Bayesian Modeling of Ordinal Survey Data: A Live Music Census Case Study

Aleksi Lahtinen

MSc thesis
May 2025

DEPARTMENT OF MATHEMATICS AND STATISTICS

Reviewers:
Prof. Leo Lahti
Prof. Kari Auranen

UNIVERSITY OF TURKU, Department of Mathematics and Statistics

MSc Thesis
Subject: Statistics
Author: Aleksi Lahtinen
Title: Bayesian Modeling of Ordinal Survey Data: A Live Music Census Case Study
Supervisor: Prof. Leo Lahti
Pages: 44 pages + 5 appendix pages
Month and year: May 2025

Surveys are a widely used method for collecting information across many fields, and they often include ordinal variables. Ordinal variables are a type of categorical data in which the categories have a natural order but the distances between them are not assumed to be equal. To appropriately analyze such data, statistical models specifically designed to account for these characteristics are often employed.

The cumulative logistic regression model is one of the most commonly used methods for analyzing ordinal data. It accounts for the specific structure of ordinal variables by modeling a latent continuous variable, which is partitioned into observed categories through threshold parameters. In the cumulative logistic model, the logit link function is used to map the latent variable to category probabilities.

Bayesian modeling provides an alternative to the classical frequentist framework by treating model parameters as random variables. The inference is based on the parameters posterior distributions, which combine prior beliefs with observed data. This enables more flexible and informative modeling, particularly in small-sample or hierarchical settings. Bayesian models are typically estimated using sampling methods such as Hamiltonian Monte Carlo.

This thesis compares Bayesian and frequentist approaches through a simulation study, including scenarios with small sample sizes, categories with low response probabilities, and hierarchical data structures. The Bayesian models, in general, showed better performance in estimating category probabilities and model parameters.

The methods were also applied to a live music audience census dataset collected across several European cities. The ordinal variables in this dataset exhibited challenges such as sparse data and categories with no responses. The data were split into training and test sets to evaluate out-of-sample performance. The Bayesian model outperformed the frequentist model in estimating category probabilities for the test set.

Keywords: ordinal data, survey data, cumulative regression model, Bayesian statistics, hierarchical modeling

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

Pro gradu -tutkielma
Pääaine: Tilastotiede
Tekijä: Aleksi Lahtinen
Otsikko: Järjestysasteikollisen kyselyaineiston bayesiläinen mallintaminen: Tutkimus elävän musiikin kartoituksesta
Ohjaaja: Prof. Leo Lahti
Sivumäärä: 44 sivua + liitteet 5 sivua
Aika: Toukokuu 2025

Kyselyt ovat laajasti käytetty tiedonkeruumenetelmä monilla eri aloilla, ja ne sisältävät usein järjestysasteikollisia muuttujia. Järjestysasteikolliset muuttujat ovat luokitteluasteikollisia muuttujia, joiden kategorioilla on luonnollinen järjestys, mutta niiden välisten etäisyyksien ei oleteta olevan yhtä suuria.

Yksi yleisimmistä järjestysasteikollisten muuttujien analyysimenetelmistä on kumulatiivinen logistinen regressiomalli. Malli huomioi muuttujien rakenteen mallintamalla jatkuvaa piilomuuttujaa, joka jaetaan havaittuihin kategorioihin kynnysarvojen avulla. Kumulatiivisessa logistisessa mallissa käytetään logit-linkkifunktiota, jonka avulla piilomuuttujan arvot muutetaan kategorioiden todennäköisyyksiksi.

Bayesiläinen mallinnus tarjoaa vaihtoehdon perinteiselle frekventistiselle lähestymistavalle käsittelemällä mallin parametreja satunnaismuuttujina. Päättely perustuu parametrien posteriorijakaumiin, jotka yhdistävät ennakkotiedon ja havainnot. Tämä mahdollistaa joustavamman ja informatiivisemman mallinnuksen erityisesti pienillä otoskoilla ja hierarkkisissa tapauksissa. Bayesiläisten mallien estimoimiseen käytetään tyypillisesti otantamenetelmiä, kuten Hamiltonian Monte Carlo -algoritmia.

Tässä tutkielmassa bayesiläistä ja frekventista menetelmiä vertaillaan simulaatiotutkimuksen avulla. Simulaatiot kattavat tilanteita, joissa otoskoko on pieni, joidenkin kategorioiden vastaustodennäköisyys on alhainen tai aineisto on hierarkkisesti jäsennelty. Bayesiläiset mallit suoriutuivat johdonmukaisesti paremmin kategorioiden todennäköisyyksien ja mallien parametrien estimoinnissa.

Menetelmiä sovellettiin myös elävän musiikin kyselyaineistoon, joka kerättiin useissa eurooppalaisissa kaupungeissa. Aineiston järjestysasteikolliset muuttujat sisälsivät haasteita, kuten pieniä otoskokoja ja tyhjiä vastauskategorioita. Aineisto jaettiin opetus- ja testiaineistoon mallien suorituskyvyn arvioimiseksi. Bayesiläinen malli antoi testiaineistossa tarkempia ennusteita kategorioiden todennäköisyyksistä kuin frekventistinen malli.

Avainsanat: järjestysasteikollinen aineisto, kyselyaineisto, kumulatiivinen regressiomalli, bayesiläinen tilastotiede, hierarkkinen mallintaminen

Acknowledgements

I would like to express my sincere gratitude to Professors Leo Lahti and Kari Auranen for their invaluable guidance and support throughout the process of writing this thesis. I also thank Juho Pelto for his thoughtful feedback. I am grateful to the entire research group for fostering a pleasant, collaborative, and intellectually stimulating environment that not only encouraged me to develop new skills but also deepened my interest in research. Special thanks go to Martin Cloonan for generously allowing me to use the live music census data. Finally, I wish to thank my friends and family for their unwavering support throughout this process and during the entirety of my studies.

Contents

1 Introduction

Surveys are conducted in many fields in both the public and private sectors, usually with the purpose of obtaining information about some specific topic. The data are collected from a target population and the survey tries to provide a close estimate of the population's characteristics by surveying only a fraction of its members [1]. A classic example of a survey is opinion polling, which seeks to measure people's political views at a given time. Survey data often contain several sources of error, such as sampling error, that should be accounted for during modeling. Additionally, surveys can often suffer from small sample sizes, which can affect modeling accuracy.

One of the common data types in surveys is categorical variables with ordinal scales. An ordinal variable differs from a nominal categorical variable in that its categories have a natural order and the distance between them cannot be assumed to be equal [2]. Examples of ordinal variables include education level (primary school, middle school, high school, etc.), opinion on policy proposal (disagree, neutral or agree) or patients' pain level (no pain, mild pain, moderate pain or severe pain). While this thesis focuses on analyzing ordinal data, it will also pay some special attention to the survey framework where they often appear in.

There are many different ways of analyzing ordinal data, but clear benefits are gained by doing so with methods that take into account the features of ordinal variables [2]. Treating ordinal data as metric and analyzing them as such has been shown to cause several problems with inference like inaccurate effect sizes and false positive results [3]. This thesis will look at ordinal regression modeling with particular focus on the cumulative logistic model. The cumulative model is one of the common models used to analyze ordinal data. It assumes that the observed categorical values arise from a continuous latent variable, which is modeled in place of the ordinal categories [2, 4]. The thesis will introduce the theoretical background of the model and also discusses some of its extensions.

The two major approaches to model fitting in statistics are the frequentist and the Bayesian frameworks. The frequentist approach defines probability through the frequencies of events in large samples, basing uncertainty on imaginary resampling of data. This means that parameters and models cannot have probability distributions, but the measurements can [5]. In contrast, the Bayesian approach makes statements about the parameters, the models and predictions in terms of probabilities. Statistical inference is based on the posterior distribution, that combines the observed data with prior information [6]. Another difference is that the Bayesian methods are often computationally more expensive and they have to be fit using sampling methods like Markov chain Monte Carlo.

This thesis will look especially at the Bayesian approach for conducting ordinal regression modeling. The main focus of this thesis is trying to show clear benefits to analyzing ordinal and survey data in the Bayesian framework. Recent literature includes recommendations and tutorials on conducting ordinal regression modeling in the Bayesian framework [7, 3], but there has been a lack of papers demonstrating the benefits of the Bayesian approach. This thesis aims to showcase scenarios that illustrate these benefits. The benefits will be shown using simulated datasets and real data.

The data that will be used in the thesis to demonstrate the benefits of Bayesian modeling are live music census data gathered for the OpenMusE (Open Music Europe) EU project [8]. The project aims to improve the European music industry by making it more competitive, fair and sustainable. The census data were gathered in five European cities and offer unique information on the live music scene of these cities. The data include several ordinal variables that can be used in demonstrating the methods with real data.

The structure of the thesis is as follows. Section 2 provides a brief overview of the key characteristics of survey and ordinal data, which motivate the modeling choices introduced later in the thesis. Section 3 presents the cumulative logistic model in detail, along with a discussion of other methods for analyzing ordinal data. The section also includes a review of recent literature on the topic. Section 4 introduces the Bayesian framework and modeling approach, including prior specification, hierarchical extensions, sampling methods and model diagnostics. Section 5 demonstrates the benefits of using the Bayesian approach for fitting the cumulative model through a simulation study. In Section 6, the Bayesian cumulative logistic model is compared to the frequentist model using the live music census data, with interpretation of the results. Finally, Section 7 concludes the thesis by summarizing the findings, discussing potential directions for future research, and addressing the limitations and challenges encountered during the thesis.

The ChatGPT AI has been used to check the language of the thesis.

2 Characteristics of survey and ordinal data

Before introducing the statistical methods used in this thesis, this section provides an overview of key aspects of survey data. It also discusses common problems encountered in survey data and how Bayesian methods can help address them. Following this, the section examines the key characteristics of ordinal data and highlights important considerations when modeling this type of data.

As the main focus of this thesis is in Bayesian modeling of ordinal data, the specific details of dealing with different sources of survey errors will be omitted. However, this section does discuss how the Bayesian framework can help account for common issues in survey data, further motivating the use of a Bayesian approach for modeling ordinal survey data.

2.1 Survey data

Surveys are a way of obtaining information used widely across academia, private sector and government. The purpose of a survey is often to understand opinions, beliefs or behavior within a specific population. Survey sizes can vary significantly depending on the target population, which can be, for example, the entire population in a country, people attending an event or some specific subgroup of people. Surveys can be conducted in many different ways like in-person, by internet or phone, or by mail. Often, many of these methods are used at the same time to reach as many people as possible. [1]

It is often unfeasible to have the whole population take a survey, which is why a subset of the population is selected instead. This is often done by using probability sampling, which ensures that every member of the sample frame, the list of units in the population that the sample will be drawn from, has a non-zero chance of being included in the sample. Probability sampling estimates the distribution of characteristics within the population, which allows generalizing the results to the whole population while saving resources. There are many different ways of doing probability sampling, the most common being simple random sampling, clustering, stratification, and proportionate or disproportionate sampling. The choice of sampling methods depends on the type of the survey, available resources and the sample frame. [1]

Using probability sampling can introduce error to the survey results. Survey error is defined as the difference between an estimate produced using the survey data and the true value of the variables in the population. There are four main types of error that should be accounted for when analyzing survey data: coverage error, sampling error, nonresponse error and measurement error (Table 1). Not accounting for these in the analysis can lead to biased estimates. The sample size is also an important factor affecting the precision of the estimates. The higher the sample size, the more precise the estimates are in general. [1]

There are many different statistical methods used to account for the sources of error, but Bayesian methods are well suited to address many of them. The error types that are usually addressed in statistical analysis are coverage, sampling and nonresponse errors. One method used to adjust for coverage and sampling error is

Error type	Description
Coverage error	Sample frame does not accurately represent the population on the characteristics of interest.
Sampling error	Error in the estimate caused by surveying only a sample of units rather than the entire sample frame. This type of error is inherent in all surveys.
Nonresponse error	This error occurs when only some of the sampled units respond, and the nonrespondents differ from the respondents in a way that affects the estimate.
Measurement error	Caused by the respondents giving inaccurate answers to the survey questions.

Table 1: Short descriptions of the different error types present in survey sampling.

multilevel regression with poststratification (MRP). It accounts for differences in key characteristics, such as age, gender, and education level, between the sample and the target population to adjust the estimates produced by the multilevel model [9]. This method works quite naturally in the Bayesian framework. While MRP is a prominent example, other common survey analysis methods, like weighting, can also be implemented in a fully Bayesian manner [10]. The Bayesian approach also allows for modeling the sampling process as a part of the model.

Nonresponse error can be caused by unit nonresponse, no answer at all, or item nonresponse, no answer to some of the questions. In addition to this, there are different types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The type of missingness determines how it should be accounted for in the modeling. Typically, MCAR and MAR are considered manageable in analysis, while MNAR can lead to problems with inference. Modeling the missingness can be integrated as a part of a Bayesian model by using, for example, selection models or pattern mixture models, which allow including the missingness process as a part of the model [11]. One of the common ways of dealing with missing observations is with imputation, which can also be implemented in a fully Bayesian way [12]. As previously noted, this thesis will not discuss in further detail the specific Bayesian methods used to adjust for sampling, coverage, and nonresponse errors, such as multilevel regression with poststratification, weighting, and imputation techniques.

2.2 Ordinal data

Ordinal data are common in surveys, often appearing in the form of Likert scale questions (Table 2). They are a type of categorical data in which the categories have a natural order or ranking, unlike nominal data, where the order is arbitrary. Additionally, the intervals between categories cannot be assumed to be equal [2]. For example, respondents may find it easier to choose "Disagree" or "Agree" rather than "Strongly agree" or "Strongly disagree." These features should be taken into account when modeling ordinal data to obtain more precise and meaningful results.

Question/Statement	Answer options				
The tax rate should be lower.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
How satisfied are you with your life?	Very dissatisfied	Dissatisfied	Neutral	Satisfied	Very satisfied
How is the public transportation in your area?	Very bad	Bad	Average	Good	Very good

Table 2: Examples of five-level Likert scale variables, including the questions and answer options.

Despite this, there is a common temptation to treat ordinal data as if they were metric. Applying metric models, which assume equal intervals between categories, can lead to several problems [3]. This approach may distort means, obscure true effects, and inflate false positive rates (Type I errors). It can also result in inaccurate estimates of response probabilities, which are often central to ordinal data analysis.

Metric models are fundamentally misspecified for ordinal responses because they ignore the ordered nature of the data and falsely assume consistent spacing between categories [3]. These issues increase the risk of biased or uninterpretable parameter estimates. To avoid such pitfalls, ordinal data should be analyzed using methods specifically designed for their structure.

The following section introduces the statistical background for the cumulative logistic model, which is one of the most commonly used regression models when analyzing ordinal data. The model takes into account the special properties that have been discussed in this section.

3 Ordinal regression modeling

This section focuses on ordinal regression modeling, with particular emphasis on the cumulative regression model. It begins with a review of recent literature related to the ordinal cumulative logistic regression model. This is followed by an introduction to the model's general framework including details on parameter estimation and model checking. The section will also discuss extensions to the basic version of the cumulative model, such as the location-scale variant. The section will conclude with a brief overview of alternative regression models for ordinal data.

3.1 Literature review of the cumulative logistic regression model

A review of the literature on ordinal regression models shows that cumulative models are widely favored for analyzing ordinal data, with the logistic version being the most prevalent in practice. While various link functions can be used (as discussed later), the logit link is by far the most commonly employed. This widespread usage provides a clear motivation for focusing on the cumulative logistic model in this thesis, rather than on alternative models.

The literature on cumulative logistic regression models can be broadly divided into methodological studies that focus on the method itself and applied studies that use the model to analyze data. In recent years, the model has seen use in wide range of different fields like public health and epidemiology [13, 14, 15], clinical research [16], mental health research [17], social sciences [18, 19, 20], credit ratings [21], technology [22], and transportation research [23]. The logistic model has been particularly popular in epidemiological research, where it is often used to examine the effect of explanatory variables on an ordinal outcome, such as the anemia severity of children [13] or effect of hydrocortisone in treating COVID-19 [14]. Another popular field of application for the model is social sciences. The model has been used to examine, for example, socioeconomic status [19] and academic performance [20]. Beyond inference, the logistic model is also used in predictive modeling, for example forecasting university rankings [18] or predicting mental health help-seeking orientations [17].

In analyses of non-hierarchical data, most studies using the logistic model adopt a frequentist approach rather than a Bayesian one. The reasons for this preference are not always explicitly stated, but possible factors could be lack of knowledge about the Bayesian approach or the established practice of using the frequentist version. Two studies [14, 23] that employ the Bayesian approach justify their choice by highlighting several advantages. One key benefit is the ability to incorporate prior information and to quantify uncertainty through posterior probabilities. Another is the capacity to make direct probability statements about parameters, avoiding the indirect interpretation of frequentist confidence intervals and the emphasis on null hypothesis testing. Additionally, they note improved inference in small-sample settings.

The hierarchical version of the cumulative logistic model is widely used in the literature, as many datasets have a nested or multilevel structure that either enables

or requires such an approach. In these scenarios both frequentist [18, 16] and Bayesian methods [13, 15, 22] are employed with similar frequency. Researches choosing the Bayesian approach cite similar advantages as the ones mentioned previously, but also additional ones like more informative and robust estimates and greater modeling flexibility. This suggests that the benefits of Bayesian modeling are more commonly recognized in hierarchical data analysis than in studies involving non-hierarchical data.

In recent years, several articles have discussed the methodological aspects of the cumulative logistic model. Many of these focus on introducing ordinal regression modeling, typically featuring the cumulative logistic model as a central example [7, 24, 25, 4]. These articles typically explain why ordinal data require dedicated modeling approaches, present mathematical formulations of various ordinal models, provide examples using R, and sometimes explore model extensions. Of the four articles cited, only the one by Bürkner and Vuorre [7] advocates for the use of Bayesian methods, while the others rely on frequentist methods. This trend suggests that even in recent methodological work, the frequentist framework remains the preferred choice.

One of the most influential articles in the recent literature on ordinal regression modeling is by Liddell and Kruschke [3], which emphasizes the importance of using ordinal models when analyzing ordinal data. They surveyed psychology journal articles and found that, despite working with ordinal outcomes, all the studies relied on metric models, a practice they demonstrated to be problematic. Liddell and Kruschke's paper has likely contributed to the recent increase in articles introducing ordinal regression models, as it is frequently cited in these articles. Liddell and Kruschke advocate for the Bayesian approach due to its flexibility, ability to provide informative posterior distributions, and robustness even with small sample sizes. However, despite this endorsement, only the article by Bürkner and Vuorre showcases the Bayesian approach.

Not all recent methodological papers focus on the foundational aspects of ordinal regression. Instead, many explore more specialized topics related to the cumulative model. One such topic is how to test the proportional odds assumption, a key requirement of the model [26, 27]. Other papers focus on challenges related to parameter interpretation, particularly within the context of logistic models [28]. These works aim to clarify how model coefficients can be meaningfully understood in applied research. In addition, some papers have introduced extensions to the basic logistic model. For example, the location-shift model allows for better modeling of variance in the data [29]. There has also been papers that have looked at ordinal modeling of continuous variables [30]. Overall, the framework of the cumulative logistic model has remained largely stable, with no major changes in the recent methodological literature.

A gap remains in the literature concerning the use of Bayesian cumulative logistic regression models, particularly in non-hierarchical contexts. In applied research, the frequentist version of the model is generally preferred, while the Bayesian alternative is employed far less frequently. This stands in contrast to hierarchical modeling, where both Bayesian and frequentist approaches are widely used. Most methodological papers introduce and advocate the use of ordinal regression models but focus primarily

on the frequentist framework. Notable exceptions include the papers by Bürkner and Vuorre [7], and Liddell and Kruschke [3], who promote the Bayesian approach for its conceptual and practical advantages. However, not even these works present direct comparisons between the two paradigms. This thesis aims to address that gap by demonstrating concrete benefits of using the Bayesian approach when fitting the ordinal regression models.

3.2 Cumulative regression model

Throughout this thesis Y denotes a random ordinal response variable and observed values are denoted with y. Bolded version of these (e.g. $\mathbf{Y}$) denotes $(n \times 1)$ vectors, n being the number of observations. Explanatory variables (i.e., predictors) are represented by $\boldsymbol{x}$, which is a $(p \times 1)$ vector and covariate effects are represented by $\boldsymbol{\beta}$, which is also a $(p \times 1)$ vector. In the following the subscript i for individual observations will be ignored for clearer notations.

In standard linear regression, the variable of interest Y is modeled directly using the predictors. However, this approach is inappropriate when Y is an ordinal variable. This is why some other method has to be used, one of them being the cumulative regression model, originally introduced for ordinal data by McCullagh in 1980 [31].

The cumulative model assumes that there is a latent variable $\tilde{Y}$ that is modeled instead the ordinal variable Y. The latent variable is partitioned into the $J = K + 1$ observable categories of Y by K thresholds α_k, $k = 1, \ldots, K$, where $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = \infty$. For a typical five-level Likert scale ordinal variable there are $K = 4$ thresholds and $J = K + 1 = 5$ categories. This means that if the observed outcome is $Y = j$, where $j = 1, \ldots, J$, the corresponding latent variable $\tilde{Y}$ has value in range $(\alpha_{j-1}, \alpha_j]$. Figure 1 visualizes the latent distribution behind the cumulative model. This latent variable formation is the reason why the cumulative model quite naturally aligns with how many ordinal variables, such as those on the Likert scale, are typically interpreted. [2, 7]

As mentioned, the latent variable $\tilde{Y}$ is modeled instead of the actual observation Y. The equation has the form of standard linear regression model:

$$\tilde{Y} = \boldsymbol{\beta}'\boldsymbol{x} + \varepsilon,$$

where $\boldsymbol{\beta}'\boldsymbol{x}$ is the linear predictor and ε is a random error term, for which $E(\varepsilon) = 0$. The distribution of the random error will be discussed in more detail in the next section. Under the latent variable structure, the cumulative probability that the ordinal variable Y falls in category j or lower is

$$\begin{aligned} P(Y \leq j) &= P(\tilde{Y} \leq \alpha_j \mid \boldsymbol{x}) = P(\boldsymbol{\beta}'\boldsymbol{x} + \varepsilon \leq \alpha_j) \\ &= P(\varepsilon \leq \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}) = F(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}), \quad j = 1, \ldots, J-1, \end{aligned} \tag{1}$$

where F denotes the cumulative distribution function (CDF) of the error term ε. This probability is conditional on the predictors $\boldsymbol{x}$, as indicated by the notation $\mid \boldsymbol{x}$. When $j = J$, the cumulative probability is $P(Y \leq J) = 1$. The name of the model arises from the fact that it models the cumulative probabilities, as shown in Equation 1. However, the interest often lies in the probability of a specific category

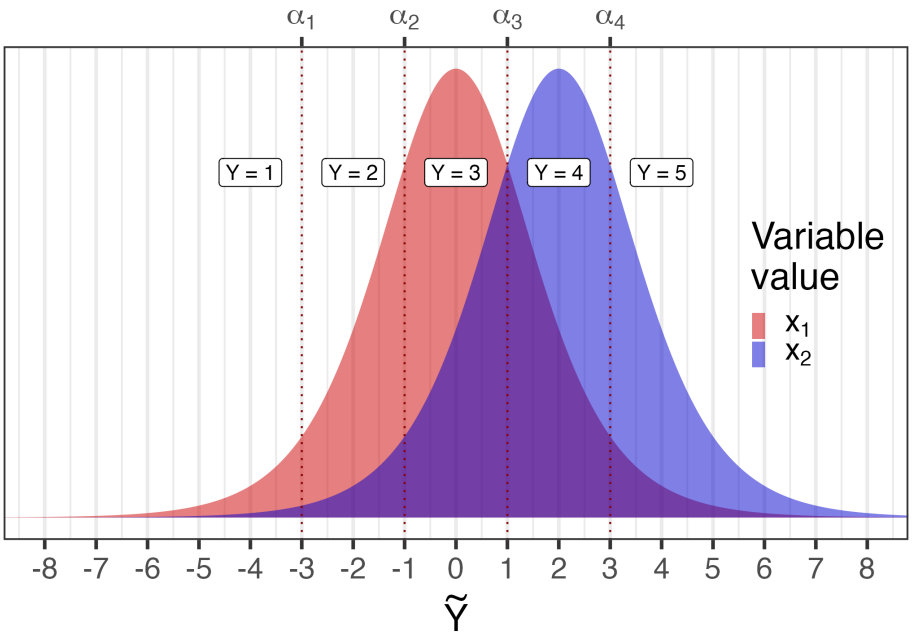

Figure 1: Demonstration of the distribution of the latent variable. The dashed lines represent the four thresholds that partition the distribution into five observable categories. The figure also illustrates how the explanatory variable affects the latent distribution: when the variable has value x_2, higher categories are more likely compared to the case where it has value x_1. The thresholds are the same for both variable values.

j instead of the cumulative probabilities. The category-specific probabilities are obtained from the following equation:

$$\begin{aligned} P(Y = j) &= P(Y \leq j) - P(Y \leq j - 1) \\ &= F(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}) - F(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}), \quad j = 1, \ldots, J. \end{aligned} \tag{2}$$

The formulation in the equation above assumes that the linear predictor $\boldsymbol{\beta}'\boldsymbol{x}$ is the same across all response categories. While this assumption may not always hold as predictors can in principle have different effects across thresholds, it is maintained for the purposes of this thesis. [2, 7]

To express the right side of the Equation 1 in terms of a linear predictor, a link function F^{-1} is applied. The link function is the inverse of the cumulative distribution function F. This gives the following equation:

$$F^{-1}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, \ldots, J - 1. \tag{3}$$

The form of the link function depends on the distribution of the error term ε. The distribution also determines the form of the cumulative distribution function F. Common choices are the logistic distribution, which lead to the cumulative logistic model, and the normal distribution, which leads to the cumulative probit model. The naming of the models is based on the link functions F^{-1}. [2, 7]

3.3 Cumulative logistic model

As discussed in the literature review section, the logistic version of the cumulative model is the most commonly used in applied research. This is also the version that McCullagh originally introduced [31]. He referred to it as the proportional odds model, a term that will be explained in detail shortly. The model is also commonly referred to as the cumulative logit model, but in this thesis, it is referred to as the cumulative logistic model.

Using the cumulative logistic model one assumes that the error term ε follows a standard logistic distribution for which the cumulative distribution function is

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

The inverse of this function defines the logit link function:

$$F^{-1}(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right),$$

which is the log-odds. This is the basis for the name cumulative logistic, or logit, model. Applying this link function to Equations 1 and 2 yields the cumulative and category-specific probabilities for the logistic model. The cumulative probability is given by:

$$P(Y \leq j) = \frac{\exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})}, \quad j = 1, \ldots, J-1.$$

The probability of observing a specific category j is

$$\begin{aligned} P(Y = j) &= P(Y \leq j) - P(Y \leq j-1) \\ &= \frac{\exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})} - \frac{\exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x})}, \quad j = 1, \ldots, J. \end{aligned}$$

The linear predictor defined in Equation 3 has the form:

$$\text{logit}(P(Y \leq j)) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, \ldots, J-1. \tag{4}$$

Consequently, in the cumulative logistic model, both the threshold parameters α_j and the regression coefficients $\boldsymbol{\beta}$ are interpreted on the log-odds scale. Specifically, α_j represents the log-odds of the outcome being in category j or below when all predictor variables are zero. Each coefficient β_k quantifies the change in log-odds of being in category j or lower for one-unit increase in the corresponding predictor x_k when all other predictors are held constant. [2, 7]

As mentioned earlier, the model is also referred to as the proportional odds model. The name reflects a key property of the model in that under Equation 4, the difference in log-odds for any category j between two covariate vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is constant:

$$\begin{aligned} &\text{logit}(P(Y \leq j \mid \boldsymbol{x}_1)) - \text{logit}(P(Y \leq j \mid \boldsymbol{x}_2)) \\ &= \log\left(\frac{P(Y \leq j \mid \boldsymbol{x}_1)/P(Y > j \mid \boldsymbol{x}_1)}{P(Y \leq j \mid \boldsymbol{x}_2)/P(Y > j \mid \boldsymbol{x}_2)}\right) = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2). \end{aligned} \tag{5}$$

This implies that the odds of observing a response in category j or below when $\boldsymbol{x} = \boldsymbol{x}_1$ are $\exp(\boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2))$ times the odds when $\boldsymbol{x} = \boldsymbol{x}_2$. In other words, the log-odds ratio is proportional to the distance between the covariate vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and it is the same across all categories (Figure 2). [2]

The proportional odds property can also be understood in terms of the threshold parameters. When holding the predictor values constant, the log-odds ratio of being in category j_1 or below versus j_2 or below depends only on the difference between the corresponding thresholds:

$$\begin{aligned} &\text{logit}(P(Y \leq j_1 \mid \boldsymbol{x})) - \text{logit}(P(Y \leq j_2 \mid \boldsymbol{x})) \\ &= \log\left(\frac{P(Y \leq j_1 \mid \boldsymbol{x})/P(Y > j_1 \mid \boldsymbol{x})}{P(Y \leq j_2 \mid \boldsymbol{x})/P(Y > j_2 \mid \boldsymbol{x})}\right) = \alpha_{j_1} - \alpha_{j_2}. \end{aligned}$$

This highlights the fact that the model assumes a constant effect of predictors across all cumulative splits, and that category differences are captured solely through the thresholds. [7]

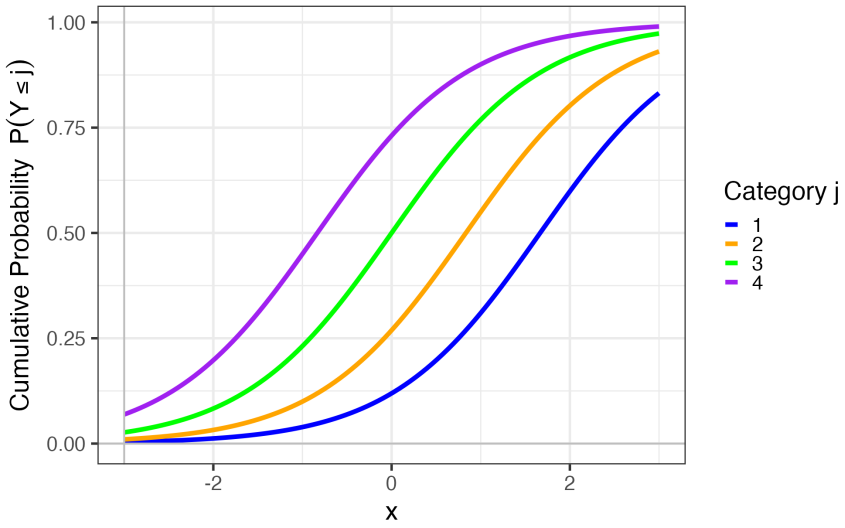

Figure 2: The cumulative probability curves in a cumulative logistic model with proportional odds. All the curves have a similar shape, indicating that the effect of the predictor is consistent across different categories

3.3.1 Likelihood function and parameter estimation

Parameter estimation is an important part of any regression model. This section will briefly outline the form of the likelihood function and maximum likelihood estimation for the cumulative logistic model defined in Equation 4 and describes how the model parameters can be estimated. In this section, the subscript i refers to individual observations.

The likelihood function and parameter estimates are obtained in a similar manner to other regression models. Let $y_{i1}, \ldots, y_{iJ}$ be binary indicator variables for observations i, where $y_{ij} = 1$, if the response falls into category j, and $y_{ik} = 0$ for all

$k \neq j$. Let $\pi_j(\boldsymbol{x}_i) = P(Y_i = j \mid \boldsymbol{X} = \boldsymbol{x}_i)$ denote the category probabilities. Assuming independence across observations, the likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \pi_j(\boldsymbol{x}_i)^{y_{ij}} \right] = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} (P(Y \leq j \mid \boldsymbol{x}_i) - P(Y \leq j-1 \mid \boldsymbol{x}_i))^{y_{ij}} \right]$$
$$= \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \left(\frac{\exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}_i)} \right)^{y_{ij}} \right], \tag{6}$$

where $\boldsymbol{\alpha}$ is a $(J - 1 \times 1)$ vector of the threshold parameters. Taking a logarithm yields the log-likelihood function:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log \left(\frac{\exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\alpha_{j-1} - \boldsymbol{\beta}'\boldsymbol{x}_i)} \right). \tag{7}$$

The maximum likelihood estimates of the model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are obtained by differentiating the log-likelihood function in Equation 7 with respect to each parameter and setting the derivatives equal to zero. These equations do not have a closed-form solution and iterative methods, such as the Fisher scoring algorithm, are typically employed. These methods, or a closer inspection of the maximum likelihood estimation process, will not be covered in this thesis. [2]

Estimates of the category-specific probabilities are obtained by inserting the maximum likelihood estimates of the model parameters into Equation 2. Standard errors for the probabilities can be derived using the delta method, and confidence intervals can be constructed as Wald intervals [24].

3.3.2 Model checking

This section briefly discusses common methods for evaluating the fit of the cumulative logistic model. As this thesis focuses on the Bayesian approach, detailed discussion and implementation of these methods is omitted, but Bayesian-specific model checking procedures will be introduced in more detail later.

Several techniques can be used to assess the adequacy of a cumulative logistic model. One approach involves goodness-of-fit tests, which compare observed counts in non-sparse contingency tables to those predicted by the fitted model. The null hypothesis that the model fits the data well can be tested using either the Pearson statistic or the likelihood-ratio statistic. However, these global tests are not valid for sparse data or models that include continuous predictors. Another approach is to compare nested models, where additional terms are introduced to test whether they improve the model fit. These comparisons can be made using the likelihood-ratio test or information criteria, such as the Akaike Information Criterion (AIC). Standardized residuals may also be examined and large residuals for specific cells suggest a local lack of fit. [2]

A key assumption of the cumulative logistic model is the proportional odds assumption (Equation 5). This assumption can be tested by comparing the proportional odds model to more flexible models that allow for category-specific effects. One commonly used test is the score test, which evaluates whether allowing such flexibility

significantly improves model fit. A small p-value indicates that the proportional odds assumption may not hold. The likelihood-ratio test and the Brant test can also be applied for this purpose. However, it may not always be feasible to use the likelihood-ratio or Brant tests, as the more complex model may suffer from structural issues, such as cumulative probabilities becoming unordered, which may prevent successful maximization of the likelihood function. For this reason, the score test is generally more applicable although it also has limitations as it can perform poorly with sparse data and tends to be overly liberal when the data are not sparse. [2]

3.4 Extensions of the cumulative model

There are many ways to extend the cumulative logistic model defined in Equation 4 to allow for greater flexibility in modeling, and this section briefly discusses several such extensions. Many of these focus on addressing additional heterogeneity or situations in which the proportional odds assumption does not hold. One such extension involves allowing category-specific effects, which leads to a nonproportional odds model. As noted in the model checking section, the inclusion of category-specific effects may result in negative or unordered probabilities. An alternative is the partial proportional odds model, which accommodates both category-specific effects and effects that adhere to the proportional odds structure [2].

Several models have also been developed to account for heterogeneity that the basic cumulative model does not capture. One example is the location-scale model, in which the variance (scale) of the latent distribution depends on covariates. This is accomplished by introducing a scaling term into the model that varies with covariate values. The location-shift model offers another approach, allowing the threshold parameters to vary based on covariates. A key advantage of the location-shift model, in comparison to the location-scale model, is that it remains within the class of generalized linear models, which allows for the use of standard estimation and diagnostic techniques. [29, 4]

Another way to extend the cumulative model is by choosing a different link function instead of the logistic one. As mentioned earlier, one of the most common alternatives is the probit link, which assumes that the error term ε follows the normal distribution. The model fits and resulting predictions of the logit and probit models are often similar, although the estimated parameters differ in scale. Another link function is the log-log link, which arises when the error term follows an extreme value distribution. The log-log model, also known as the proportional hazards model, differs from the logistic and probit models in that its distribution is not symmetric. A further extension is the Cauchit link, which is based on the Cauchy distribution. The Cauchy distribution has heavier tails than either the logistic or normal distributions, which makes it more sensitive to extreme values. [2, 7]

The cumulative model can also be extended within the hierarchical framework, similar to other regression models. This is done by adding cluster-specific random effects to the model. The random effect can be included in the form of an intercept or slope term, and multiple random effects may be added when the data are nested at several levels [2, 4]. A more detailed explanation hierarchical modeling and its benefits will be presented in the Bayesian modeling section.

3.5 Other ordinal regression models

The cumulative model is not the only regression model used to analyze ordinal data. This section briefly highlights other ordinal regression models to give a general picture of the other options. Specific details of these models are not included. Notably, all the models discussed can also be implemented within the Bayesian framework.

One alternative model is the adjacent-categories model, which is also commonly used for ordinal data. This approach models a series of conditional binary comparisons, each representing the probability that the response falls in category j given that it lies in either category j or $j+1$. The model thus treats the decision process as occurring between adjacent pairs of categories, and each decision is associated with its own latent variable distribution. An advantage of the adjacent-categories model is that it can accommodate category-specific effects without the risk of leading to unordered probabilities, as can occur in the cumulative model. While it is most commonly estimated using a logistic distribution, the model can also be specified with alternative distributions such as the normal distribution. [7, 4]

The sequential model is, alongside the cumulative and adjacent-categories models, one of the main regression approaches used for analyzing ordinal data. It is particularly suitable when the response process is inherently stepwise, such that a higher category can only be reached by passing through all lower categories. In this framework, a latent continuous variable governs the transition between adjacent categories j and $j+1$. The model uses binary submodels to express the probability of observing $Y \geq j$ given that $Y \geq j-1$, reflecting a conditional progression through ordinal levels. Like other ordinal models, it is often implemented with the logistic, normal, or extreme-value distributions. [7]

Mixture models provide another approach for analyzing ordinal data, particularly when there is heterogeneity in response behavior. These models assume that respondents come from different latent classes, for instance, one group whose responses reflect content-driven preferences, and another characterized by uncertainty or random responding. Mixture models incorporate two sets of parameters: one that governs the structured, content-based part of the response, and another that models the likelihood of uncertain or random behavior. This allows for the separation of informative and uninformative responses within a unified framework. [4]

4 Bayesian framework

This section presents the Bayesian framework as applied to the cumulative ordinal model, starting with a detailed comparison between the frequentist and Bayesian approaches. It then presents the fundamental components of Bayesian modeling, with particular emphasis on adapting the cumulative logistic model to this framework. The section continues with an introduction to Bayesian hierarchical modeling, illustrating how multilevel structures can be naturally incorporated in the Bayesian context. Hamiltonian Monte Carlo, the primary sampling method used in this thesis for drawing from posterior distributions, is also introduced. The section concludes with an overview of Bayesian model checking and model comparison techniques.

4.1 Comparing Bayesian and frequentist approaches

The Bayesian framework, which will be compared to the frequentist one in this section, provides a fundamentally different perspective on statistical modeling. It is important to identify the differences between the two approaches to better understand the advantages offered by the Bayesian approach.

In the frequentist framework, parameters are treated as fixed but unknown quantities, and inference relies on repeated sampling, typically using methods such as maximum likelihood estimation and confidence intervals. In contrast, the Bayesian framework treats parameters as random variables and bases inference on their posterior distributions, which combine prior beliefs and observed data using Bayes' theorem [6, 5]. This allows Bayesian models to make direct probability statements about parameters and future observations. For example, a Bayesian credible interval expresses a 95% probability that a parameter lies within a given range, whereas a frequentist confidence interval describes the expected behavior of the interval across repeated samples, which can be less intuitive to interpret. However, frequentist confidence intervals have guaranteed long-run frequency coverage, whereas Bayesian credible intervals may lack this property, particularly if the prior is poorly specified.

One of the most fundamental features of Bayesian modeling is the possibility to specify prior distributions for model parameters. Priors allow incorporating prior knowledge, domain expertise, or regularization into the modeling process by constraining the range of plausible parameter values [5]. This is a key distinction from the frequentist approach, which does not formally include prior information in model estimation. The ability to set priors can be viewed as either a strength or a potential drawback. On the one hand, priors make it possible to incorporate meaningful prior beliefs about the topic of research. On the other hand, poorly specified, or intentionally biased, priors can lead to misleading inferences. For this reason, it is important to choose priors transparently, to justify their use, and to assess their influence on posterior results. A more detailed discussion of priors will be provided later in the thesis.

The Bayesian framework generally allows for more complex and flexible modeling than the frequentist approach. This flexibility stems from the Bayesian model's ability to incorporate multiple levels of uncertainty and to combine information from different sources. It also provides a conceptually straightforward way to handle

models with many parameters. In particular, the Bayesian framework is well-suited for hierarchical modeling, where group-level variation is captured through hyperpriors and structured dependencies. [6, 5]

The main cost of the Bayesian models' flexibility is computational. While frequentist models can often be fit in a matter of seconds, Bayesian models can take much longer, especially when dealing with complex hierarchical structures. This computational burden arises because samples typically cannot be drawn directly from the posterior distribution. Instead, the posterior must be approximated using sampling algorithms, such as Hamiltonian Monte Carlo. Significant advances in Bayesian computation over the past decade has made it much easier and faster to fit these models than was previously possible. [6]

Despite the many benefits of Bayesian modeling, it can still feel intimidating to many practitioners. This hesitation may arise from the need to specify prior distributions, concerns about longer computation times, or a general lack of familiarity with the Bayesian framework. One of the aims of this thesis is to demonstrate the advantages of the Bayesian approach, particularly in the context of ordinal regression modeling, and to encourage its adoption as a practical and powerful alternative to traditional frequentist methods.

4.2 Bayesian modeling

In the following, the notation $\boldsymbol{\theta}$ is used to represent the model parameters. For the cumulative logistic model, this includes both the threshold parameters and the covariate effects. The notation $p(\cdot)$ denotes a marginal probability distribution, while $p(\cdot \mid \cdot)$ is used for conditional distributions. Unobserved outcomes are denoted with a $\tilde{\boldsymbol{y}}$.

Bayesian model structure and inference

Bayesian inference about parameters $\boldsymbol{\theta}$ is based on the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, which represents the probability distribution of the parameters given the observed data. To begin the modeling process, a joint probability distribution (or joint density) for $\boldsymbol{\theta}$ and $\boldsymbol{y}$ has to be defined. This can be written as the product of the prior distribution $p(\boldsymbol{\theta})$ and the sampling distribution (or likelihood) $p(\boldsymbol{y} \mid \boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Applying Bayes' theorem yields the posterior density of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{y})}.$$

In the equation above, the denominator $p(\boldsymbol{y})$ is the marginal likelihood or prior predictive distribution, which for continuous $\boldsymbol{\theta}$ can be written as

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{8}$$

Since the marginal likelihood does not depend on the parameters $\boldsymbol{\theta}$, it is often treated as a normalizing constant. For this reason, Bayesian inference frequently works with the unnormalized posterior:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{9}$$

This expression makes clear that the observed data $\boldsymbol{y}$ affects the posterior only through the likelihood function $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, which is treated as a function of $\boldsymbol{\theta}$ for fixed data. [6]

Predictions are straightforward to obtain in the Bayesian framework. Given observed data $\boldsymbol{y}$, an unknown but observable quantity $\tilde{\boldsymbol{y}}$ can be predicted using the posterior predictive distribution, which is defined as

$$\begin{aligned} p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) &= \int p(\tilde{\boldsymbol{y}}, \boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} = \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}, \end{aligned} \tag{10}$$

where the final step relies on the assumption that $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ are conditionally independent given $\boldsymbol{\theta}$. Equation 10 shows that the posterior predictive distribution is obtained by averaging the conditional predictive distribution $p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta})$ over the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. In other words, predictions are made by integrating over parameter uncertainty. Both the posterior predictive distribution and the prior predictive distribution (Equation 8) play a central role in Bayesian model assessment and checking. These concepts will be explored further in a later section. [6]

Choosing prior distributions

Much can be said about how to choose prior distributions for the parameters $\boldsymbol{\theta}$ of the Bayesian model defined in Equation 9. At a fundamental level, the prior represents what is known or believed about the parameters before observing any data. In practice, there are many approaches and motivations for choosing priors, and some degree of subjectivity is always involved. For this reason, it is important to justify prior choices explicitly to ensure transparency and reproducibility. This section discusses different types of priors, the motivations behind them, and the contexts in which each type may, or may not, be appropriate.

Noninformative priors and weakly informative priors are commonly used in Bayesian modeling. The former are typically flat or vague, providing minimal influence on the posterior and often serving to regularize inference without imposing strong assumptions. Many noninformative priors are improper, meaning they do not integrate to one, though they can still yield proper posterior distributions. However, even proper noninformative priors can sometimes place substantial mass on implausible or impossible values for a given parameter. An example of a weakly informative and proper prior is the normal distribution $N(0, 100)$, which spreads probability mass broadly across the parameter space while still offering some regularization. [6]

It is important to recognize that a prior that appears noninformative may, in fact, be highly informative depending on the scale of the parameter. For example, the prior $N(0, 100)$ could be strongly informative if the true parameter values are expected to

be in the range of thousands. Improper or overly vague priors can lead to biased posterior distributions, especially when applied to transformed parameters (e.g., in log or inverse scales). Additionally, noninformative priors often lead to results that closely resemble those from frequentist methods. While this is not problematic in itself, it means that the Bayesian framework is not being fully leveraged. Specifically, the potential to incorporate prior information to improve inference and precision remains underutilized. [32]

Weakly informative priors are more informative than noninformative priors, as the name suggests. They are commonly used to regularize the posterior, helping to constrain extreme or implausible parameter values without exerting a strong influence on the results. Importantly, they are proper distributions, unlike many noninformative priors, which makes them a safer and more stable choice in practice [6]. An example of a weakly informative prior might be the normal distribution $N(0, 1)$ (e.g., for the estimate of a standardized variable), which centers the parameter near zero while still allowing moderate variability. Using weakly informative priors is often preferable to noninformative priors, as they better leverage the Bayesian framework [32]. They provide more accurate and stable estimates, help avoid biased posteriors, and remain conservative in their assumptions.

At the other end of the spectrum are informative priors, which incorporate a substantial amount of prior knowledge into the analysis, taking full advantage of the Bayesian approach. These priors are typically based on previous research, expert opinion, or external data and are intended to reflect strong, well-founded beliefs about the likely values of the parameters. For example, prior $N(2, 0.1)$ may be appropriate if the parameter is believed to be close to 2 with high certainty. Informative priors can improve estimation accuracy when reliable prior information is available. However, they must be used cautiously, as they can introduce bias if the prior is misspecified, thereby unduly influencing the posterior distribution [32].

Historically, conjugate priors have played an important role in Bayesian modeling. A prior distribution is said to be conjugate to a given likelihood function $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ if the resulting posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ belongs to the same distribution family as the prior $p(\boldsymbol{\theta})$. The main advantage of conjugate priors is their computational convenience, which made them especially useful before the development of modern sampling algorithms. Because conjugate priors lead to analytically tractable posterior distributions, they were widely used in early Bayesian applications. While advances in computational methods have reduced their necessity, conjugate priors are still used in some contexts. In addition to their simplicity, they offer the benefit of being interpretable as additional data. [6]

Regardless of the type of prior used, the degree to which a prior is informative always depends on the scale and context of the data. A prior that is weakly informative in one context might be highly informative in another. For this reason, priors should be chosen with careful consideration of the actual data and modeling goals. Whether the prior is used to regularize the posterior or to encode prior knowledge, its choice should always be justified and documented. Additionally, it is good practice to perform prior sensitivity analyses, that is, to fit the model using multiple priors of varying strength, in order to assess how robust the posterior distribution is to the prior assumptions [32].

4.3 Cumulative logistic model in the Bayesian framework

The cumulative logistic model can be defined within the Bayesian framework by using the posterior formulation established in Equation 9, where the likelihood function is combined with the prior distributions. In the cumulative logistic model, priors are placed on the covariate effects $\boldsymbol{\beta}$ and the threshold parameters $\boldsymbol{\alpha}$. The same prior can be used for all covariate effects or thresholds, or each parameter can be assigned its own prior. The general form of the prior families is introduced in this section, while some of the hyperparameter values will be specified later.

To begin, the prior distributions for the model parameters must be defined. Assuming that the priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent, the joint prior distribution can be expressed as

$$p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = p(\boldsymbol{\beta})p(\boldsymbol{\alpha}) = \prod_{k=1}^{p} f(\beta_k) \prod_{j=1}^{J-1} f(\alpha_j),$$

where $f(\cdot)$ denotes the probability density function of the prior distribution. As the parameters are assumed to be independent, the priors can be specified and sampled separately. The priors used in this thesis for the parameters are

$$\beta_k \sim N(0, 1), \quad k = 1, \ldots, p, \tag{11}$$
$$\alpha_j \sim N(\mu_{\alpha_j}, 1), \quad \alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}, \quad j = 1, \ldots, J-1, \tag{12}$$

where the hyperparameters μ_{α_j} will be defined later. The prior $N(0, 1)$ is chosen for the coefficient parameters because it is weakly informative. It reflects the assumption that the log-odds are likely centered around zero, implying no strong prior belief about the direction or size of the effect, while still regularizing extreme estimates. The prior for the thresholds is a truncated normal distribution to enforce the required ordering of thresholds [33]. The normal distribution is chosen as it allows for a flexible specification of both weakly informative and informative priors. It is also a popular choice in the literature for modeling both covariate effects and thresholds in cumulative models [33, 27]. The likelihood function of the cumulative logistic model was defined previously in Equation 6. By combining this likelihood with the prior distributions, the unnormalized posterior distribution takes the following form:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha})p(\boldsymbol{\beta}, \boldsymbol{\alpha}) = p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\alpha}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_j(\boldsymbol{x}_i)^{y_{ij}} \prod_{k=1}^{p} f(\beta_k) \prod_{j=1}^{J-1} f(\alpha_j), \end{aligned} \tag{13}$$

where $\pi_j(\boldsymbol{x}_i) = P(Y_i = j \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$ is the probability of a specific category and $f(\cdot)$ is the probability density function of the normal distribution.

Directly sampling from the posterior distribution defined in Equation 13 is generally not feasible, as the normalizing constant is analytically intractable [6]. Therefore, sampling-based methods are required to approximate the posterior distribution. These methods will be discussed in more detail later in the thesis.

Fitting the cumulative logistic model within the Bayesian framework shares many of the same advantages and limitations as other Bayesian regression models.

However, the Bayesian approach also offers distinct benefits when applied specifically to cumulative models. One such advantage is the ability to specify a prior distribution for the threshold parameters $\boldsymbol{\alpha}$. As previously discussed, surveys often suffer from low response rates, and respondents often tend to select middle-category answers. This can result in some response categories receiving no observations at all. In the frequentist framework, such situations can lead to convergence issues or unreliable probability estimates for the unobserved categories. In contrast, the Bayesian approach allows for the use of prior distributions to regularize the thresholds, resulting in more stable and interpretable estimates of category probabilities in such scenarios.

The Bayesian framework also allows for fitting more flexible models. For example, location-shift and location-scale models often encounter estimation problems in the frequentist setting, particularly with small sample sizes, where parameters may be poorly identified or even non-estimable. In contrast, the Bayesian approach handles these models more effectively, providing stable estimates even in hierarchical or small-data contexts. Moreover, Bayesian models can easily accommodate customized threshold structures, such as allowing each level of a predictor (e.g., gender) to have its own set of threshold parameters. This added flexibility makes the Bayesian framework especially well-suited for modeling real-world ordinal data where assumptions like proportional odds or constant thresholds may not hold.

4.4 Hierarchical Bayesian modeling

The literature review revealed that many applications of the cumulative logistic model involved hierarchical data structures. This structure should be accounted for in modeling and the Bayesian framework offers a natural way to model this type of data. This section discusses the motivations and advantages of using hierarchical models. The general form of hierarchical models is presented and the cumulative logistic model is incorporated into this framework.

Hierarchical modeling, also known as multilevel modeling, is used when the data are structured across multiple levels of observational units [6]. This is commonly the case with survey data, where responses may be collected from units nested within broader clusters. For example, survey data may be gathered from individuals in different cities, regions, or countries, making it a natural candidate for hierarchical modeling. Additionally, when individuals answer multiple survey questions, their responses can be modeled hierarchically to account for within-person variation.

Hierarchical models are designed to learn from both individual groups and the overall population, sharing information across clusters based on the observed variation between them. This process, known as partial pooling, allows for more stable and accurate estimates, especially when group sizes are imbalanced. In contrast, single-level models may overfit or underfit in the presence of repeated or clustered observations. Hierarchical models also handle sampling imbalances effectively, ensuring that groups with larger sample sizes do not disproportionately influence the inference. Another key benefit is that they provide explicit estimates of between-group variation, such as differences in effects across cities. In general, when the structure of the data supports it, using hierarchical models can lead to more precise and more robust inferences. [5]

This thesis focuses on hierarchical models with varying intercepts, i.e., models in which each cluster in the data has its own intercept parameter. These intercepts are treated as random variables drawn from a common distribution, whose parameters, known as hyperparameters, are themselves given prior distributions called hyperpriors. This nested structure is what gives hierarchical (or multilevel) models their name (Figure 3). 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. 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. The benefits are particularly pronounced for smaller clusters, which may otherwise yield unstable estimates. [5]

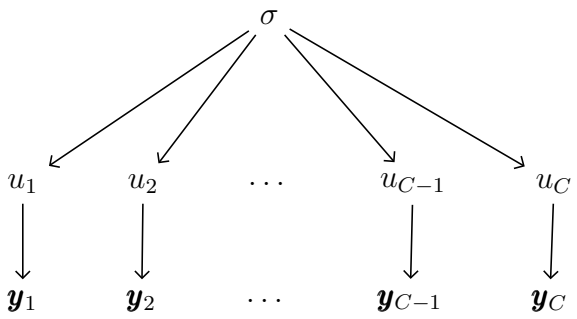

Figure 3: Illustration of the parameter structure in a two-level hierarchical model with C clusters. At the top level is the hyperparameter that defines the distribution from which cluster-specific intercepts are drawn. These intercepts are then used to specify distributions for the observations within each cluster.

Although hierarchical structures can be extended beyond two levels, this thesis focuses on two-level hierarchical models. The posterior distribution of the model parameters in such a model is

$$p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi}),$$

where $p(\boldsymbol{\phi})$ is the hyperprior for the hyperparameters $\boldsymbol{\phi}$ [6]. This structure closely resembles the general Bayesian posterior in Equation 9, but with the addition of a distribution for the hyperparameters.

The cumulative logistic model can be extended into the hierarchical framework by introducing a cluster-specific varying intercept. The linear predictor for this model is

$$\text{logit}(P(Y_c \leq j)) = \alpha_j + u_c - \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, \ldots, J, c = 1, \ldots, C,$$

where u_c is the varying intercept term for cluster c and C is the number of clusters. These intercepts are modeled as

$$u_c \sim N(0, \sigma), \quad c = 1, 2, \ldots, C, \tag{14}$$
$$\sigma \sim \text{half-normal}(0, 1), \tag{15}$$

where σ is the standard deviation of the distribution of the cluster-specific intercepts. The half-normal distribution is a normal distribution constrained to be non-negative.

While exponential priors are often used to regularize variance components, they may be too vague when only a few clusters are available, as is the case in the live music census dataset. For this reason, a more informative half-normal prior is used instead [5]. The full posterior distribution is then:

$$p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}, \boldsymbol{\sigma} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}) p(\boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\boldsymbol{u} \mid \boldsymbol{\sigma}) p(\boldsymbol{\sigma}),$$

which is similar to the posterior in Equation 13, but with the addition of the hyperprior and the distribution of varying intercepts.

Sampling from the posterior distribution in hierarchical models follows the same principles as in non-hierarchical models. However, hierarchical models introduce additional layers of parameters, such as hyperparameters and group-level effects, which increase the complexity of the posterior distribution. Once again, direct sampling is typically not feasible due to the intractability of the posterior's normalizing constant. Consequently, specialized sampling algorithms, discussed in the next section, are used to approximate the joint posterior distribution of all model parameters. [6, 5]

4.5 Sampling methods

Inferring from the Bayesian model is necessary for their practical application, and scalable solutions are essential. It is often either impossible or computationally inefficient to draw samples directly from the posterior distribution, which is why simulation-based methods are required. Hamiltonian Monte Carlo has been shown to provide a general and efficient inference mechanism for Bayesian models [6, 5]. This section provides an overview of this method and summarizes how it can be used to draw posterior samples for the types of models described in the previous section.

Markov chain Monte Carlo

There are many different types of simulation methods used for posterior sampling, one of the most prominent being Markov chain Monte Carlo (MCMC). The central idea of MCMC is to construct a Markov chain using a transition distribution that eventually yields samples approximating the target distribution, i.e., the posterior distribution of $\boldsymbol{\theta}$. An MCMC algorithm starts at an initial value $\boldsymbol{\theta}^0$ and generates a sequence of values $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \ldots$, where each new sample $\boldsymbol{\theta}^t$ is drawn from a transition distribution $T_t(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}^{t-1})$. This transition distribution depends only on the previous state, which gives the Markov chain its defining property. In practice, MCMC algorithms are typically run with multiple chains, each initialized from different starting points to improve exploration of the parameter space. These chains are run for a sufficient number of iterations until the distribution of samples is expected to have converged to the posterior. It is also common practice to discard an initial portion of the samples, called the warm-up period or burn-in period, to allow the chains to stabilize. For example, the first 50% of the samples may be discarded to ensure that only samples from the stationary distribution are retained. [6]

Two basic MCMC methods are the Metropolis-Hastings algorithm and the Gibbs sampler, with the Gibbs sampler being a special case of the Metropolis-Hastings

algorithm. These methods operate by sampling a proposal $\boldsymbol{\theta}^*$ from a proposal distribution at time t, denoted $J_t(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{t-1})$. After proposing a new value, a ratio of posterior densities is computed:

$$r = \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y}) J_t(\boldsymbol{\theta}^{t-1} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{t-1} \mid \boldsymbol{y}) J_t(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{t-1})}.$$

It is worth noting that the Metropolis-Hastings algorithm depends only on the ratio of posterior densities r, so the normalizing constant cancels out and is not required. Consequently, the intractability of the normalizing constant poses no problem. The proposed value is then accepted or rejected according to the following rule:

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^*, & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{t-1} & \text{otherwise.} \end{cases} \tag{16}$$

This update step ensures that proposals with higher posterior density than the current state are always accepted, while those with lower density are accepted with a probability proportional to r. This mechanism defines the Metropolis-Hastings acceptance rule and various extensions and refinements exist beyond this basic formulation. [6]

Hamiltonian Monte Carlo

Both Metropolis-Hastings algorithm and the Gibbs sampler are often computationally inefficient, especially with high-dimensional models. This inefficiency arises from their random walk behavior, which causes slow exploration of the posterior distribution. To address this, a more advanced algorithm known as Hamiltonian Monte Carlo (HMC) is commonly used. HMC introduces an auxiliary momentum variable ρ_j for each model parameter θ_j, enabling the sampler to make more informed transitions through the parameter space. This results in more rapid and efficient exploration of the posterior. The posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is augmented with an independent distribution $p(\boldsymbol{\rho})$ over the momentum variables, producing the following joint distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\rho} \mid \boldsymbol{y}) = p(\boldsymbol{\rho}) p(\boldsymbol{\theta} \mid \boldsymbol{y}).$$

In practice, the interest lies in the posterior samples of $\boldsymbol{\theta}$, and the momentum variables $\boldsymbol{\rho}$ serve solely to improve the efficiency of the sampling algorithm. One key requirement of HMC is the ability to compute the gradient of the log-posterior density, which is used to update the momentum variable and simulate Hamiltonian dynamics. This gradient must typically be derived analytically, and its accuracy is critical to the performance of the algorithm. [6, 34]

HMC is also a variant of the Metropolis-Hastings algorithm, and its sampling iterations follow a structure similar to those used in Gibbs and Metropolis-Hastings algorithms. Each iteration begins by sampling a new momentum vector $\boldsymbol{\rho}$ from a multivariate normal distribution $N(0, \boldsymbol{M})$, where $\boldsymbol{M}$ is typically a diagonal mass matrix. After sampling $\boldsymbol{\rho}$, the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\rho})$ is jointly updated using a series of L leapfrog steps. These steps simulate a discretized trajectory through the

parameter space and are scaled by a step size ε. Each leapfrog step consists of first updating $\boldsymbol{\rho}$:

$$\boldsymbol{\rho} \leftarrow \boldsymbol{\rho} + \frac{1}{2}\varepsilon \frac{d \log p(\boldsymbol{\theta} \mid \boldsymbol{y})}{d\boldsymbol{\theta}}.$$

This is followed by updating the value of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \varepsilon \boldsymbol{M}^{-1}\boldsymbol{\rho}.$$

The leapfrog step ends by updating the value of $\boldsymbol{\rho}$ again. After the L leapfrog steps, the ratio of densities is calculated:

$$r = \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{t-1} \mid \boldsymbol{y})p(\boldsymbol{\theta}^{t-1})},$$

where $(\boldsymbol{\theta}^{t-1}, \boldsymbol{\rho}^{t-1})$ represent the values before the leapfrog steps, and $(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$ the proposed values after the L leapfrog steps. The updated parameter value $\boldsymbol{\theta}^t$ is then accepted or rejected according to the rule in Equation 16. The momentum variables $\boldsymbol{\rho}$ do not need to be retained after the iteration, as new values are drawn at the beginning of each sampling step. [6, 34]

In this thesis, all Bayesian models were fit using Stan (*Software for Bayesian Data Analysis*) [35], which implements an extension of HMC known as the No-U-Turn Sampler (NUTS). One of the limitations of the standard HMC is that it requires the user to manually specify values for $\boldsymbol{M}$, ε, and L. NUTS addresses this by automatically tuning these parameters during sampling, removing the need for manual calibration [36]. Due to the complexity of the NUTS algorithm, further technical details are omitted.

Convergence diagnostics

When using HMC, it is essential to verify that the sampling chains have converged to the target posterior distribution. This step is critical to ensure that any inferences drawn from the samples are reliable and trustworthy. Convergence can be assessed through visual diagnostics (e.g., trace plots) or quantitative metrics, including the $\widehat{R}$ statistic and effective sample size. In models with many parameters, visual inspection may become impractical, making convergence metrics especially valuable.

The $\widehat{R}$ diagnostic is computed separately for each model parameter. It compares the variance across all chains to the average variance within each chain. If the chains have not mixed well, the between-chain variance will be higher than the within-chain variance, resulting in a higher $\widehat{R}$ value. Additionally, the diagnostic compares the first and second halves of the chains to help detect non-convergence within chains. A value of $\widehat{R} < 1.01$ is generally considered desirable for all model parameters. [37]

MCMC draws tend to be autocorrelated within chains, which reduces the effective amount of information and increases uncertainty in parameter estimates. The effective sample size (n_{eff} or ESS) quantifies how much autocorrelation impacts the efficiency of sampling. It estimates the equivalent number of independent draws obtained from the posterior. Like $\widehat{R}$, n_{eff} is calculated separately for each parameter and is typically computed by dividing the total number of draws by a measure of

chain autocorrelation. The closer n_{eff} is to the actual number of iterations, the less autocorrelated the samples are. As a general rule, n_{eff} should be at least 50 for $\widehat{R}$ to be considered reliable. [37]

The overview of MCMC, Metropolis, and HMC algorithms presented in this section is intentionally surface-level, but sufficient to convey the core concepts. The purpose of including this section is to give the reader a general understanding of how posterior samples are obtained in practice. For readers interested in more details, the HMC algorithm is explained in article by Neal [34], and the No-U-Turn Sampler, is described in article by Homan and Gelman [36]. Further discussion of convergence diagnostics, including $\widehat{R}$ and effective sample size, can be found in article by Vehtari et al. [37].

4.6 Bayesian model checking and comparison

Model checking and comparison are essential components of regression analysis. The Bayesian paradigm offers a variety of evaluation techniques that are broadly applicable across regression models. This section discusses a selection of these methods, including posterior predictive checks, leave-one-out cross-validation (LOO-CV), and diagnostics for the proportional odds assumption.

A distinctive aspect of Bayesian model checking is the use of posterior predictive checks (PPCs). The basic idea behind PPCs is that data generated from the model should resemble the observed data. This can be assessed by drawing replicated datasets (replicants) from the posterior predictive distribution, and comparing these to the actual observations. One common approach is to plot the distributions of these replicants alongside the observed data, which allows for visual assessment of model fit. The posterior predictive distribution is given by

$$p(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{y}) = \int p(\boldsymbol{y}^{\text{rep}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta},$$

where $\boldsymbol{y}^{\text{rep}}$ denotes a replicated version of the observed dataset $\boldsymbol{y}$. This differs from the predictive distribution shown in Equation 10, as it aims to generate new realizations of the same process that produced the observed data, rather than predicting future or unseen observations. Because each draw from the posterior leads to a different replicant, it is good practice to generate multiple replicates for comparison. [6]

In the case of the cumulative logistic model, posterior predictive checks can be used to compare predicted ordinal category counts to the observed counts. A natural way to do this is by using bar plots. Since multiple replicants are typically generated, intervals can be plotted for each category to visualize uncertainty. This provides a visual summary of how well the model captures variation in the observed outcomes. Posterior predictive counts can also be compared across levels of categorical or binary covariates, providing a useful diagnostic for assessing fit at the subgroup level. Additional visualization tools include density plots, error plots, and summary statistics to compare replicated and observed values.

In addition to visual inspection, posterior predictive distributions can be used in test quantities $T(\boldsymbol{y}, \boldsymbol{\theta})$. If the test quantity does not depend on the model parameters, it is denoted by $T(\boldsymbol{y})$. A test quantity is a scalar summary of the data

and the parameters that allows for formal comparison between the observed data and replicated datasets. The model is considered to fit well if the observed test statistic $T(\boldsymbol{y})$ is consistent with the distribution of the replicated test statistic $T(\boldsymbol{y}^{\text{rep}})$, estimated from replicated datasets: $T(\boldsymbol{y}^{\text{rep }1}), \ldots, T(\boldsymbol{y}^{\text{rep }S})$, where S is the number of replicates. The choice of test quantity depends on the context and should reflect important aspects of the model and data. A given test quantity can also be used to compute a Bayesian p-value, which estimates the probability that a replicated dataset exhibits more extreme behavior than the observed data, according to the chosen quantity. A Bayesian p-value near 0 or 1 suggests poor model fit, while values near 0.5 are ideal. [6]

One test quantity that can be used for the cumulative logistic model is the predicted count in each ordinal category. The test statistic for category j is defined as

$$T_j(\boldsymbol{y}) = \sum_{i=1}^{n} 1(y_i = j), \quad j = 1, \ldots, J,$$

where $1(y_i = j)$ equals 1 if observation y_i falls in category j, and 0 otherwise. The test statistic is computed for both the observed data and each replicated dataset. Model fit is then evaluated separately for each category.

Another approach to model evaluation and comparison is cross-validation, which estimates the out-of-sample predictive performance of a model using only within-sample fits. As mentioned earlier, all Bayesian models in this thesis were fit using Stan, which implements Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO). While traditional LOO-CV is computationally expensive, requiring the model to be refit n times, PSIS-LOO offers a fast approximation by reweighting posterior draws using importance sampling. The model's predictive performance is measured using the expected log pointwise predictive density:

$$\widehat{\text{elpd}}_{\text{psis-loo}} = \sum_{i=1}^{n} \log \left(\frac{\sum_{s=1}^{S} w_i^s p(y_i \mid \boldsymbol{\theta}^s)}{\sum_{s=1}^{S} w_i^s} \right),$$

where w_i^s are the Pareto-smoothed importance weights, and $p(y_i \mid \boldsymbol{\theta}^s)$ is the posterior value of observation y_i under posterior draw $\boldsymbol{\theta}^s$. A higher value of $\widehat{\text{elpd}}_{\text{psis-loo}}$ indicates better model fit and predictive accuracy. [38]

When using PSIS-LOO, it is also important to consider the estimated Pareto shape parameter $\hat{k}$, which comes from the generalized Pareto distribution used in calculating the importance weights. Ideally, all $\hat{k}$ values should be below 0.5, which indicates high-quality importance sampling. Values between 0.5 and 1.0 are still acceptable but suggest that the estimates may be less stable. Values exceeding 1.0 are considered problematic, as they indicate that the PSIS-LOO estimate may not be reliable. [38]

The model's performance can also be estimated using a train-test split. In this approach, the data are split into a training set and a test set. The model is fitted on the training set, after which its performance is evaluated using the test set. For the cumulative logistic model, this performance can be evaluated, for example, by

assessing how well the model estimates the category-specific probabilities in the test set.

Testing the proportional odds assumption is essential when fitting cumulative logistic models. In the Bayesian framework, one approach is to fit an alternative ordinal regression model from a different family, for example the adjacent-category logistic model, which allows for category-specific effects. This model can then be compared to the cumulative logistic model using PSIS-LOO [7]. If the model that relaxes the proportional odds assumption shows clearly better predictive performance, the proportional odds assumption should be questioned. For a more complete comparison, one should also fit a version of the alternative model that maintains the proportional odds assumption to disentangle the effects of the modeling family from the odds assumption. Another approach is to use the highest density intervals (HDIs) and regions of practical equivalence (ROPE) to evaluate whether category-specific effects differ significantly, thereby testing the assumption directly [27]. In this thesis, the assumption will be tested using the LOO-based comparison approach, as it is conceptually straightforward.

The model checking and comparison methods discussed in this section are not exhaustive, but they have been selected for their practical applicability, widespread use, and availability of reliable software tools. It is important to emphasize that the convergence diagnostics discussed in the previous section are also a critical part of the Bayesian model checking process. These checks should always be performed before conducting posterior predictive checks or model comparisons. If convergence has not been achieved, any inferences drawn from the model may be invalid.

5 Simulation study

To demonstrate the advantages of using the Bayesian approach for fitting the cumulative logistic model, this section presents a simulation study. Simulations allow models to be fit in a controlled setting where the true data-generating processes are known, enabling a direct comparison between the frequentist and Bayesian approaches. The comparison will focus on parameter estimates and category probability estimates, as these are usually the most important parts of inference based on cumulative ordinal models. Also the times required to fit the models will be reported. All estimates will be compared against the true parameter values and true category probabilities used in the simulation. The specific R code used for data generation and analysis is provided in a GitHub repository [39]. Supplementary tables and figures related to the simulation study are included in Appendix A.

5.1 Models

To generate the data, the structure of the linear predictor and the link function must be specified. All simulations use a five-category ordinal outcome, consistent with the ordinal survey data from the live music census. Threshold values were computed by transforming predefined category probabilities using the inverse logistic function. These thresholds were used to discretize simulated values of a latent continuous variable $\tilde{Y}$ into ordinal responses. The form of the regression model used in the simulations was the following:

$$\text{logit}(P(Y \leq j)) = \alpha_j - \beta_1 x_{i1}, \quad i = 1, \ldots, n, j = 1, \ldots, 4,$$

where x_{i1} is a binary group indicator: it takes the value 1 if individual i belongs to group 1, and 0 if they belong to group 0. Thus, the model includes one explanatory variable that distinguishes between two groups, labeled 0 and 1. In all simulations, the covariate effect was set to $\beta_1 = \log(3) \approx 1.098$, implying that the odds of responding in a higher category are greater for group 1. Although additional explanatory variables could be added, these simulations focused on a simple model with one covariate to ensure clarity and interpretability.

One of the simulations was conducted with a hierarchical data structure, reflecting the structure of the live music census, which was collected across five cities. This means that the form of the linear predictor included a random intercept term. The simulated data included five clusters, and the linear predictor took the following form:

$$\text{logit}(P(Y_c \leq j)) = \alpha_j + u_c - \beta_1 x_{i1}, \quad i = 1, \ldots, n, j = 1, \ldots, 4, c = 1, \ldots 5,$$

where u_c is the cluster-specific random intercept. In the hierarchical scenario, the cluster-specific effects were first simulated and then used as a part of the data generation process.

The Bayesian models included the prior distributions defined in Equations 11 and 12, while the priors for the hierarchical model were those in Equations 14 and 15. The hyperparameter values used for the threshold priors were $\mu_{\alpha_1} = -1.4$, $\mu_{\alpha_2} = -0.4$, $\mu_{\alpha_3} = 0.4$, and $\mu_{\alpha_4} = 1.4$. These values were based on the assumption

that each ordinal category has an equal probability when all explanatory variables are set to zero. For a five-category ordinal variable, this corresponds to 20% probability per category. Using Equation 4, the cumulative probabilities can be transformed into log-odds, giving the threshold values. These are weakly informative priors, as they do not assume that any specific category is more or less likely. These threshold hyperparameters are used across all the Bayesian models in this thesis.

5.2 Implementation

The ordinal data were simulated using code adapted from Gambarota and Altoè [25]. Ordinal data can be simulated either through a multinomial distribution or using a latent variable formulation. The latent approach was used here, as it is both more computationally efficient and consistent with the underlying theoretical framework of ordinal regression. Each simulation scenario included 50 generated datasets.

All models were implemented in R (version 4.4.1) [40]. Frequentist cumulative logistic models were fitted using the `ordinal` package (version 2023.12-4.1) [41], which provides a flexible interface and additional tools not available in alternatives such as the `MASS` package [42]. The confidence intervals for probability estimates of the frequentist models were calculated using function `emmeans()` from package `emmeans` (version 1.10.7) [43]. Bayesian models were fit using Stan (version 2.36.0) [35], which compiles and executes models in C++. To avoid manual C++ coding, the `brms` package (version 2.22.0) [44] was used. This package allows specifying Bayesian regression models using a syntax similar to functions `lm()` and `glm()`, while handling Stan code generation and sampling in the backend.

The frequentist models were fit using the `clm()` function from the `ordinal` package, while the Bayesian models were fit using the `brm()` function from the `brms` package. It is important to note that `clm()` does not produce an estimate for categories with zero observations. While the `polr()` function from the `MASS` package can technically accommodate such cases, it was not used here due to its limited functionality and because the resulting estimates would still be unstable. The Bayesian hierarchical model was also fit using the `brm()` function, while the frequentist hierarchical model was fit using the `clmm()` function from the same `ordinal` package.

The performance of the models in estimating category probabilities was assessed quantitatively. This was done using the mean squared error (MSE). MSE was calculated by comparing the estimated category probabilities to the true probabilities used to generate the data. The Bayesian estimates were based on the posterior expected values. It was computed either at the overall level, at different levels of a categorical explanatory variable or for the response categories separately. In addition to category probabilities, the accuracy of parameter estimates was also evaluated. This was done by checking whether the true parameter value falls within the estimated 95% interval, specifically the confidence interval for the frequentist model and the credible interval for the Bayesian model.

5.3 Results

The first simulation was conducted under an ideal scenario with a relatively large number of observations. A total of 600 observations were generated, 300 for each group defined by the binary covariate. All categories were assigned equal probabilities at the baseline level of the covariate, that is, when $x_{i1} = 0$. All the convergence diagnostics for the Bayesian model indicated good performance. Unless stated otherwise, this is assumed to be the case for the other Bayesian models as well. The frequentist models were faster to fit than the Bayesian models were to sample from the posterior (Table 3).[1] It is important to note that the initial model compilation time is not included in the Bayesian sampling time. The model compilation times can be found in Appendix A.

In this setting, both models performed equally well in estimating the model parameters. The results were nearly identical for both methods. The frequentist model had a slightly lower MSE, while the Bayesian model showed slightly better coverage of the true coefficient value (Table 3).

Model	Bayesian MSE	Frequentist MSE	Bayesian coefficient coverage (%)	Frequentist coefficient coverage (%)	Average Bayesian sampling time	Average frequentist fitting time
Optimal scenario	0.00401	0.00396	100	98	3.4 s	73 ms
One category with a low response probability	0.00438	0.00478	100	96	1.5 s	39 ms
Two categories with low response probabilities	0.00608	0.00741	94	94	1.8 s	38 ms
Hierarchical scenario	0.00244	0.00269	96	95	53.4 s	0.14 s

Table 3: Performance metrics of the two modeling approaches across the four simulation scenarios. The Bayesian model generally performs better, particularly in scenarios with categories that have low probabilities. It achieves lower MSE in all but the first scenario and better coverage of the true coefficient value. The frequentist model was consistently faster to fit, especially in the last scenario. The table includes MSE, coefficient coverage, and sampling/fitting time for both models across the four scenarios.

The second simulation scenario was designed to reflect a situation observed in the live music census data, where one of the response categories receives no responses. This can occur when a category has a low response probability and the sample size is small. In this scenario, the base probability for the first category was set to 0.05, and 50 observations were generated per simulation, with 25 per group. As a result, some simulations resembled the live music census data, where the first category received no responses. Both models converged slightly faster under this scenario (Table 3).

Overall, the Bayesian model appears to recover the true category probabilities more accurately in this scenario (Figure 4). The median MSE is generally lower, and its distribution is more concentrated toward smaller values. The Bayesian model also shows a lower overall MSE compared to the frequentist model and achieves

[1] All models in this thesis were fit on a 2024 MacBook Air with an Apple M3 chip and 16 GB of RAM.

better coverage of the true coefficient values (Table 3). Although it is not visible in Figure 4, it is important to note that in scenarios where the first category received no responses, the frequentist model does not provide a probability estimate for that category. Therefore, even though the frequentist model appears to have a slightly lower MSE for the first category in general, this limitation should be taken into account.

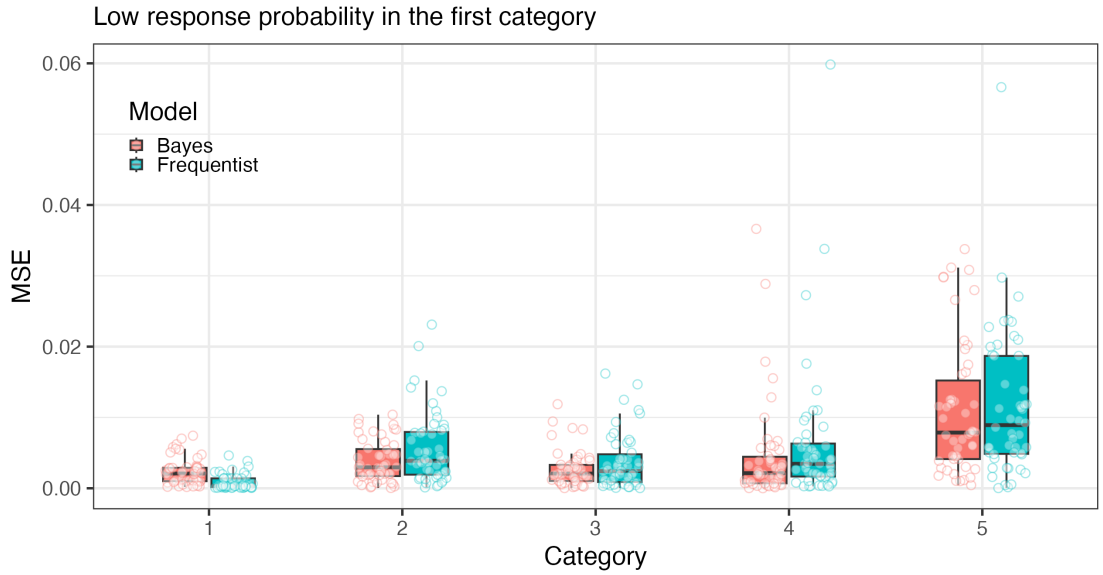

Figure 4: Scenario where the first category has a low response probability and the sample size is small. The Bayesian model shows better performance in estimating category probabilities. Box plots of the MSE values for both models across the five response categories in the 50 simulated datasets.

The live music census data also include a situation where two of the ordinal response categories have no observations. To reflect this, the first two categories were each assigned a base probability of 0.05 in the simulation, and 50 observations were again generated per simulation. This setup can lead to scenarios where no responses are observed in those categories. The frequentist models fit without issue, whereas the Bayesian model occasionally produced warnings about divergent transitions, indicating potential convergence issues. These were resolved by increasing the target acceptance rate (`adapt_delta`) from the default 0.95 to 0.99. As in previous scenarios, the frequentist model was faster to fit (Table 3).

In this scenario, the Bayesian model again outperformed the frequentist model in estimating category probabilities (Figure 5). It achieved lower MSE in most categories. As before, when there were no observations in the first two categories, the frequentist model did not provide estimates for them. Both models showed similar coverage of the true coefficient values, but the Bayesian model had a clearly lower overall MSE (Table 3).

The final simulation demonstrates a hierarchical data scenario. The total sample size is 200, divided into five clusters with sizes varying from 20 to 60. The category probabilities were same as in the previous scenario. In this setting, the Bayesian hierarchical model took significantly longer to sample than the time required for the

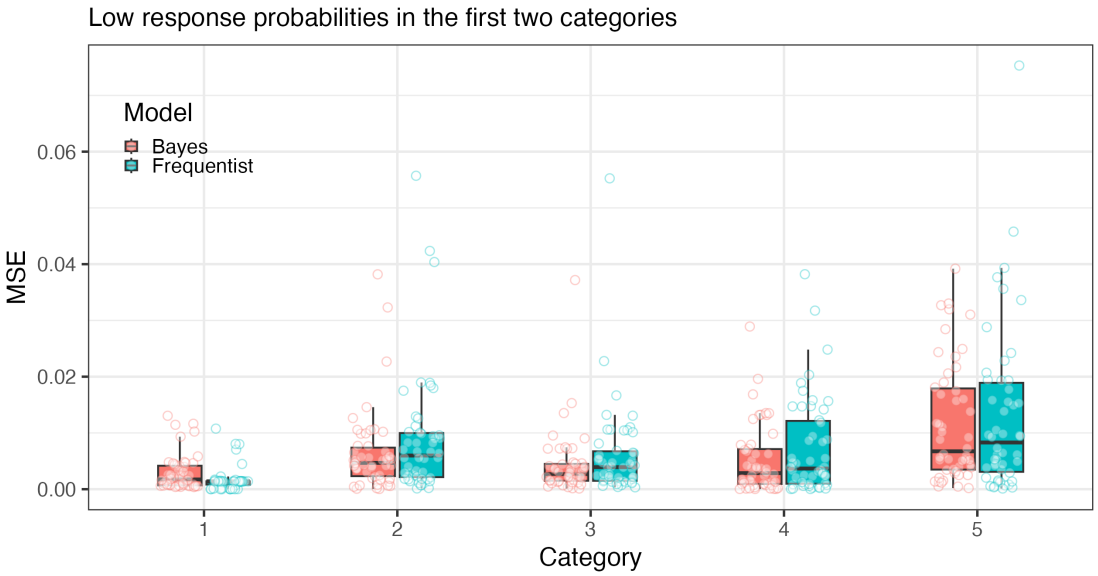

Figure 5: Scenario where the first two categories have low response probabilities and the sample size is small. The Bayesian model provides more accurate probability estimates. Box plots of the MSE values for both models across the five response categories in the 50 simulated datasets.

frequentist model to fit (Table 3). This increase was due to the additional complexity introduced by the hierarchical structure. Additionally, the acceptance rate for the Bayesian model had to be increased to 0.99 again.

The performance of the two models was generally comparable, with the Bayesian model performing slightly better in this scenario (Figure 6). Both the overall MSE and the coverage of the true coefficient value were marginally better in the Bayesian results (Table 3).

Although the performance differences were small, the Bayesian model offers several practical advantages in hierarchical settings. In the Bayesian framework, cluster-specific probability estimates and credible intervals can be directly obtained from the posterior samples. In contrast, in the frequentist framework, such estimates must be calculated manually, and no R package appears to currently provide built-in functionality for this. Moreover, confidence intervals typically require resampling methods such as bootstrapping. The Bayesian approach also demonstrated greater robustness in simulations with fewer clusters. While five clusters were used here to match the structure of the live music census data, additional tests with fewer clusters, such as four or three, showed that the frequentist model struggled with convergence and model fitting.

In summary, the simulation study demonstrated that the Bayesian cumulative logistic model generally outperformed the frequentist alternative, particularly in scenarios involving sparse data or possible response categories with no observations. While both models performed similarly under ideal conditions, the Bayesian model provided more accurate category probability estimates and better coverage of the true coefficient values in more challenging scenarios. However, it was consistently slower to fit. It is important to note that the simulations were limited in scope, relying on

a relatively simple data-generating process and a small number of scenarios.

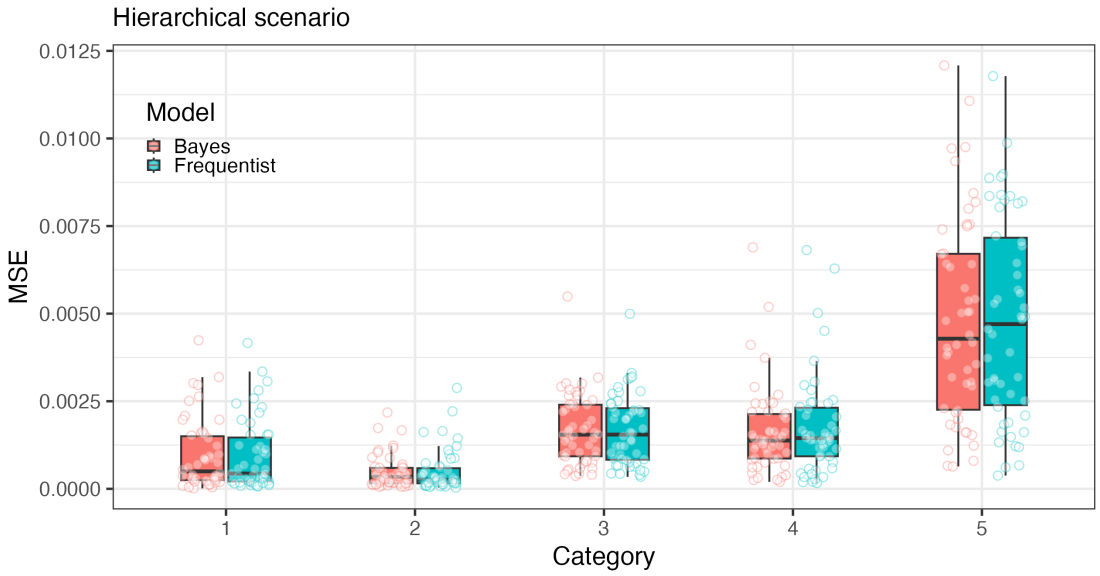

Figure 6: Scenario with a hierarchical data structure. The Bayesian approach performs slightly better in the hierarchical setting. Box plots of the MSE values for both models across the five response categories in the 50 simulated datasets.

6 Analysis of live music census data

This section applies the cumulative logistic model to real-world ordinal data from the live music census. The goal of using Bayesian modeling is to address several common challenges in analyzing ordinal survey responses. These include estimating meaningful and robust category probabilities even in the presence of sparse data or categories with no responses, obtaining more interpretable and coherent measures of uncertainty, and leveraging hierarchical structures to improve inference across clusters. These issues frequently arise in applied research but are not always well handled by standard frequentist approaches. To demonstrate how Bayesian modeling can help overcome these challenges, both Bayesian and frequentist cumulative logistic models are fit to the data and compared. The models are evaluated using a train-test split, and their performance is assessed based on category probability and parameter estimates. In the final part of the analysis, the Bayesian model is used to summarize and interpret overall patterns in the data, including differences across cities, age groups, and genders. The code used in this section is also available in the GitHub repository [39].

6.1 Data description

The live music census data used in this analysis were collected in five European cities: Helsinki, Vilnius, Mannheim, Lviv, and Heidelberg. Three different surveys were conducted: two for music venues in these cities, and one for the audience attending live music performances on October 11, 2024. These surveys were part of the OpenMusE EU project, which aims to make the European music industry more competitive, fair, sustainable, and transparent [8]. This is a unique dataset, with only few prior examples of similar censuses, one of them being the UK Live Music Census conducted in 2017 [45]. What makes the data particularly distinctive is that they were collected simultaneously across multiple countries.

The live music audience census was conducted over a 24-hour period, starting at noon on October 11, 2024, across the five participating cities. The survey was distributed to attendees of live music performances and included questions about spending behavior, experience of the night, music preferences, opinions on the local live music scene, and more. After data cleaning, the dataset included responses from 409 individuals. There were slightly more female respondents, and a few participants identified with another gender identity (Table 4). The first two age groups had a similar number of respondents, while the older age groups were less represented. In terms of geography, Helsinki had the highest number of responses, followed by Vilnius, Mannheim, and Lviv. The smallest group was from Heidelberg, which had only seven respondents.

In addition to the audience survey, two venue surveys were conducted: one prior to and one following the audience census. The pre-census survey collected information on general operations, including which days and months venues were open, the types of events and activities hosted, and the challenges faced by venues. The post-census survey focused on details specific to the night of October 11, including attendance, venue capacity, staff composition, and financial data. The first venue survey received

Variable	Category	Total responses	Scene rating responses	Night rating responses
Gender	Female	218	202	140
	Male	174	160	113
	Another gender identity	17	16	8
Age group (years)	18-29	169	151	89
	30-49	156	148	109
	50-64	63	60	49
	>64	21	19	14
City	Helsinki	164	156	137
	Vilnius	118	108	37
	Mannheim	68	61	51
	Lviv	52	47	33
	Heidelberg	7	6	3

Table 4: Demographic breakdown of the live music audience census data. The "Total responses" column indicates the number of respondents in each demographic category who completed the survey. The "Scene rating responses" and "Night rating responses" columns indicate the number of respondents in each category who provided a valid answer to the questions "How would you rate the live music scene in your city?" and "How was your night?", respectively. Note the smaller sample sizes for these specific questions due to item nonresponse.

72 responses, and the second received 56.

This thesis focuses on analyzing two ordinal five-point Likert scale questions from the audience survey:

- "How would you rate the live music scene in your city?"
- "How was your night?"

Both questions used the same response scale (Table 5). The question "How was your night?" received no responses in the first two categories. The other question, regarding the live music scene, includes some responses in the lower categories, although certain cities recorded no responses in them. Both questions exhibit item nonresponse, but the night rating question clearly had a lower overall response rate, particularly in some cities and demographic categories (Table 4). As showcased in the simulation section, Bayesian methods should perform better in analyzing these questions.

Although the audience survey includes two other ordinal variables, one on event attendance frequency and another on preferred venue size, these use a different answer scale than the Likert scale. To keep the scope of the analysis manageable, these variables are not analyzed. Likewise, although the venue surveys include ordinal variables, they are excluded for the same reason.

Question	Answer options					No response
	Very bad	Bad	Average	Good	Very good	
How would you rate the live music scene in your city?	6	12	89	212	59	31
How was your night?	0	0	21	109	131	148

Table 5: Response counts for the different categories of the ordinal variables analyzed. The "How was your night?" question received no responses in the first two categories, while the "How would you rate the live music scene in your city?" question had only a few responses in those categories.

6.2 Model validation and predictive performance

The simulation section demonstrated that the Bayesian model outperformed the frequentist alternative in several simulated data scenarios. This section extends that comparison to the live music audience census. As the true category probabilities are not known, the Bayesian and the frequentist approaches will be compared using a train-test split. In this setup, models are fit using the training set and then assessed using estimated values from the test set. The Helsinki data are used as the test set, while all other observations form the training set.

For the question about the live music scene, the training set contains 222 observations, and the test set contains 156. For the question about how the person's night went, the training set has 124 observations, and the test set has 137. Model performance is evaluated by assessing how well the estimated intervals from each model cover the category probabilities observed in the test set, and by calculating the mean squared error between the predicted and test set probabilities. Unless otherwise stated, all Bayesian models converged well, based on the standard diagnostics. The proportional odds assumption was also tested for all the analyzed models using the method stated in the section about model checking.

The first analysis focuses on the question: "How would you rate the live music scene in your city?" Multiple model specifications were tested using the `loo()` function from the `brms` package, which implements the LOO-CV method discussed before. Based on these comparisons, a model including gender and age group as explanatory variables was selected. This version slightly outperformed the hierarchical model. While a hierarchical model would have been theoretically justified, since the data were collected from several different cities, the frequentist version of the hierarchical model encountered convergence issues during testing. To enable a comparison between the Bayesian and frequentist approaches, the non-hierarchical version was chosen for the analysis. The frequentist model was fitted using the same set of explanatory variables as the Bayesian model.

Although the frequentist intervals are generally narrower, the Bayesian model demonstrates better coverage of both the training and test sets (Figure 7). The frequentist intervals appear overconfident, while the Bayesian model more effectively

captures the uncertainty in the estimates. The frequentist confidence intervals can also be unreliable in groups with sparse observations. This is the case with "Very good" category for respondents identifying with another gender identity, where the frequentist lower bound is negative and thus not interpretable. The MSE is also lower for the Bayesian model (Table 6), indicating better performance in estimating category probabilities. Similar patterns were observed in both the age group and aggregate comparisons, where the Bayesian model consistently performed better (Appendix A).

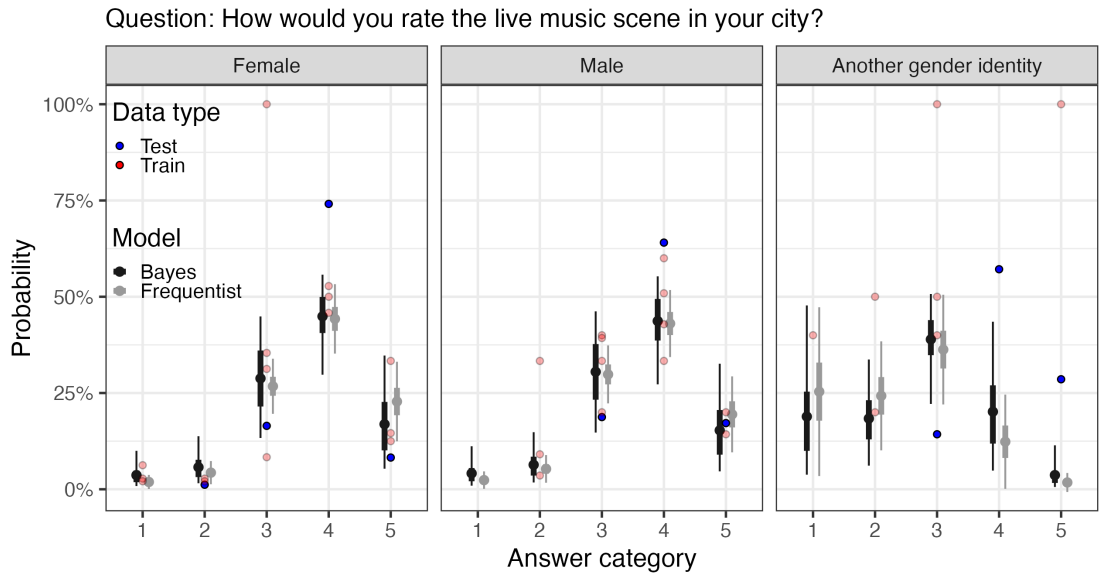

Figure 7: Estimated category probabilities from non-hierarchical models for the "How would you rate the live music scene in your city?" question. The Bayesian model better captures the observed category probabilities in both the training and test sets. The Intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at different levels of the explanatory variable (gender). 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 (Helsinki).

Question	Bayesian MSE	Frequentist MSE
How would you rate the live music scene in your city?	0.033	0.042
How was your night?	0.036	0.067

Table 6: The MSE values for both questions from the Bayesian and the frequentist models. The Bayesian model has lower MSE for both questions, indicating better overall accuracy in estimating category probabilities. The MSE values were calculated by comparing the estimated category probabilities to the corresponding observed probabilities in the test set.

For the question "How was your night?", multiple sets of explanatory variables

were again tested and compared using LOO-CV. This time, all models performed similarly, with differences in LOO-CV scores well within one standard error of each other. Based on the principle of parsimony, the simplest model would typically be selected. However, for comparison purposes, the model was fit using the same explanatory variables (gender and age group) as in the first model.

Once again, the Bayesian model more accurately captures the category probabilities observed in the test set (Figure 8). Notably, the frequentist model does not provide estimates for the first two categories ("Very bad" and "Bad"), as there were no observations in those categories within the training set. The MSE is substantially lower for the Bayesian model (Table 6), further indicating a better overall fit. Several of the frequentist intervals for the another gender identity group are also problematic, with lower bounds falling below 0% and upper bounds exceeding 100%, both of which are implausible for probability estimates. Similar patterns of Bayesian model performing better were observed in both the aggregate and age group comparisons (Appendix A).

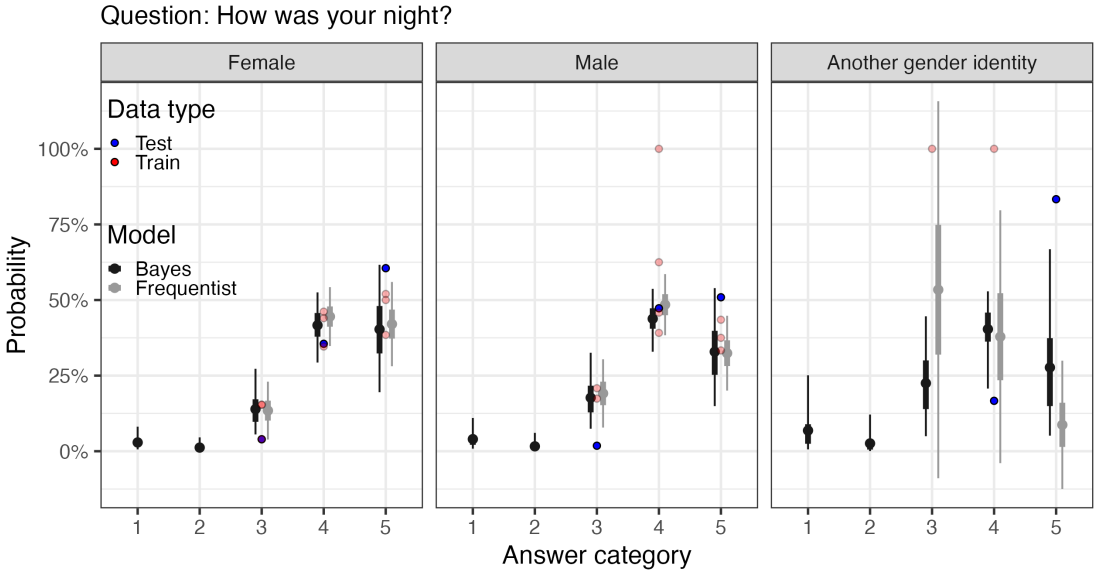

Figure 8: Estimated category probabilities from non-hierarchical models for the "How was your night?" question. The Bayesian model better captures the observed category probabilities in both the training and test sets, and also provides estimates for categories with no observed responses. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at different levels of the explanatory variable (gender). 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 (Helsinki).

Overall, these results further demonstrate that the Bayesian model seems to provide more reliable estimates for category probabilities, particularly in unseen or sparse test data scenarios. Its estimates are more robust, less sensitive to zero counts, and free from the structural issues encountered in the frequentist intervals. The coefficient estimates from the Bayesian model are also more precise, with narrower intervals than those from the frequentist model (Figure 9). This increased precision

results from the prior distribution on the coefficients, which introduces regularization by shrinking estimates toward zero.

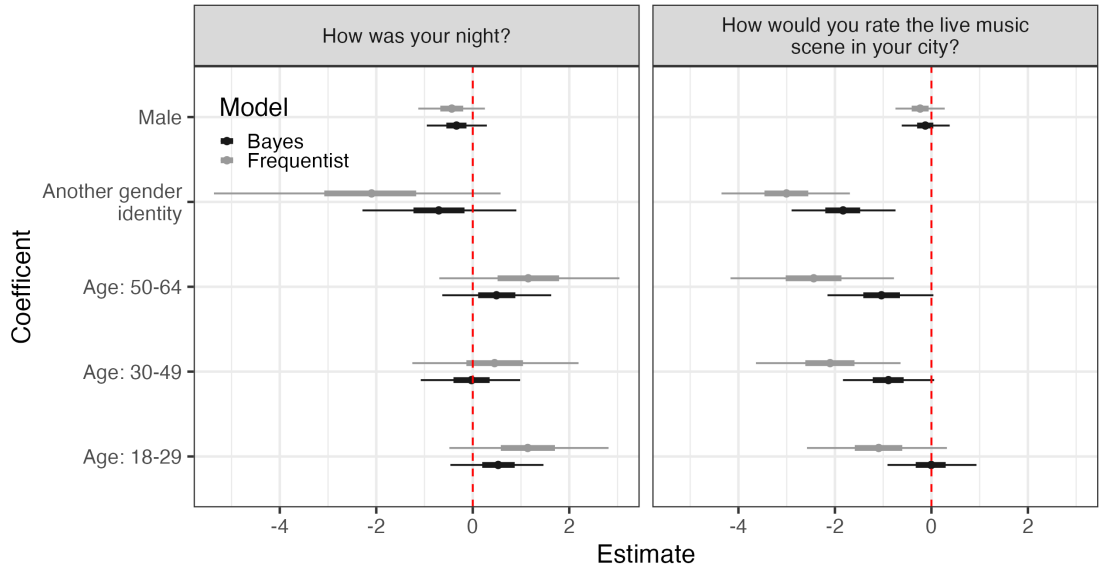

Figure 9: Coefficient (log-odds) estimates for both questions from the Bayesian and frequentist models. The Bayesian model's estimates are more precise. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models. The reference categories are female for gender and 65 and over for age group.

6.3 Interpretation of the Bayesian hierarchical model results

To conclude this section on modeling the live music census data, Bayesian cumulative logistic models are fitted to the full dataset. The purpose of this final analysis is to examine the overall effects of the explanatory variables and to interpret city-specific patterns. In applied use of the cumulative logistic model, this type of summary analysis is typically of primary interest. This analysis was conducted for both survey questions analyzed in the previous section.

The first model was fitted to the question "How would you rate the live music scene in your city?", and as before, LOO-CV was used to compare alternative model specifications. The best-fitting model was a hierarchical model that included a city-specific varying intercept and fixed effects for gender and age group. This differs from the model selected in the previous section, where the hierarchical version was not favored. The inclusion of the full dataset and an additional city appears to make the hierarchical structure more appropriate in this setting. For the second question, "How was your night?", the same model structure was selected, also based on LOO-CV results.

Interpreting the effect of covariates in ordinal regression models can be challenging due to the non-linear nature of the outcome. To simplify interpretation, the effects of gender and age group are also evaluated using the predicted mean response, calculated by multiplying each category's probability estimate by its corresponding

response value and summing across categories. This is done in addition to looking at the coefficient estimates.

The mean responses across covariate levels suggest that the explanatory variables have little or no effect on the rating (Figure 10). Respondents identifying with another gender identity rated the live music scene slightly lower on average, while those aged 18–29 reported slightly higher ratings. Overall, the predicted mean response across groups falls between "Average" and "Good". This lack of strong group differences is also evident in the coefficient estimates (Figure 11). All the credible intervals include zero, with the another gender identity group having the largest negative log-odds, and the 18–29 age group having the largest positive log-odds. These patterns are consistent with the small differences observed in the predicted mean responses.

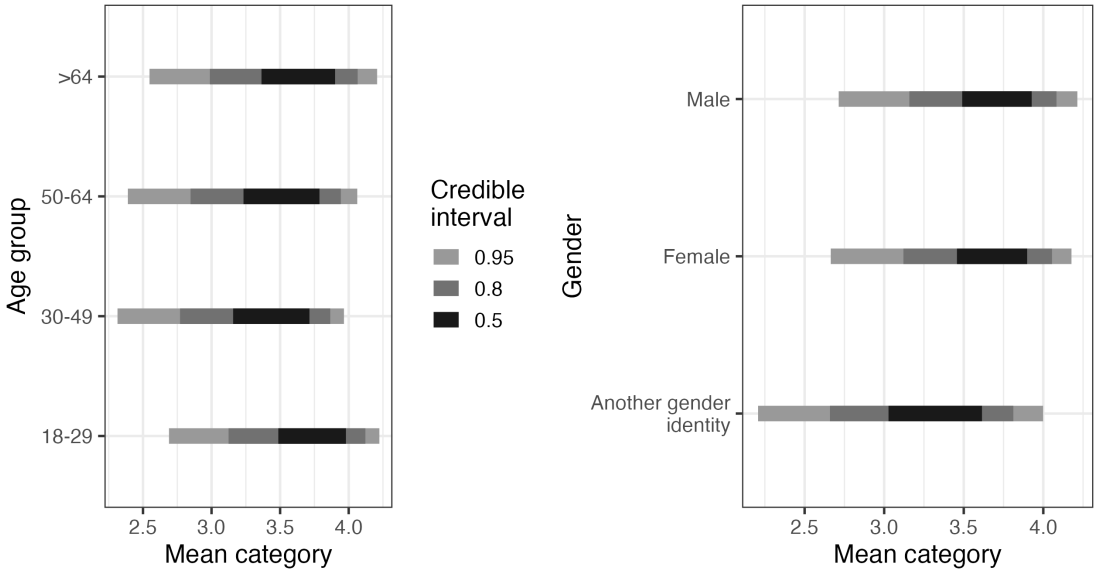

Figure 10: Mean response to the question "How would you rate the live music scene in your city?" by age group and gender, estimated using the Bayesian hierarchical model. The plot shows small differences between groups, consistent with the coefficient estimates, which were close to zero and had credible intervals that included zero (Figure 11). Shaded areas represent the 50%, 80%, and 95% credible intervals.

City-specific probability estimates show some variation between the cities (Figure 12). These results suggest that respondents in Helsinki and Lviv were slightly more likely to rate their city's live music scene as "Good" or "Very good". In contrast, responses from Heidelberg show higher probabilities for the lower rating categories, though this is likely a consequence of the very small sample size for that city. Correspondingly, the credible intervals for Heidelberg are substantially wider, reflecting the greater uncertainty in the estimates due to limited data.

The mean response estimates from the second model, fitted to the "How was your night?" question, do not indicate any substantial effects (Figure 13). In this case, there appear to be no meaningful differences between groups, and even the slight variation observed in the first model is no longer present. This is also reflected in the coefficient estimates, which are now even closer to zero than in the first model (Figure 11). The predicted mean response across groups falls between "Good"

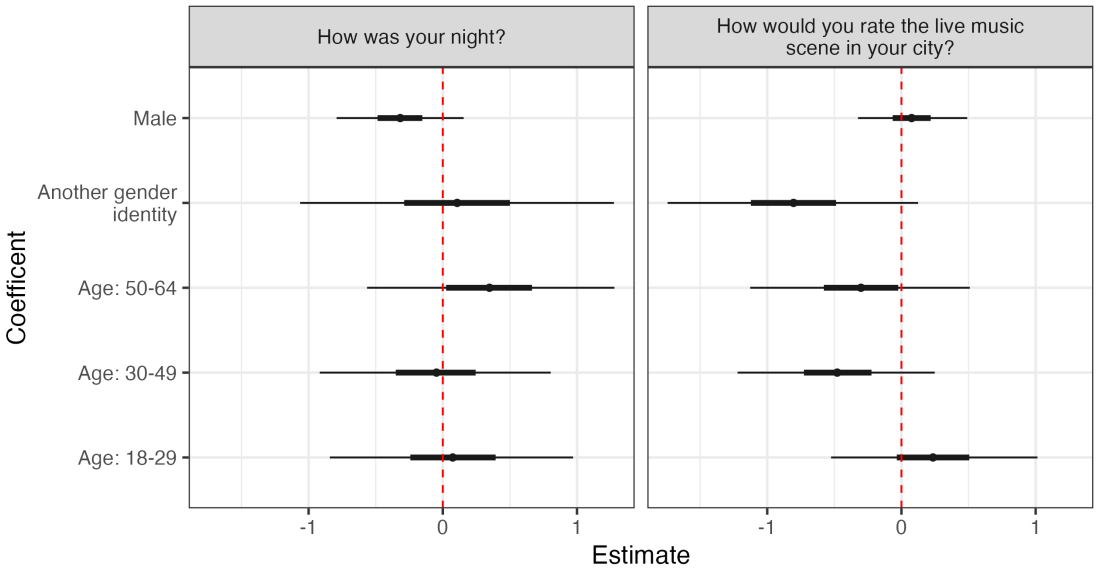

Figure 11: Coefficient (log-odds) estimates for both questions, with 50% and 95% credible intervals from the Bayesian hierarchical models. Most estimates are close to zero, and all intervals include zero, indicating no substantial group-level effects. The reference categories are female for gender and 65 and over for age group.

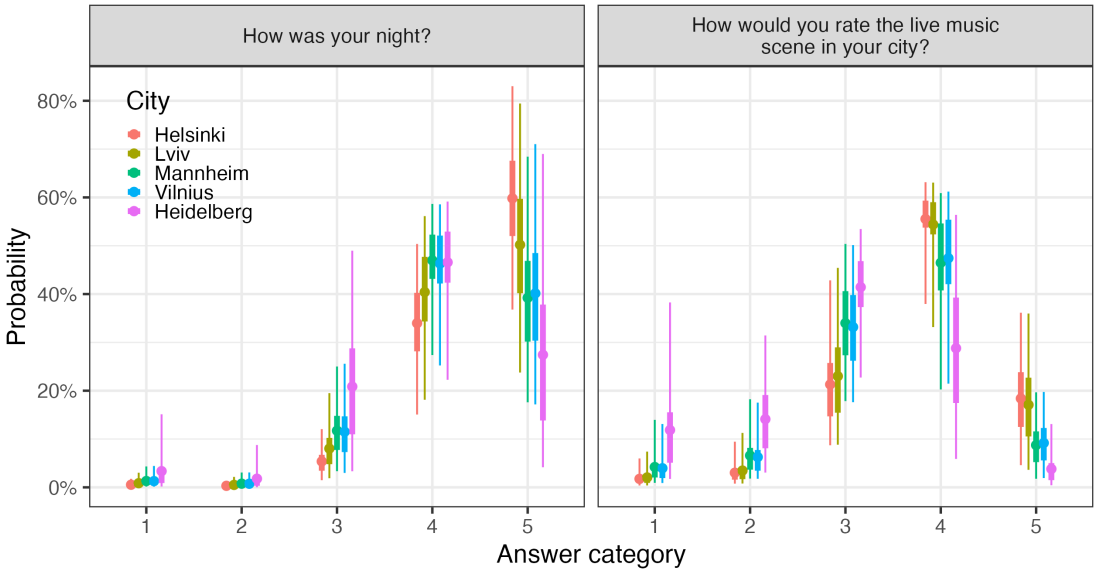

Figure 12: City-level differences in category probability estimates for both survey questions from the Bayesian hierarchical models. City-specific category probability estimates show variation across cities, with respondents from Helsinki and Lviv more likely to select the more positive response categories. Estimates are shown with 50% and 95% credible intervals for both questions.

and "Very good". The city-specific estimates for this model show some variation (Figure 12). As with the other question, respondents in Helsinki and Lviv were somewhat more likely to select higher response categories. Once again, the wider

credible intervals for Heidelberg reflect the limited number of observations from that city.

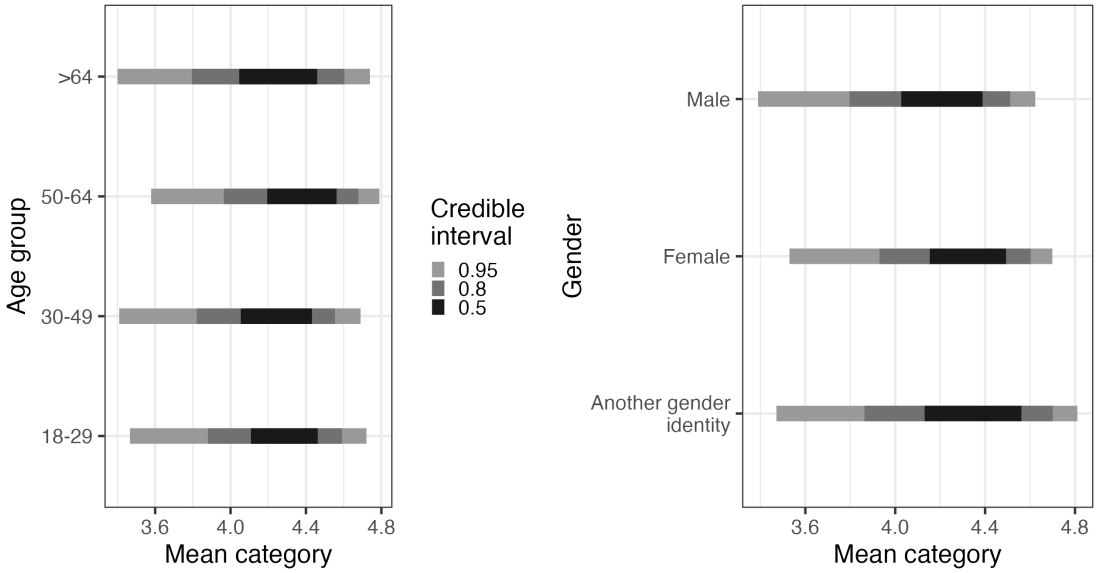

Figure 13: Mean response to the question "How was your night?" by age group and gender, estimated using the Bayesian hierarchical model. The plots shows no notable differences between groups, which is consistent with the coefficient estimates that were close to zero and had credible intervals including zero (Figure 11). Shaded areas represent the 50%, 80%, and 95% credible intervals.

Overall, the results from both models do not indicate any significant effects of the covariates on the responses. However, there is some city-specific variation, with Helsinki and Lviv consistently showing a higher probability of respondents selecting the more positive response options. The Bayesian framework enabled estimation even for categories with no responses, and allowed for the calculation of credible intervals for city-specific category probabilities, both of which would have been more difficult or unstable under a frequentist approach.

7 Discussion

The motivation for this thesis was to demonstrate the benefits to choosing the Bayesian framework when analyzing ordinal survey data, particularly when dealing with small sample sizes and sparse categories. In these situations traditional frequentist methods often fall short. The analysis focused on the cumulative logistic model, a widely used approach in the literature for modeling ordinal outcomes. The thesis introduced the theoretical background of ordinal regression, reviewed key aspects in Bayesian modeling, which included hierarchical extensions, prior specification, sampling methods, and model checking. These methods were applied to both simulated and real-world survey data. The goal was to illustrate how Bayesian inference can lead to improvements in estimation, uncertainty quantification, and flexibility in model structure.

The results suggest that the Bayesian approach may have certain advantages. In the simulation studies, the Bayesian models provided better coverage of the true probability values compared to the frequentist models (Figures 4 and 5, and Table 3). Similarly, in the train-test split analysis using the live music audience census data, the Bayesian model more accurately captured the category probability estimates, particularly for the test set (Figures 7 and 8, and Table 6). This highlights the robustness and applicability of the Bayesian model even in real-world applications.. The coefficient estimates from the Bayesian model also had narrower intervals compared to those from the frequentist model (Figure 9).

Despite these strengths, it is important to acknowledge the limitations of the model. The cumulative logistic model assumes proportional odds, which was not violated in the data analyzed here. However, this assumption may not hold in other datasets or settings, which could limit the model's broader applicability. Also, the Bayesian models used weakly informative priors, selected for their generality and simplicity. Their influence, particularly in small or sparse subgroups, was not explored in depth, and alternative priors could potentially yield different results.

This thesis relied on a single dataset and focused on just two ordinal variables, which was sufficient to demonstrate the modeling approach but may limit the generalizability of the findings. Future research should apply these methods to a broader range of datasets and ordinal variables to better assess their robustness. For example, the live music audience census includes additional ordinal variables, such as preferred venue size and monthly attendance frequency, that could be incorporated into further analyses. Similarly, the accompanying venue surveys contain several ordinal items that are also suitable for modeling. These variables were excluded here to keep the thesis scope manageable, but future research could revisit them.

While the simulation study illustrated key differences between Bayesian and frequentist approaches, it had some limitations. The scenarios were simplified, with a limited number of predictors and a known, correctly specified data-generating process, which may not reflect the complexity of real-world data. All simulations used fixed priors, and the effect of alternative prior choices was not explored. The sample sizes and number of response categories were also limited, and the study focused solely on the cumulative logistic model.

The models applied in this thesis were deliberately kept simple, with the most

complex case involving a hierarchical model with varying intercepts. The goal was to showcase the advantages of the Bayesian approach in models that are commonly used in applied research. Future studies could explore more advanced extensions, including location-scale or location-shift models, as well as more complex hierarchical structures. Another avenue for future work would be to explore the use of informative or structured priors, such as Dirichlet priors, particularly in scenarios with small sample sizes.

Further research could also involve fitting other ordinal regression models within the Bayesian framework, such as the adjacent-categories or sequential models, which are based on different assumptions about the latent process underlying ordinal responses [7]. These models could then be compared to their frequentist counterparts to evaluate differences in performance. Additionally, the cumulative model could be extended by testing alternative link functions. Another interesting regression model type to test would be mixture models, which allow for modeling of latent subpopulations within the ordinal data [4].

The analysis of the live music census data revealed relatively minor differences between levels of the explanatory covariates (Figure 11). There were some differences across cities, with respondents in Helsinki and Lviv generally giving higher ratings (Figure 12). However, the overall picture was consistent: respondents were, on average, satisfied with their city's live music scene and their night out.

While the focus of this thesis was on modeling ordinal data, there are also many Bayesian methods for addressing survey error, which were discussed briefly in the section about survey data. Further work could investigate how Bayesian ordinal regression performs when combined with methods for adjusting for common sources of survey error, including coverage bias, nonresponse, and sampling variance. This could include techniques like multilevel regression with poststratification [9] or Bayesian implementations of traditional weighting methods [10].

There was also a notable degree of item nonresponse in the audience census data. The question about rating the city's live music scene had 31 missing responses, meaning 7.6% of respondents did not answer. The question about rating the night had 148 missing responses, or 36.1% nonresponse. These missing values may affect the validity of the inferences drawn, particularly if the missingness is not completely random. In future work, the modeling approach could be extended by incorporating Bayesian imputation methods to account for item nonresponse and improve the robustness of the results [12].

In conclusion, this thesis has demonstrated both practical and theoretical advantages of using the Bayesian approach for modeling ordinal data. In simulated and real-world settings, it showed improved accuracy, robustness, and interpretability. These findings support Bayesian methods as a strong alternative to traditional approaches for analyzing ordinal survey data.

References

[1] D. A. Dillman, J. D. Smith, and L. M. Christian. *Internet, Phone, Mail, and Mixed-Mode Surveys : The Tailored Design Method*. John Wiley & Sons, Incorporated, 2014. ISBN: 9781394260645. DOI: https://doi.org/10.1002/9781394260645.

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

[3] T. M. Liddell and J. K. Kruschke. "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.

[4] G. Tutz. "Ordinal regression: A review and a taxonomy of models". In: *WIREs Computational Statistics* 12.2 (2022). DOI: https://doi.org/10.1002/wics.1545.

[5] R. McElreath. *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.

[6] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis: Third edition*. Chapman and Hall/CRC, 2013. ISBN: 9780429113079. DOI: https://doi.org/10.1201/b16018.

[7] P.-C. Bürkner and M. Vuorre. "Ordinal Regression Models in Psychology: A Tutorial". In: *Advances in Methods and Practices in Psychological Science* 2.1 (2019). DOI: https://doi.org/10.1177/2515245918823199.

[8] OpenMusE. *Project Info*. Accessed: 14 April 2025. URL: https://www.openmuse.eu/.

[9] M. Kołczyńska, P.-C. Bürkner, L. Kennedy, and A. Vehtari. "Modeling Public Opinion Over Time and Space: Trust in State Institutions in Europe, 1989-2019". In: *Survey Research Methods* 18 (2024). DOI: https://doi.org/10.18148/srm/2024.v18i1.8119.

[10] R. J. Little. "Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference". In: *Survey Methodology* 48.2 (2022). URL: http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00001-eng.htm.

[11] Z. Ma and G. Chen. "Bayesian methods for dealing with missing data problems". In: *Journal of the Korean Statistical Society* 47.3 (2018), pp. 297–313. DOI: https://doi.org/10.1016/j.jkss.2018.03.002.

[12] Sanju, V. Kumar, and P. Kumari. "Evaluating the Performance of Bayesian Approach for Imputing Missing Data under different Missingness Mechanism". In: *Sankhya B* 86.2 (2024), pp. 713–723. DOI: https://doi.org/10.1007/s13571-024-00344-w.

[13] T. S. Alamneh, A. W. Melesse, and K. A. Gelaye. "Determinants of Anemia Severity Levels among Children Aged 6–59 Months in Ethiopia: Multilevel Bayesian Statistical Approach". In: *Scientific Reports* 13.1 (2023). DOI: https://doi.org/10.1038/s41598-022-20381-7.

[14] D. C. Angus, L. Derde, F. Al-Beidh, D. Annane, Y. Arabi, A. Beane, W. van Bentum-Puijk, L. Berry, Z. Bhimani, M. Bonten, et al. "Effect of Hydrocortisone on Mortality and Organ Support in Patients With Severe COVID-19". In: *JAMA* 324.13 (2020), pp. 1317–1329. DOI: https://doi.org/10.1001/jama.2020.17022.

[15] M. Á. Beltrán-Sánchez, M.-A. Martinez-Beneito, and A. Corberán-Vallet. "Bayesian Modeling of Spatial Ordinal Data from Health Surveys". In: *Statistics in Medicine* 43.21 (2024), pp. 4178–4193. DOI: https://doi.org/10.1002/sim.10166.

[16] S. E. Lee, C. Vincent, V. S. Dahinten, L. D. Scott, C. G. Park, and K. Dunn Lopez. "Effects of Individual Nurse and Hospital Characteristics on Patient Adverse Events and Quality of Care: A Multilevel Analysis". In: *Journal of Nursing Scholarship* 50.4 (2018), pp. 432–440. DOI: https://doi.org/10.1111/jnu.12396.

[17] G. Stewart, A. Kamata, R. Miles, E. Grandoit, F. Mandelbaum, C. Quinn, and L. Rabin. "Predicting Mental Health Help Seeking Orientations among Diverse Undergraduates: An Ordinal Logistic Regression Analysis". In: *Journal of Affective Disorders* 257 (2019), pp. 271–280. DOI: https://doi.org/10.1016/j.jad.2019.07.058.

[18] R. de Freitas Souza, F. G. Lima, and H. L. Corrêa. "Multilevel Ordinal Logit Models: A Proportional Odds Application Using Data from Brazilian Higher Education Institutions". In: *Axioms* 13.1 (2024). DOI: https://doi.org/10.3390/axioms13010047.

[19] M. E. Lelisho, A. A. Wogi, and S. A. Tareke. "Ordinal Logistic Regression Analysis in Determining Factors Associated with Socioeconomic Status of Household in Tepi Town, Southwest Ethiopia". In: *The Scientific World Journal* 2022 (2022). DOI: https://doi.org/10.1155/2022/2415692.

[20] R. Sesay, M. Kpangay, and S. Seppeh. "An Ordinal Logistic Regression Model to Identify Factors Influencing Students Academic Performance at Njala University". In: *International Journal of Research and Scientific Innovation* 08 (2021), pp. 91–100. DOI: https://doi.org/10.51244/IJRSI.2021.8104.

[21] L. Zanin. "Estimating the Effects of ESG Scores on Corporate Credit Ratings Using Multivariate Ordinal Logit Regression". In: *Empirical Economics* 62.6 (2022), pp. 3087–3118. DOI: https://doi.org/10.1007/s00181-021-02121-4.

[22] C. M. Homan, G. Serapio-Garcia, L. Aroyo, M. Diaz, A. Parrish, V. Prabhakaran, A. S. Taylor, and D. Wang. "Intersectionality in Conversational AI Safety: How Bayesian Multilevel Models Help Understand Diverse Perceptions of Safety". In: *arXiv preprint* (2023). DOI: https://doi.org/10.48550/arXiv.2306.11530.

[23] M. Rezapour, S. S. Wulff, M. M. Amirarsalan, and K. Ksaibati. "Application of Bayesian Ordinal Logistic Model for Identification of Factors to Traffic Barrier Crashes: Considering Roadway Classification". In: *Transportation Letters* 13.4 (2021), pp. 308–314. DOI: https://doi.org/10.1080/19427867.2020.1728041.

[24] R. H. B. Christensen. "Cumulative Link Models for Ordinal Regression with the R Package Ordinal". In: *R package vignette* (2018). URL: https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf.

[25] F. Gambarota and G. Altoè. "Ordinal regression models made easy: A tutorial on parameter interpretation, data simulation and power analysis". In: *International Journal of Psychology* 59.6 (2024), pp. 1263–1292. DOI: https://doi.org/10.1002/ijop.13243.

[26] A. Liu, H. He, X. M. Tu, and W. Tang. "On Testing Proportional Odds Assumptions for Proportional Odds Models". In: *General Psychiatry* 36.3 (2023). DOI: https://doi.org/10.1136/gpsych-2023-101048.

[27] J. Xu, S. G. Bauldry, and A. S. Fullerton. "Bayesian Approaches to Assessing the Parallel Lines Assumption in Cumulative Ordered Logit Models". In: *Sociological Methods & Research* 51.2 (2022), pp. 667–698. DOI: https://doi.org/10.1177/0049124119882461.

[28] A. Agresti and C. Tarantola. "Simple Ways to Interpret Effects in Modeling Ordinal Categorical Data". In: *Statistica Neerlandica* 72.3 (2018), pp. 210–223. DOI: https://doi.org/10.1111/stan.12130.

[29] G. Tutz and M. Berger. "Separating Location and Dispersion in Ordinal Regression Models". In: *Econometrics and Statistics* 2 (2017), pp. 131–148. DOI: https://doi.org/10.1016/j.ecosta.2016.10.002.

[30] Q. Liu, B. E. Shepherd, C. Li, and F. E. Harrell. "Modeling Continuous Response Variables Using Ordinal Regression". In: *Statistics in Medicine* 36.27 (2017), pp. 4316–4335. DOI: https://doi.org/10.1002/sim.7433.

[31] P. McCullagh. "Regression Models for Ordinal Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980). DOI: https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

[32] N. Lemoine. "Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses". In: *Oikos* 128.7 (2019), pp. 912–928. DOI: https://doi.org/10.1111/oik.05985.

[33] T. J. McKinley, M. Morters, and J. L. N. Wood. "Bayesian Model Choice in Cumulative Link Ordinal Regression Models". In: *Bayesian Analysis* 10.1 (2015). DOI: https://doi.org/10.1214/14-ba884.

[34] R. Neal. "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Chapman & Hall/CRC, 2011, pp. 116–162. DOI: https://doi.org/10.1201/b10905.

[35] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.32.6. 2024. URL: https://mc-stan.org/.

[36] M. D. Homan and A. Gelman. "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *The Journal of Machine Learning Research* 15.1 (2014). URL: https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf.

[37] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)". In: *Bayesian Analysis* 16.2 (2021), pp. 667–718. DOI: https://doi.org/10.1214/20-BA1221.

[38] A. Vehtari, A. Gelman, and J. Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.5 (2017), pp. 1413–1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

[39] A. Lahtinen. *Bayesian Modeling of Ordinal Data*. Accessed: 25 May 2025. URL: https://github.com/Allaht2/ordinal_bayes.

[40] 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/.

[41] R. H. B. Christensen. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1. 2023. DOI: https://doi.org/10.32614/CRAN.package.ordinal.

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

[43] R. V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.7. 2025. DOI: https://doi.org/10.32614/CRAN.package.emmeans.

[44] P.-C. Bürkner. "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1 (2017), pp. 1–28. DOI: https://doi.org/10.18637/jss.v080.i01.

[45] A. Behr, E. Webster, M. Brennan, M. Cloonan, and J. Ansell. "Making Live Music Count: The UK Live Music Census". In: *Popular Music and Society* 43.5 (2020), pp. 501–522. DOI: https://doi.org/10.1080/03007766.2019.1627658.

A Additional figures and tables

Scenario	$P(Y=1)$	$P(Y=2)$	$P(Y=3)$	$P(Y=4)$	$P(Y=5)$
Optimal scenario	0.2	0.2	0.2	0.2	0.2
One category with a low response probability	0.05	0.225	0.25	0.25	0.225
Two categories with low response probabilities	0.05	0.05	0.325	0.325	0.25
Hierarchical scenario	0.05	0.05	0.325	0.325	0.25

Table A1: Base probabilities used for the different scenarios in the simulation study.

Scenario	Compilation time
Optimal scenario	34.0s
One category with a low response probability	35.9s
Two categories with low response probabilities	36.7s
Hierarchical scenario	89.5s

Table A2: Compilation time of the Bayesian models in the simulation study.

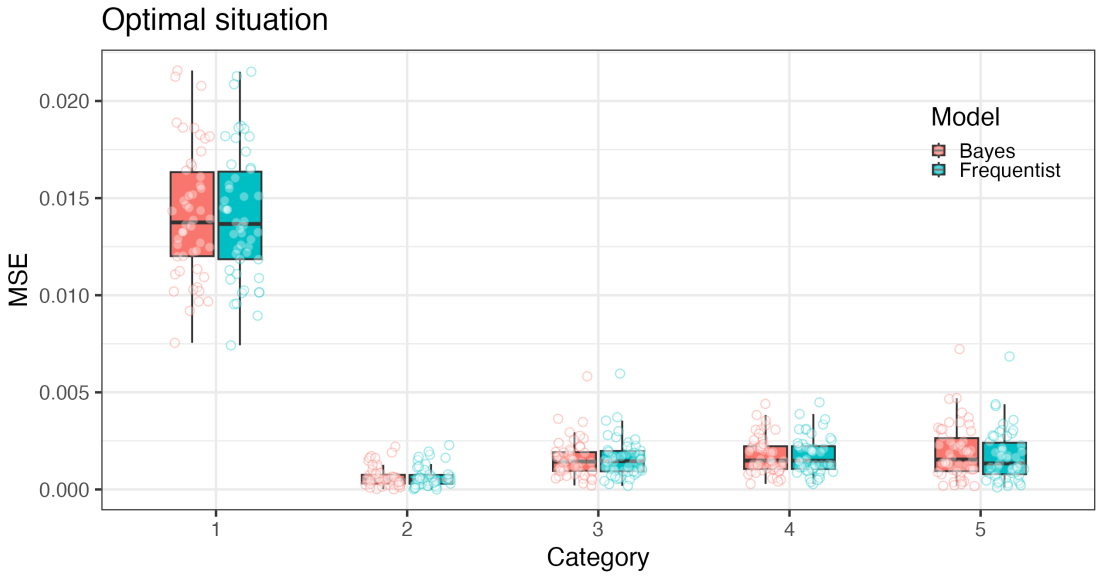

Figure A1: Scenario where all the categories have the same response probability. Both models perform equally well. Box plots of the MSE values for both models across the five response categories in the 50 simulated datasets.

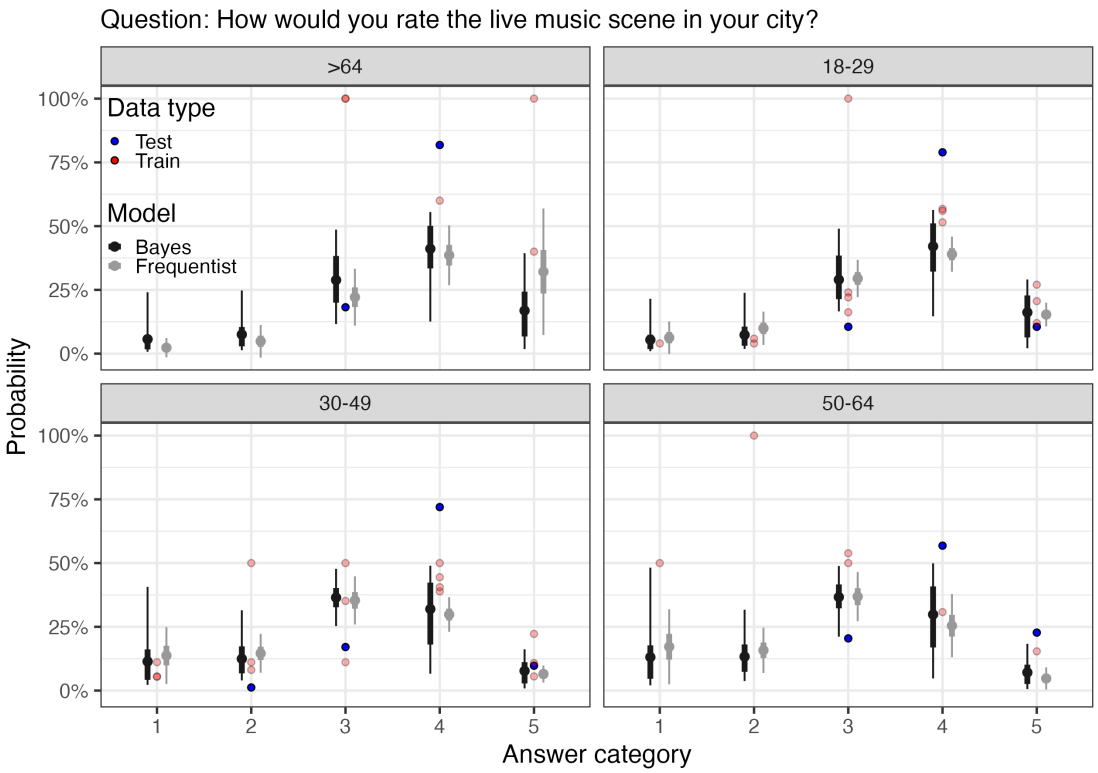

Figure A2: Estimated category probabilities from non-hierarchical models for the "How would you rate the live music scene in your city?" question. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at different levels of the explanatory variable (age group). The red dots indicate estimates from the training set, and blue dots represent estimates from the test set (Helsinki).

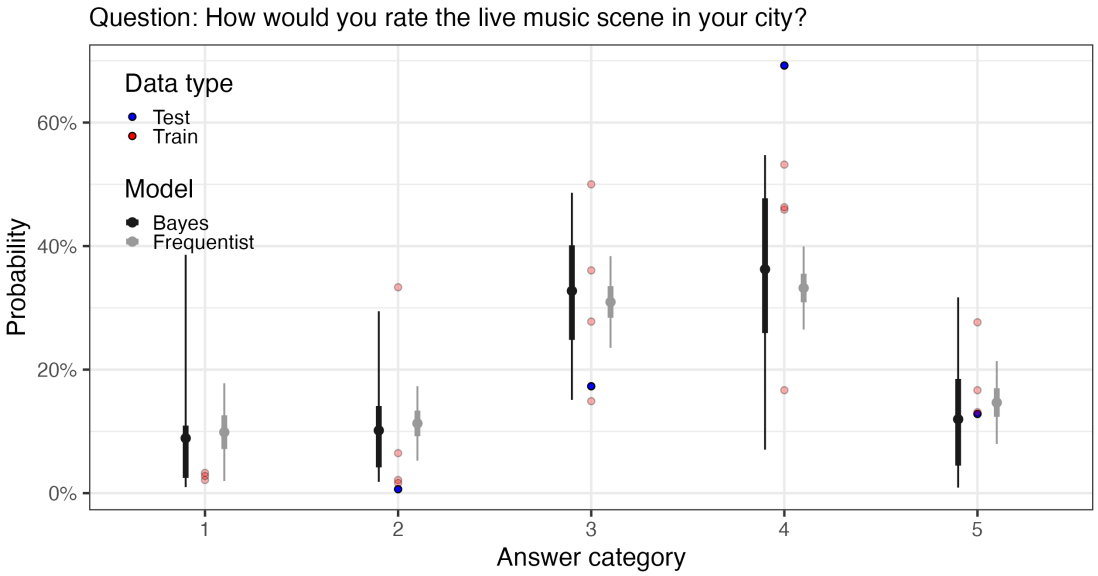

Figure A3: Estimated category probabilities from non-hierarchical models for the "How would you rate the live music scene in your city?" question. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at the aggregate level. The red dots indicate estimates from the set data, and blue dots represent estimates from the test set (Helsinki).

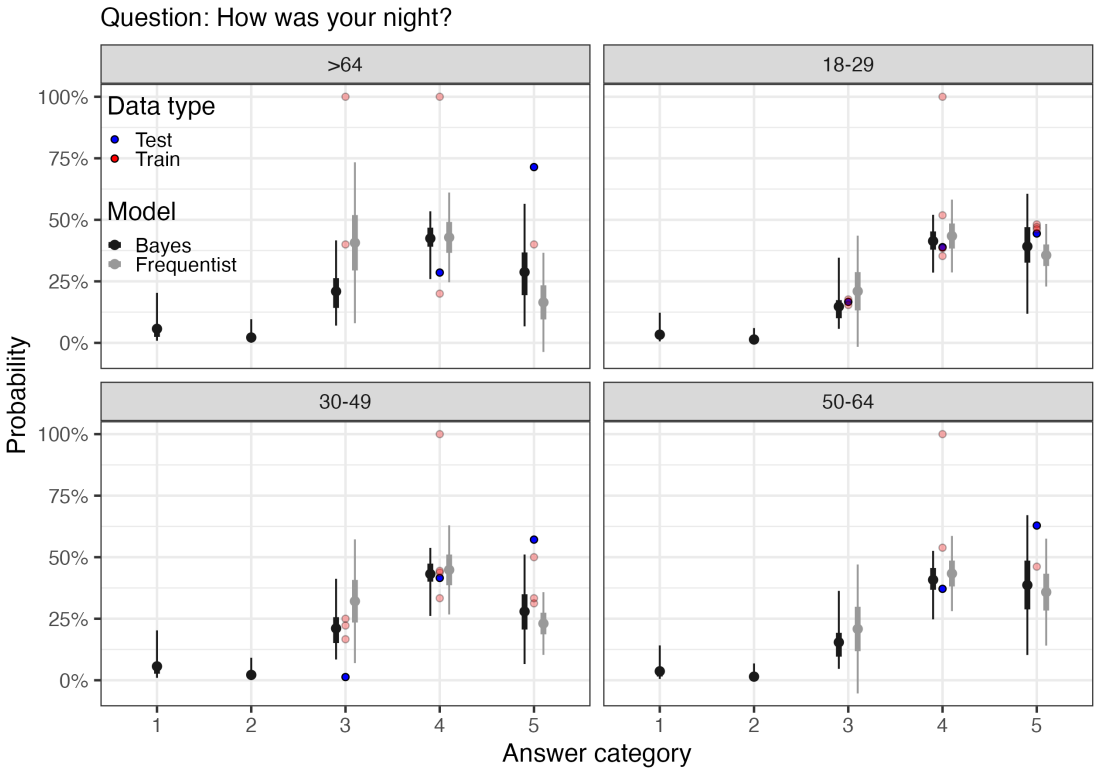

Figure A4: Estimated category probabilities from non-hierarchical models for the "How was your night?" question. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at different levels of the explanatory variable (age group). The red dots indicate estimates from the training set, and blue dots represent estimates from the test set (Helsinki).

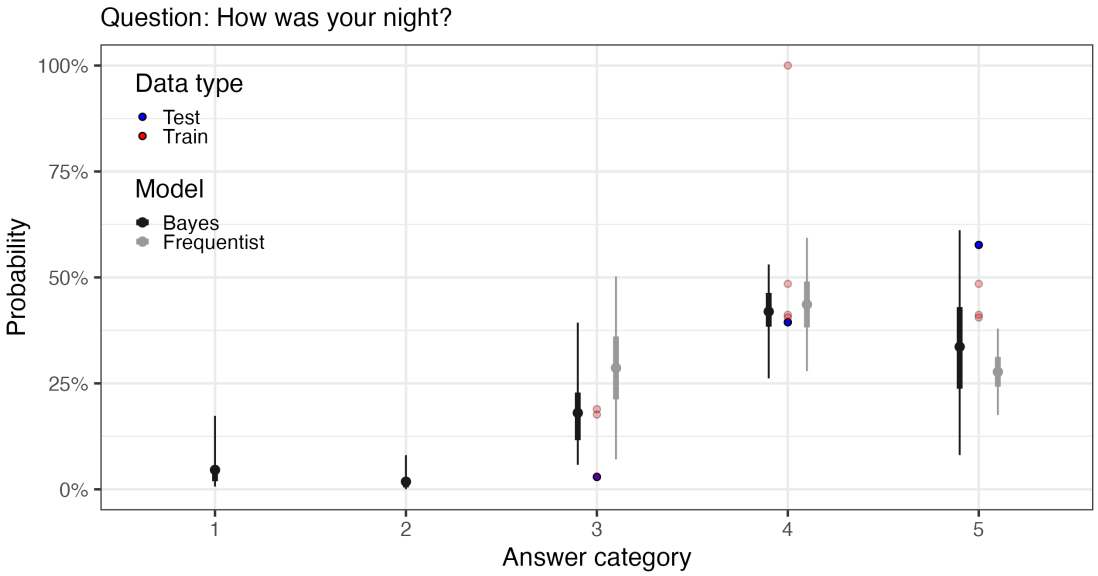

Figure A5: Estimated category probabilities from non-hierarchical models for the "How was your night?" question. The intervals correspond to 50% and 95% credible and confidence intervals from the Bayesian and frequentist models at the aggregate level. The red dots indicate estimates from the training set, and blue dots represent estimates from the test set (Helsinki).