent one level, while individuals responding to the survey within those countries form another. In such cases, one level is nested within another. In addition to varying-intercept models, another important class of models includes varying-slope models [15]. These allow for the effect of an explanatory variable to differ across clusters, enabling the analysis of whether and how relationships vary between clusters.

The cumulative logistic model can be easily adapted into the hierarchical framework. This is done by adding a cluster-specific intercept or slope term in to the model Equation 1. We will be using the varying-intercept model, which means that every cluster in the data, country in the case of the data used here, has its own intercept term. Now the linear predictor of the model is

$$\text{logit}(P(Y_c \leq \alpha_j)) = \alpha_j + u_c - \sum_{i=1}^{p} \beta_i x_i, \quad j = 1 \ldots, J-1, c = 1, \ldots, C, \tag{2}$$

where $u_c$ is the cluster-specific intercept term. The intercept term has its own prior distribution, which is $u_c \sim N(0, \sigma)$, and the hyperparameter $\sigma$ has its own hyperprior $\sigma \sim \text{half-normal}(0, 1)$. The half-normal distribution is a normal distribution constrained to be non-negative. We selected this hyperprior to impose greater regularization, which is important given the small number of clusters in the data.

To summarise, hierarchical modelling accounts for the multilevel structure of survey data and yields more reliable estimates when group sizes differ. It also allows examination of both differences and commonalities across groups.

### 2.4 Probabilistic programming

As we previously discussed, the frequentist approach to modelling ordinal survey data can encounter several difficulties, particularly in studies with small sample sizes. For instance, when certain response categories receive no observations, the frequentist method may fail to properly estimate all threshold parameters (Equation 1) and corresponding category probabilities. Similarly, it can struggle to estimate the effects of categorical explanatory variables with few observations, as well as their associated category probabilities. Moreover, the frequentist approach often performs poorly when fitting hierarchical models (Equation 2) with only a few clusters.

Probabilistic programming provides a more robust and flexible alternative that addresses many of these challenges faced by the frequentist approach. One key benefit is the ability to incorporate prior knowledge into the model through the use of priors. This prior information is combined with

observed data to form a posterior distribution, which enables more informative inference [23]. It also allows for a more comprehensive quantification of uncertainty regarding the model and its parameters. Additionally, the probabilistic framework facilitates the construction of more flexible and complex models. In scenarios where frequentist methods may fail to converge or yield unstable estimates, probabilistic models can often be fit successfully. However, probabilistic modelling is generally more computationally intensive and time-consuming than frequentist approaches. Additionally, fully utilizing probabilistic models typically requires a higher level of statistical expertise.

Probabilistic models offer a natural way to model hierarchical data. In the varying-intercepts model, the intercept terms are treated as random variables drawn from a common distribution, whose parameters, known as hyperparameters, are themselves given prior distributions called hyperpriors [15]. When a varying-intercepts model is combined with regularizing priors, it enables partial pooling, meaning that information is shared across clusters without assuming that they are identical [23]. This approach leads to shrinkage, where cluster-level estimates are pulled toward a global mean. On average, such estimates are more accurate than those from models with no pooling, where each cluster is treated independently [23]. The benefits are particularly pronounced for smaller clusters, which may otherwise yield unstable estimates.

The cumulative model defined in Equation 1 can be fitted using classical frequentist methods, which remains the standard approach in much of the applied literature [9; 11; 12; 16]. However, the advent of probabilistic programming tools has made it equally feasible to fit such models within a Bayesian framework. In this study, we fit the cumulative model using both approaches to compare their respective parameter category probability and parameter estimates.

Our analysis is implemented using Stan [5], which allows for the implementation of complex probabilistic models. Specifically, we implement the model described in Equation 1 using the R programming language [25] and the `brms` package [3], which provides a high-level interface for fitting probabilistic models. This includes the cumulative ordinal model and other regression models used to analyse ordinal data [4]. For comparison, the classical version of the model can be fit using the `ordinal` package [7]. Alternative package for fitting classical ordinal models is `MASS` package [30]. The complete source code for all analyses is available in an online repository.[3]

Overall, probabilistic programming combines prior information with observed data to yield more stable estimates and richer uncertainty quantification.

## 3 Simulation study

Before comparing the probabilistic and classical approaches using the cultural participation and access data, we conduct a brief simulation study. We use simulations to evaluate the performance of the two approaches in a controlled setting where the true data-generating process is known. Specifically, we compare how well each model estimates category probabilities across different sample sizes. We assessed model performance by calculating the mean squared error (MSE) between the real values and the estimates produced by each approach.

We adapted the code used to generate the simulated datasets from the implementation by Gambarota and Altoè [13]. We generated datasets of varying sizes, with sample sizes ranging from 20 to 400, and created 100 datasets for each sample size. The regression model used in the simulations was a simple cumulative ordinal model with a single binary explanatory variable and an ordinal outcome variable with five categories. All simulations used the same covariate effect, $\beta_1 = \log(2)$, and the same category thresholds. The probabilistic model was fitted using the priors specified in the previous section.

Our results show that the probabilistic approach consistently outperforms the classical model

[3] Link to Zenodo repository that contains the code and scripts used in the paper: https://doi.org/10.5281/zenodo.17453413

in estimating the real category probabilities, particularly at smaller sample sizes (Figure 1). When $N = 20$, the probabilistic model's category probability estimates are more tightly concentrated around the real values. As the sample size increases, the performance of both models converges, and from $N = 200$ onward, their estimates are largely similar. Nevertheless, even at larger sample sizes, the probabilistic model produces slightly narrower uncertainty intervals, indicating more precise estimates. It is important to note, however, that the MSE values for both models are quite small, especially at higher $N$, and the resulting inferences are practically very similar. While these differences may be modest in magnitude, the probabilistic model appears more robust overall.

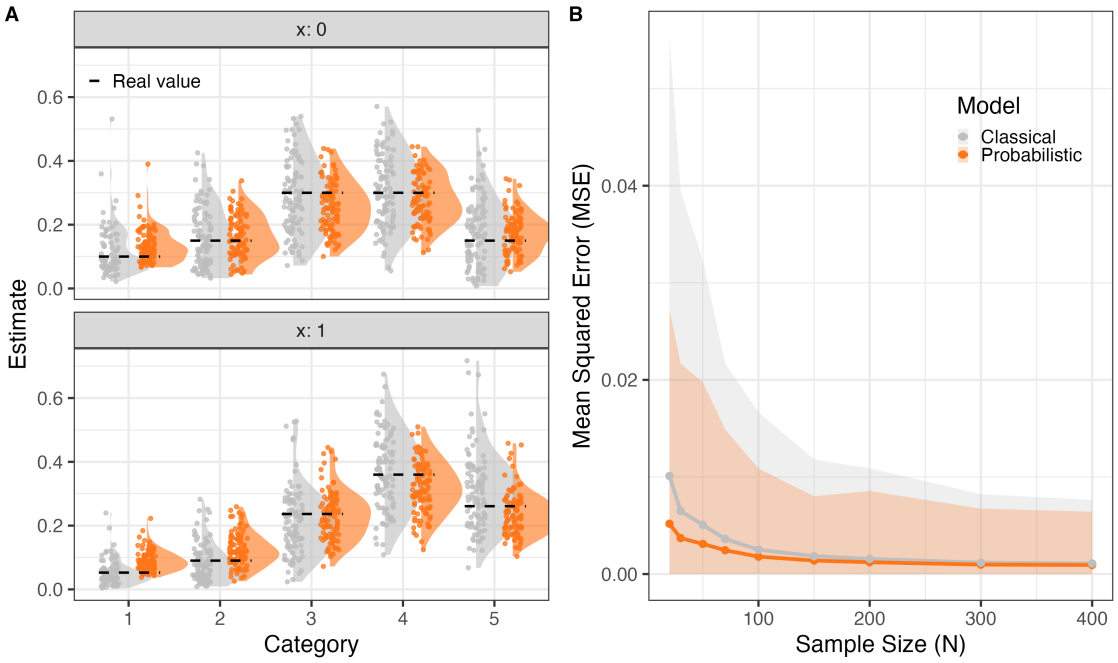


**Figure 1:** Performance of the two models across different sample sizes. (A) Category probability estimates from 100 simulated datasets with $N = 20$, shown at the different levels of the explanatory variable, for both models. The top plot displays the estimated category probabilities for observations where the explanatory variable $x = 0$, while the bottom plot shows estimates for $x = 1$. The probabilistic model's estimates are more tightly concentrated around the true values (indicated by both plots). (B) Mean squared error (MSE) values and their 95% intervals across different sample sizes ($N$).

## 4 Analysis of cultural access and participation survey

Let us next move to compare the two approaches using the cultural access and participation survey data (Table 1). To compare the two modelling approaches, we split the dataset into training and test sets. We fitted the models using the training set, while their performance was evaluated using the test set. Specifically, we calculated category probabilities in the test set and compared them to the corresponding model estimates. Model performance was quantified using MSE, calculated as the squared difference between the estimated and observed category probabilities. In addition to predictive accuracy, we also compared the coefficient estimates and their uncertainty intervals from both approaches.

For this evaluation, we used data from Germany as the hold-out test set and trained the models on data from the remaining countries. Germany was chosen because it represents a mid-range case in terms of both sample size and response distributions, offering a balanced and informative

basis for assessing out-of-sample performance. Holding out an entire country, rather than applying a random split across all observations, provides a more rigorous test of the models' capacity to generalize their estimates to a previously unseen data, here, a completely new city.

We tested the proportional odds assumption for all the cumulative logistic models using the LOO-CV approach described earlier in the paper. All models satisfied the proportional odds assumption. Additionally, for some of the probabilistic models, the target acceptance probability had to be increased from 0.95 to 0.99 to ensure stable sampling. Aside from this adjustment, model fitting proceeded without any issues.

### 4.1 Comparison of the modelling approaches

We used the LOO-CV to select the models that were fitted to the data. This method is implemented in Stan and `brms` and it allows for efficient way to compare probabilistic models to each other. While traditional LOO-CV is computationally expensive, requiring the model to be refit $n$ times, Stan implements a Pareto-smoothed importance sampling LOO-CV, which offers a fast approximation by reweighting posterior draws using importance sampling [29].

As we noted in the methods section, the explanatory variables of interest in this analysis are gender and age group. We used LOO-CV to compare models fitted to the full test set that included either one of these variables, both variables, or a hierarchical version. The hierarchical model that included both gender and age group provided the best fit for both statements. Classical frequentist versions of the models were also fitted and the results are compared to the probabilistic models.

We encountered convergence issues in some of the classical models during the fitting process. For the statement "Music is/was an important part of life in my family", we had to exclude certain observations due to sparse category counts. Specifically, for the "music audience" subgroup, observations with the gender identity another were removed; for "musicians", those selecting don't want to say were excluded; and for "music professionals", observations from the age group 70 and over were removed. In the model fitted to the whole population, both another and don't want to say gender categories had to be excluded. These problems arose because these categories contained only one or two observations, making it impossible for the classical models to estimate the corresponding parameters reliably.

For the statement "Music connects me with people from different cultural backgrounds," the hierarchical model could not be fitted for the "music audience" subgroup under the frequentist approach due to estimation issues. Consequently, we evaluated this subgroup using the non-hierarchical model instead. Under the probabilistic approach, however, the hierarchical model was successfully fitted. For the "music professionals" subgroup, the hierarchical model was used in both approaches, but in the frequentist model, observations from the age group 70 and over were removed due to insufficient data.

### 4.2 Category probabilities

We begin the model comparison by examining the estimated category probabilities and their associated intervals produced by the two approaches. First, we assess the overall estimates across the different subgroups to evaluate and compare the performance of each method. We also discuss any notable differences between the subgroups and the full population. Following this, we examine category probability estimates at different levels of the explanatory variables within specific subgroups.

#### Overall probability estimates

Although the classical model generally produces narrower intervals, the probabilistic model in general achieves better coverage of both the training and test sets across the subgroups and the

whole population for both statements (Figures 2 and 3). The intervals from the classical model often appear overly confident, whereas the probabilistic model more effectively captures uncertainty in the estimates.

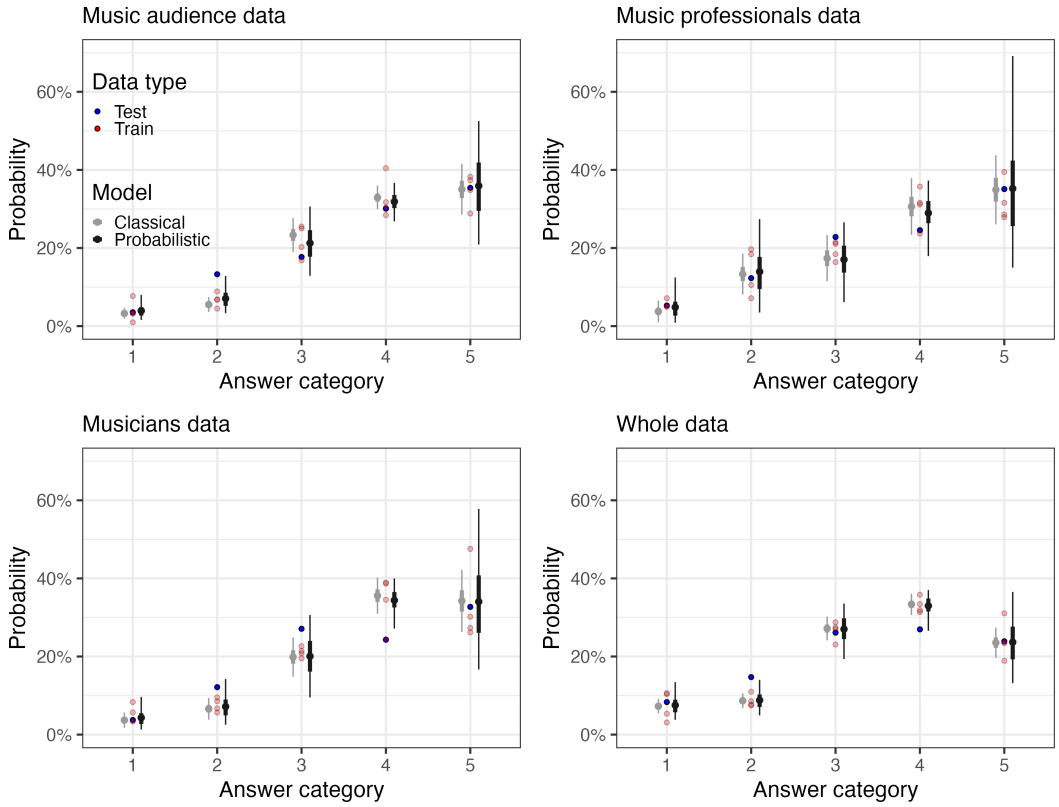


**Figure 2:** Comparing the fit between new data and model predictions for the statement "Music is/was an important part of life in my family". Estimated category probabilities from the probabilistic and classical models across the subgroups and the whole data. The intervals correspond to 50% and 95% credible and confidence intervals from the probabilistic and classical models. The red dots indicate the category probability values from the training set, and the blue dots represent the category probability values from the test set (Germany).

The MSE values are also lower for the probabilistic model across both statements and all subgroups, except for the full population in the statement "Music connects me with people from different cultural backgrounds," where the classical model performs slightly better (Table 3). For many subgroups, the MSE from the probabilistic model is clearly smaller compared to the classical model.

We observe notable differences in category probabilities between the subgroups and the whole population. For the statement "Music is/was an important part of life in my family" all subgroups show a higher proportion of responses in the fifth category, "Strongly agree" (Figure 2). In the "music professionals" and "music audience" subgroups, this category is has the highest estimated probability, while in the "musicians" subgroup, the fourth and fifth categories have similar probabilities. In contrast, responses from the whole population have higher probabilities in the third and fourth categories.

For the statement "Music connects me with people from different cultural backgrounds" differences between subgroups and the full population are also evident (Figure 3). The "music audience" and "music professionals" subgroups have a higher probability foor selecting category five. In contrast, the category probabilities for the "musicians" subgroup closely resemble those of the whole

| Statement | Subgroup | Probabilistic MSE | Classical MSE |
|---|---|---|---|
| Music is/was an important part of life in my family | Music audience | 0.0011 | 0.0020 |
| | Music professionals | 0.0011 | 0.0014 |
| | Musicians | 0.0036 | 0.0043 |
| | Whole data | 0.0015 | 0.0016 |
| Music connects me with people from different cultural backgrounds | Music audience | 0.0021 | 0.0031 |
| | Music professionals | 0.0069 | 0.0092 |
| | Musicians | 0.00097 | 0.0026 |
| | Whole data | 0.00079 | 0.00078 |

**Table 3:** MSE values for both statements across all subgroups and the full dataset, comparing the probabilistic and frequentist models. The probabilistic model yields MSE for all subgroups and statements, except for the full dataset on the statement "Music connects me with people from different cultural backgrounds," where the classical model performs slightly better. This indicates that the probabilistic model does better at estimating the category probabilities. MSE was calculated by comparing the estimated category probabilities to the observed proportions in the test set.

population.

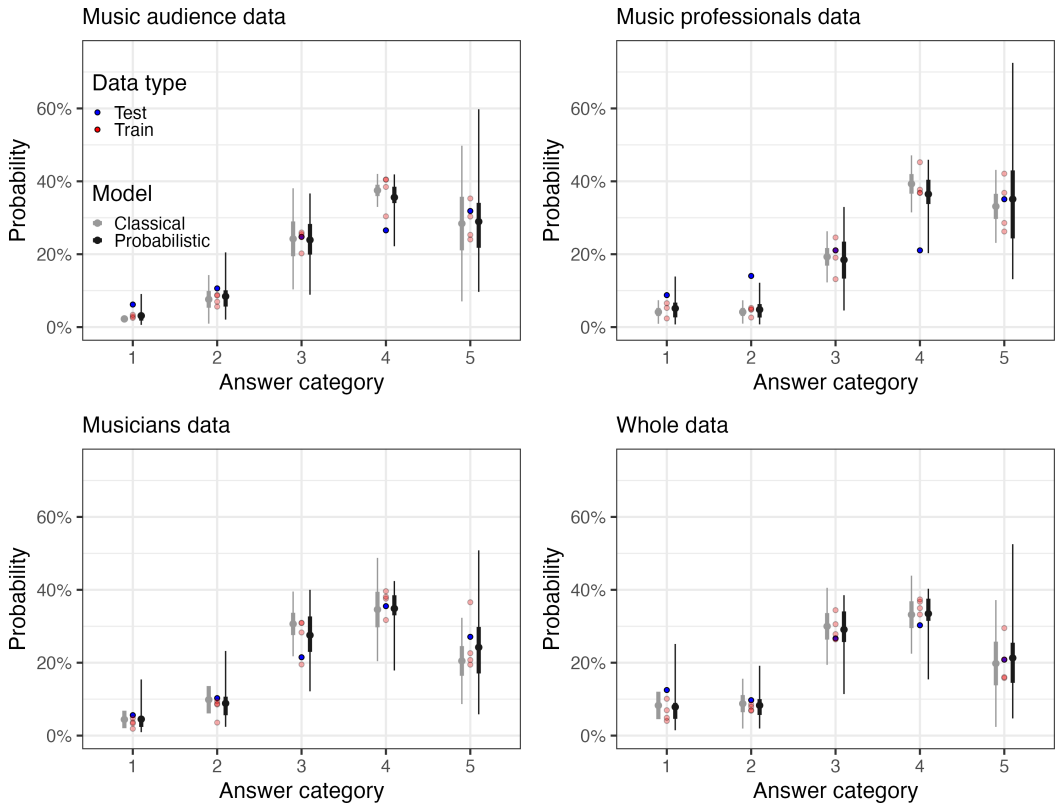


**Figure 3:** Comparing the fit between new data and model predictions for the statement "Music connects me with people from different cultural backgrounds". Estimated category probabilities from the probabilistic and classical models across the subgroups and the whole data. The intervals correspond to 50% and 95% credible and confidence intervals from the probabilistic and classical models. The red dots indicate the category probability values from the training set, and the blue dots represent the category probability values from the test set (Germany).

### Probability estimates at different levels of a categorical variable

We previously observed that the probabilistic model performs better in estimating overall category probabilities. We now turn to examining category probabilities across different levels of the explanatory variable age group. This variable includes six levels (Table 1), with the "70 and over" group having relatively few responses, particularly within the "music professionals" subgroup. The combination of a small sample size and item nonresponse results in only a single observation for this age group in that subgroup for both statements. As noted earlier, such sparsity causes estimation issues for the classical model.

Overall, both approaches perform similarly in estimating the category probabilities, with the probabilistic model yielding slightly wider intervals (Figure 4). More importantly, only the probabilistic model is able to produce estimates for the "70 and over" age group, which also align well with the observed values in the test set. This is the case with both of the statements.

An additional important point is that the confidence intervals from the classical model occasionally extend below zero at certain levels of the explanatory variable age group, rendering them invalid for probability estimates. For instance, this occurs for categories one and two within the 40–49 age group for the statement "Music connects me with people from different cultural backgrounds" (Figure 4). This issue arises because the classical model relies on approximate methods for interval estimation, which can sometimes produce improper intervals. In contrast, the proba-

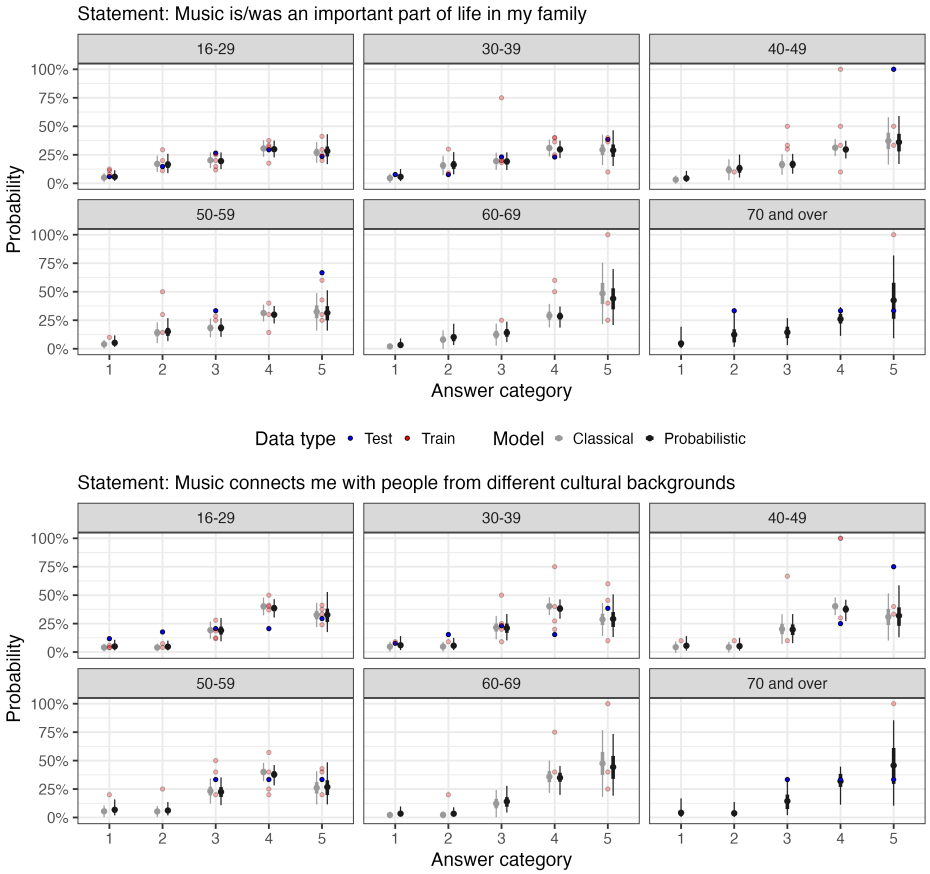


**Figure 4:** Comparing the fit between new data and model predictions at different levels of the explanatory variable age group. Estimated category probabilities from the probabilistic and classical models for both statements within the "music professionals" subgroup. The red dots indicate the category probability values from the training set, and the blue dots represent the category probability values from the test set (Germany). The 70 and over age group includes only a few observations, resulting in some categories lacking probability values.

bilistic model avoids this problem, as its intervals are derived from posterior predictive distribution samples, ensuring valid probability bounds.

### 4.3 Parameter estimates

The coefficient estimates from the probabilistic models are slightly more precise, with narrower intervals compared to those from the classical model (Figure 5). This modest gain in precision results from the use of prior distributions, which introduce mild regularization by shrinking the estimates toward zero. Additionally, the probabilistic model provides estimates for the "70 and over" age group coefficient for both statements within the "music professionals" subgroup, something the classical model fails to do due to data sparsity.

Although the estimates from the probabilistic models are slightly narrower, the overall conclusions regarding the parameter effects are consistent across both modelling approaches. A key advantage of the probabilistic approach is its ability to produce estimates for groups with sparse observations, where classical models can often fail. Furthermore, in probabilistic modelling, the significance of a coefficient is typically assessed using the highest density interval (HDI) and the

region of practical equivalence (ROPE) [18]. These methods make fuller use of the posterior distribution, allowing for more nuanced and informative inferences about the effects of the parameters.

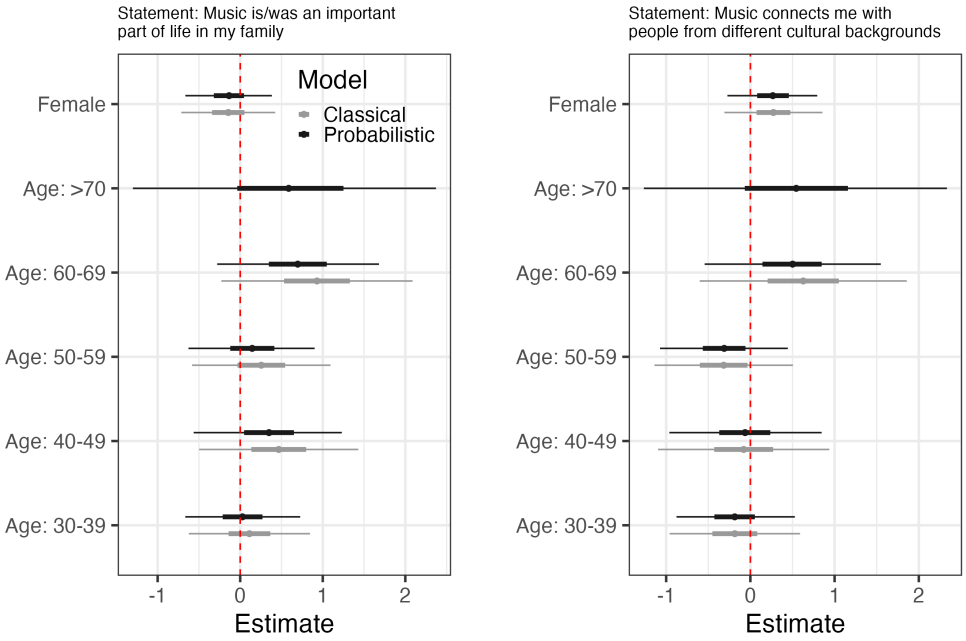


**Figure 5:** Coefficient (log-odds) estimates from the probabilistic and classical models for both statements within the "music professionals" subgroup. The intervals correspond to 50% and 95% credible and confidence intervals from the probabilistic and classical models. The reference categories are male for gender and 16-29 for age group.

### 4.4 Country-specific estimates

Because we fitted hierarchical probabilistic models, we are able to examine differences across the countries. In these models, country-specific intercepts influence the estimated category probabilities and their corresponding intervals. For the "music professionals" subgroup responding to the statement "Music is/was an important part of life in my family," there is little variation between the countries (Figure 6). However, for the same subgroup responding to the statement "Music connects me with people from different cultural backgrounds" some variation is evident. In particular, respondents from Italy and France appear less likely to select the fifth category compared to those in other countries.

Additionally, we can generate estimates for the Germany test set, which was not included in model training. To do this, we set the varying-intercept term, specific to each country cluster, to zero. This effectively treats Germany as having an mean profile across all countries, making the estimate a pooled average of the other country-specific estimates and their associated uncertainty intervals.

Country-specific estimates in the classical model can be obtained by including a country-specific random effect in the model formula and then calculating the corresponding estimates. However, deriving uncertainty intervals for these estimates is more challenging than in the probabilistic approach. Within the probabilistic framework, we obtained country-level estimates and intervals by grouping posterior draws by country, which allowed for a clearer and more coherent quantification of uncertainty.

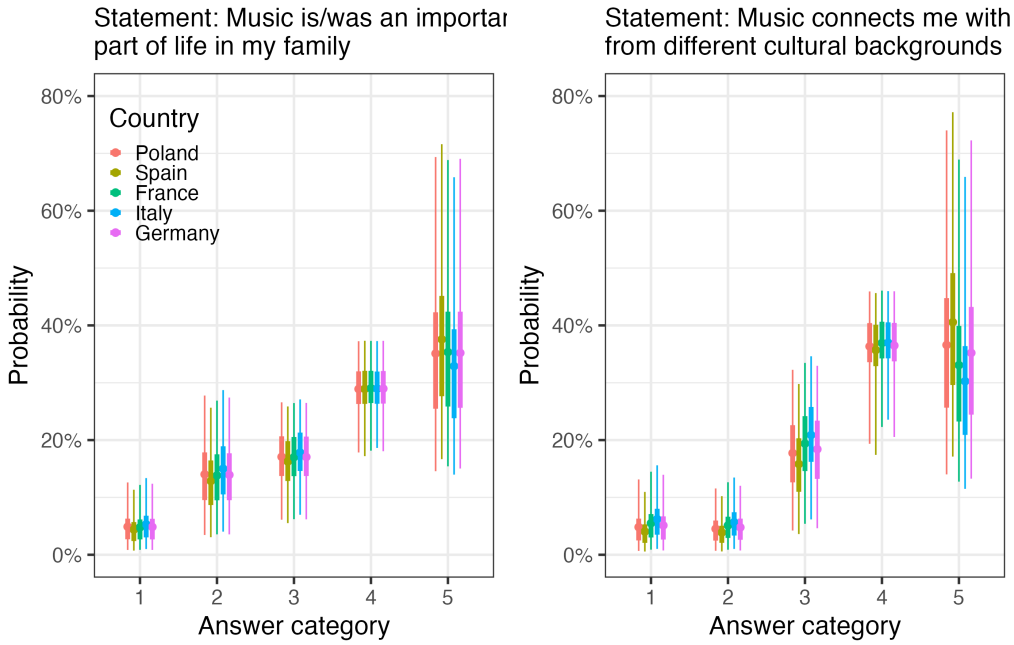


**Figure 6:** Country-level differences in estimated category probabilities for both statements within the "music professionals" subgroup, based on the probabilistic hierarchical model. Country-specific estimates exhibit variation across countries, particularly for the statement "Music connects me with people from different cultural backgrounds." Estimates are shown with 50% and 95% credible intervals for both statements.

## 5 Conclusions

Our results show that the probabilistic approach to modelling ordinal survey data can offer advantages over the classical approachin survey studies typical for computational humanities and social sciences applications. It yields more accurate category probability estimates, better coverage of observed values, and consistently lower MSE values. Although the absolute MSE improvements are modest, they remain statistically meaningful and particularly valuable for sparse or imbalanced subgroups where estimation is difficult. The probabilistic model also provides stable estimates for categorical predictors with limited data and consistently converges when fitting hierarchical models, unlike the classical model, which sometimes failed to do so. The probabilistic approach also enables robust and flexible analysis of country-level differences. From a cultural policymaking perspective, these advantages are matter, as sparsely represented groups, such as older adults or gender-diverse populations,are frequently under-represented in mainstream cultural activities. Reliable data for these groups are essential for developing equitable and effective policy interventions.

Our simulation study demonstrated the advantages of using the probabilistic approach, particularly in settings with small sample sizes. However, the simulation was based on a simple model with only one explanatory variable, and the threshold values and covariate effects were held constant across all datasets. A more comprehensive comparison could be pursued in future work by incorporating more complex models, varying parameter values, and exploring different data-generating scenarios.

Despite the demonstrated strengths of the probabilistic approach, several limitations should be acknowledged. The comparison with classical methods was based on a limited set of ordinal variables and a single dataset, which may constrain the generalizability of the findings. The probabilistic models employed a single weakly informative prior specification chosen for simplicity; however, a preliminary sensitivity analysis was conducted by varying the prior variances while

keeping the expected values fixed. Although this analysis indicated that the results were robust to these changes, a more comprehensive investigation of prior influence would be a good direction for future research.

Although all ordinal response categories in our dataset contained observations, small-sample scenarios may include empty categories, as seen in the Live Music Census data [8] also from the OpenMusE project. Such cases cause then classical methods to fail to estimate the threshold parameters, whereas probabilistic approaches can provide more stable inference. Future work could extend our analysis by exploring different link functions (e.g., probit or log-log) [1] and other ordinal regression frameworks, such as adjacent-category or sequential models [4]. These extensions would help benchmark the strengths and limitations of probabilistic modelling across diverse settings, supporting its broader adoption for surveys with small samples, missing responses, or data aggregated from multiple sources.

## Acknowledgements

This work received funding from the European Union funded under Grant No. 101095295 (OpenMusE) and Strategic Council of Finland Grant No. 352604 (Out of Despair).

## References

[1] Agresti, Alan. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Incorporated, 2010. ISBN: 9780470594001. DOI: https://doi.org/10.1002/9780470594001.

[2] Behr, Adam, Webster, Emma, Brennan, Matt, Cloonan, Martin, and Ansell, Jake. "Making Live Music Count: The UK Live Music Census". In: *Popular Music and Society 43*, no. 5 (2020), pp. 501–522. DOI: https://doi.org/10.1080/03007766.2019.1627658.

[3] Bürkner, Paul-Christian. "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software 80* (2017), pp. 1–28. DOI: 10.18637/jss.v080.i01.

[4] Bürkner, Paul-Christian and Vuorre, Matti. "Ordinal Regression Models in Psychology: A Tutorial". In: *Advances in Methods and Practices in Psychological Science 2*, no. 1 (2019). DOI: https://doi.org/10.1177/2515245918823199.

[5] Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew D., Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus, Guo, Jiqiang, Li, Peter, and Riddell, Allen. "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software 76*, no. 1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01.

[6] Carroll, Nicholas, Sadowski, Adam, Laila, Amar, Hruska, Valerie, Nixon, Madeline, Ma, David W.L., Haines, Jess, and Guelph Family Health Study, on behalf of the. "The Impact of COVID-19 on Health Behavior, Stress, Financial and Food Security among Middle to High Income Canadian Families with Young Children". In: *Nutrients 12*, no. 8 (2020). DOI: https://doi.org/10.3390/nu12082352.

[7] Christensen, Rune Haubo B. "Cumulative Link Models for Ordinal Regression with the R Package Ordinal". 2018. URL: https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf.

[8] Cloonan, Martin and Tuukkanen, Roosa. "Helsinki Live Music Census 2024". Tech. rep. University of Turku, 2025. URL: https://www.utu.fi/sites/default/files/public%3A/media/file/Helsinki-Live-Music-Census-2024-Report-EN.pdf.

[9] Colvin, R. M. and Jotzo, Frank. "Australian voters' attitudes to climate action and their social-political determinants". In: *PLOS ONE 16*, no. 3 (2021), pp. 1–18. DOI: 10.1371/journal.pone.0248268.

[10] Edwards, James, Borgstedt, Silke, and Barth, Bertram. “New Music Recommendation Algorithm Facilitates Audio Branding”. In: *Marketing Review St Gallen 4* (2019), pp. 888–894.

[11] Fenta, Haile Mekonnen, Workie, Demeke Lakew, Zike, Dereje Tesfaye, Taye, Belaynew Wassie, and Swain, Prafulla Kumar. “Determinants of stunting among under-five years children in Ethiopia from the 2016 Ethiopia demographic and Health Survey: Application of ordinal logistic regression model using complex sampling designs”. In: *Clinical Epidemiology and Global Health 8*, no. 2 (2020), pp. 404–413. DOI: https://doi.org/10.1016/j.cegh.2019.09.011.

[12] Fodeman, Ari D., Snook, Daniel W., and Horgan, John G. “Picking Up and Defending the Faith: Activism and Radicalism Among Muslim Converts in the United States”. In: *Political Psychology 41*, no. 4 (2020), pp. 679–698. DOI: https://doi.org/10.1111/pops.12645.

[13] Gambarota, Filippo and Altoè, Gianmarco. “Ordinal regression models made easy: A tutorial on parameter interpretation, data simulation and power analysis”. In: *International Journal of Psychology 59*, no. 6 (2024), pp. 1263–1292. DOI: https://doi.org/10.1002/ijop.13243.

[14] Gassman-Pines, Anna, Ananat, Elizabeth Oltmans, and Fitz-Henley, John. “COVID-19 and Parent-Child Psychological Well-Being”. In: *Pediatrics 146*, no. 4 (2020). DOI: https://doi.org/10.1542/peds.2020-007294.

[15] Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, and Rubin, Donald B. *Bayesian Data Analysis: Third edition*. Chapman and Hall/CRC, 2013. ISBN: 9780429113079. DOI: https://doi.org/10.1201/b16018.

[16] Hrusa, Gili, Spigt, Mark, Dejene, Tariku, and Shiferaw, Solomon. “Quality of Family Planning Counseling in Ethiopia: Trends and determinants of information received by female modern contraceptive users, evidence from national survey data, (2014- 2018)”. In: *PLOS ONE 15*, no. 2 (2020), pp. 1–18. DOI: 10.1371/journal.pone.0228714.

[17] Hussen, Nuru Mohammed and Workie, Demeke Lakew. “Multilevel Analysis of Women’s Education in Ethiopia”. In: *BMC Women’s Health 23* (2023). DOI: 10.1186/s12905-023-02380-6.

[18] Kruschke, John K. “Rejecting or Accepting Parameter Values in Bayesian Estimation”. In: *Advances in Methods and Practices in Psychological Science 1*, no. 2 (2018), pp. 270–280. DOI: https://doi.org/10.1177/2515245918771304.

[19] Lahti, Leo, Mäkelä, Eetu, and Tolonen, Mikko. “Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming”. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 280–289. URL: https://ceur-ws.org/Vol-2723/short46.pdf.

[20] Lee, Seung Eun, Vincent, Catherine, Dahinten, V. Susan, Scott, Linda D., Park, Chang Gi, and Dunn Lopez, Karen. “Effects of Individual Nurse and Hospital Characteristics on Patient Adverse Events and Quality of Care: A Multilevel Analysis”. In: *Journal of Nursing Scholarship 50*, no. 4 (2018), pp. 432–440. DOI: https://doi.org/10.1111/jnu.12396.

[21] Liddell, Torrin M. and Kruschke, John K. “Analyzing ordinal data with metric models: What could possibly go wrong?” In: *Journal of Experimental Social Psychology 79* (2018), pp. 328–348. DOI: https://doi.org/10.1016/j.jesp.2018.08.009.

[22] Liu, Anqi, He, Hua, Tu, Xin M, and Tang, Wan. “On Testing Proportional Odds Assumptions for Proportional Odds Models”. In: *General Psychiatry 36*, no. 3 (2023). DOI: https://doi.org/10.1136/gpsych-2023-101048.

[23] McElreath, Richard. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2nd ed.)* Chapman and Hall/CRC, 2020. ISBN: 9780429029608. DOI: https://doi.org/10.1201/9780429029608.

[24] OpenMusE. “Project Info”. Accessed: 19 June 2025. URL: https://www.openmuse.eu/.

[25] R Core Team. “R: A Language and Environment for Statistical Computing”. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[26] Tesema, Getayeneh Antehunegn, Worku, Misganaw Gebrie, Tessema, Zemenu Tadesse, Teshale, Achamyeleh Birhanu, Alem, Adugnaw Zeleke, Yeshaw, Yigizie, Alamneh, Tesfa Sewunet, and Liyew, Alemneh Mekuriaw. “Prevalence and Determinants of Severity Levels of Anemia among Children Aged 6–59 Months in Sub-Saharan Africa: A Multilevel Ordinal Logistic Regression Analysis”. In: *PLOS ONE 16*, no. 4 (2021). DOI: https://doi.org/10.1371/journal.pone.0249978.

[27] Tiihonen, Iiro, Lahti, Leo, and Tolonen, Mikko. “Print Culture and Economic Constraints: A Quantitative Analysis of Book Prices in Eighteenth-Century Britain”. In: *Explorations in Economic History 94* (2024). DOI: 10.1016/j.eeh.2024.101614.

[28] Tiihonen, Iiro, Tolonen, Mikko, and Lahti, Leo. “Probabilistic Analysis of Early Modern British Book Prices”. In: *Proceedings of the Conference on Computational Humanities Research 2021*. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 39–48. URL: http://ceur-ws.org/Vol-2989/short_paper9.pdf.

[29] Vehtari, Aki, Gelman, Andrew, and Gabry, Jonah. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing 27*, no. 5 (2017), pp. 1413–1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

[30] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Fourth. New York: Springer, 2002. ISBN: ISBN 0387954570. DOI: https://doi.org/10.1007/978-0-387-21706-2.