Abdulhakim Alhowaish
January 20, 2025

New Module for Sample Size Estimation
GASTAT Project proposal

This is a proposal for the General authority of Statistics (GASTAT) project about finding a new module for sample size estimation in order to find an outcome with fewer Type I and Type II errors. This proposal suggests using Bayesian Hierarchical models(BHMs) with configurable parameters designed for specific target populations (social, economic, spatial). It starts with background literature. Then, it explores different techniques and applications of BHMs, what to use and when to use them. Eventually, it raises attention to challenges in applying BHMs in general fields that concern GASTAT. A sample case study is provided as a proof of concept in the end.

1 Sample size estimation: Background literature and challenges

Using the population of interest would theoretically eliminate both types of error. However, there are practical constraints such as time, cost, and feasibility that make sampling a realistic and more convincing alternative. In fact, a highly accurate and reliable inference from some samples of a population can be obtained. In the end, it is intuitive to use representatives to represent a larger identical sample.

1.1 Two main pillars

Estimation using sampling theory has two main constituents: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT)[1].

LLN (weak) states that as the sample size n increases, the sample mean $\bar{X}_n$ converges to the expected value μ

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mu$$

CLT states that if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) random variables with mean μ and variance σ^2, then the sampling distribution of the standardized sample mean approaches a standard normal distribution as $n \to \infty$

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

where:

- $X_1, X_2, \ldots, X_n$ are iid random variables with expected value μ
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean,
- μ is the population mean,
- σ is the standard deviation,
- $\xrightarrow{d}$ denotes convergence in distribution,
- $\mathcal{N}(0,1)$ is the standard normal distribution.

1.2 Type I and Type II errors

In detection theory, Type I and Type II errors describe the possible errors that can occur when making a decision based on a statistical test. Referring to hypothesis testing, the two types of error can be summarized in the following table.

Decision / Truth	H_0 **True**	H_0 **False**
Reject H_0	Type I Error (False positive)	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error (False negative)

1.3 Relation to Type I and Type II errors

Consider the scenario of tossing a coin. The coin is fair if $p = 0.5$

LLN (weak) states that as you increase the number of coin tosses (n), the sample proportion of heads ($\hat{p}$) will converge to the true probability p.

$$\bar{p} = \frac{1}{n} \sum heads \xrightarrow{n \to \infty} p$$

The larger the sample size, the lower the probability of failure to detect a biased coin (to overcome **Type II Error**).

CLT explains how the sample mean of heads ($\hat{p}$) is distributed. For large n, the sampling distribution of $\hat{p}$ approaches a normal distribution.

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Proper use of the normal approximation (via CLT) ensures the significance level is maintained (to overcome **Type I Error**).

By carefully designing the experiment and selecting a representative sample, we can reliably estimate the targeted parameters and assess potential risks, balancing precision with practical feasibility. This allows you to calculate confidence intervals and perform hypothesis tests about p.

2 Bayesian Hierarchical Models

BHMs are widely used in many fields to estimate parameters. Especially those that require complex data structures with multiple levels of variability and uncertainty. Excluding basic statistical models, most Bayesian models are hierarchical. Many complex systems and processes can be approximated using multiple layers of simpler models. With the sufficient amount of data samples and the assumption of having independent identically distributed random variables, LLN and CLT ensure that most parameters converge to population mean and most parameter distributions converge to normal distribution. Bayesian statistics benefit from prior knowledge about the environment and nature of the observed datasets, unlike the frequentest statistics. Let us assign a random variable X to the observed data samples and θ to the knowledge we know about the environment and nature of the data samples. Bayesian models are built around knowing three main distributions: prior $p(\theta)$, likelihood $p(\mathbf{x}|\theta)$ and posterior $p(\theta|\mathbf{x})$ distributions.

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{p(\mathbf{x})}$$

BHMs are when we go further step by assuming a distribution of the prior of the prior. Also it can be connecting different Bayesian distributions to become one large distribution (like in the case of mixture models). This process can have its own advantages and disadvantages. However, it has been examined in fields that are related to GASTAT project and proven to have satisfiable performance.

2.1 Applications and Techniques

From spatial and spatiotemporal modeling to health and epidemiology. Psychology, social science, marketing, finance, biostatics, and many other disciplines can represent data through different techniques of BHMs. The next section illustrate some examples of BHMs techniques that are related to GASTAT project.

2.1.1 Mixed effects and mixture models

While both mixed effects models and mixture models deal with variability in data, they have distinct goals, assumptions, and applications. Mixed effects models explain variability in the data between fixed - deterministic - and random effects. On the other hand, mixture models deal with random subgroups that have distinct distributions due to heterogeneity, which are called latent subgroups or hidden clusters. Data that intrinsically have mixed effects and latent subgroups can be modeled with a mixture of the two models. Mathematical formulation for all three models as follows[2].

Mixed effects model:

$$y_{ij} = X_{ij}\beta + Z_{ij}u_j + \epsilon_{ij}$$

Mixture model:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x \mid \theta_k),$$

Mixture of mixed effects model:

$$y_{ij} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(X_{ij}\beta_k + Z_{ij}u_{jk}, \sigma_k^2),$$

where:

- y_{ij}: The response variable for observation i in group j.
- $X_{ij}\beta$: The fixed effects part (β fixed effect coefficients).
- $Z_{ij}u_j$: The random effects part (u_j are group-specific random effects, assumed to follow a normal distribution: $u_j \sim N(0, \sigma_u^2)$).
- ϵ_{ij}: The residual error, assumed to follow a normal distribution.
- $f(x)$: The overall probability density function of the data.
- K: The number of components (subgroups).
- π_k: The mixing proportion for the component k, with $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$.
- $f_k(x \mid \theta_k)$: The probability density function of the k-th component, parameterized by θ_k.
- σ_k^2: Variance of residual error in subgroup k.

To illustrate the use of both models in the GASTAT project paradigm, the mixed effects models can be used in survey responses while accounting for variability across regions or demographic groups. Similarly, mixture models can be used in population segmentation such as income distributions or housing affordability.

2.1.2 Generalized linear models (GLMs)

Generalized Linear Models (GLMs) are an extension of ordinary linear regression[[3]. Unlike traditional linear models, GLMs use a link function $\mathbf{g}$ (e.g. logit, identity) to connect the response variable $\mathbf{Y}$ (i.e. randomly distributed and usually follows an exponential family, e.g. normal, binomial, Poisson) to a linear predictor $\boldsymbol{\eta}$.

$$\begin{aligned} g(\mu_i) = \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \\ \mu_i &= \mathbb{E}[\mathbf{Y}_i] \end{aligned}$$

GLMs are flexible and can map a variety of datasets, from modeling the probability of disease prevalence (binary) to counting visits to a certain hospital during chosen period (Poisson).

2.1.3 Neural networks (NNs)

In the Bayesian Hierarchical Model (BHM) framework, NNs can be treated as flexible prior or likelihood components, allowing probabilistic reasoning and uncertainty quantification in predictions. BHMs extend neural networks by incorporating hierarchical structures, allowing NNs to model dependencies across groups or levels of data. Using spatial data, urban growth can be predicted using NNs.

2.1.4 Nonparametric methods

Nonparametric methods in the Bayesian framework allow flexible modeling without specifying fixed parametric forms. These methods are particularly useful for complex and high-dimensional data with unknown underlying distributions. Gaussian processes (GP) and Dirichlet Process Mixture Models (DPMMs) are famous nonparametric methods. To mention some uses related to GASTAT applications, GPs are applied for spatio-temporal modeling, DPMMs are used in clustering regions with similar social activity, and mainly any distribution-free relationships from survey responses[4].

2.2 Challenges

BHMs have many advantages that make them strong candidates for sampling and estimation. They are flexible, which makes them incorporate multiple levels of variability and uncertainty. Also, with the aid of Monte Carlo Markov Chains (MCMC), they can generate synthesis samples that have a similar behavior to improve estimates for smaller groups with limited data. Furthermore, they are interpretable and naturally provide uncertainty estimates for all parameters. However, there are some real challenges that we need to be aware of while using them. 'Sampling is distribution dependent, and varies greatly within the same class'. Meaning, that the main body of the distribution may converge to Gaussian within some rate; but the remote parts do not. Depending on the distribution, remote parts determine many crucial features of the inference. Another challenge is how fast the distribution of the data converges to the Gaussian distribution. Some complex distributions (i.e. power laws, subexponential, and heavy-tail distributions) take much time to converge, and some may never converge. So the nature of the data and picking the model of interest plays a huge rule on having the right conclusions. [5]

2.3 Case Study

This is a case study to show how to use a Bayesian Hierarchical model to infer hidden relations about social data. Specifically, it applies the mixture model technique. The data is chosen from a common data repository called **COUNT**[6]. It is implemented using the R language. The case study will be uploaded as a separate PDF file.

3 Conclusions

In this project proposal for the GASTAT project 'New module for sample size estimation', I tried to propose a famous statistical general method, BHMs, to make inferences about targeted data. Then I introduced BHMs and their different techniques. I highlighted what best suits GASTAT related projects. I mentioned challenges on the way to apply BHMs. Finally, a case study is attached as a separate PDF file. This case study is provided as a proof of concept of one of the BHMs techniques on public dataset. The goal of this proposal is to collaborate with the GASTAT team to make an inference about the social, economic, and spatial data collected.

References

[1] George Casella and Roger L. Berger. *Statistical Inference*. 2nd. Duxbury Advanced Series, 2002.

[2] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[3] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd. Chapman and Hall/CRC, 2020.

[4] Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd. Chapman and Hall/CRC, 2014.

[5] Nassim Nicholas Taleb. *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications*. STEM Academic Press, 2020.

[6] R. Fair. "A Theory of Extramarital Affairs". In: *Journal of Political Economy* 86 (1978), pp. 45–61.