Proposal

Tammi T. Tran, Terell Johnson, Kesson Nyankpani

July 2024

1 Proposal

1.1 Overview

In today's digital age, social media and online platforms have revolutionized the way we access information. While this democratization of information has its advantages, it has also led to the widespread dissemination of misinformation across various fields. One critical area affected by this phenomenon is personal finance. Understanding the extent to which young adults are misinformed about financial matters is crucial, as this demographic represents the future of our economy. Our research project aims to quantify the percentage of young adults who hold misconceptions about finances, particularly in the realm of investing and stocks. By leveraging Bayesian statistics and probabilistic programming with PyMC, we plan to create a model that can accurately estimate this probability. Our approach focuses on collecting data from students through a test in the form of a Google document. We plan to aim for the collection of 300 responses to questions about stocks and investing.

The test will consist of basic questions about stocks and investing, designed to assess the knowledge and misconceptions held by young adults. With at least 300 responses it would provide a robust dataset for our analysis. The collected data will be analyzed using Bayesian statistics and probabilistic programming with PyMC. This approach will allow us to create a model that estimates the likelihood of financial misinformation among the respondents. By analyzing the results, we aim to identify patterns and correlations that reveal the influence of social media platforms, news outlets, educational institutions, and peer groups on financial understanding.

Financial literacy is essential for making informed decisions that can lead to economic stability and growth. However, misinformation can lead to poor financial decisions, resulting in negative consequences for individuals and the broader economy. By quantifying the extent of financial misinformation among young adults, this project seeks to highlight the areas where educational interventions are needed most. Ultimately, this project aims to contribute to a more financially literate and informed generation, capable of making sound investment decisions.

1.2 Proposed Data and Analysis Plan

Data Set:
Our data set will consist of the scores obtained from a Google document-based test that we will administer to a targeted group of young adults. This test is designed to evaluate their knowledge and understanding of investing and stock market concepts. It will include basic questions such as whether the respondents are familiar with investing in the S&P 500, among other fundamental financial concepts.

Details of the Data Set:

1. Format: Google document-based test
2. Target Respondents: Young adults (primarily students)
3. Number of Respondents: Aiming for at least 300 responses
4. Content: Questions related to investing, stock market basics, and financial literacy

Why We Are Using This Data Set:
We chose this data set because it directly measures the knowledge and misconceptions about investing among our target demographic. The test scores will provide quantifiable data that can be analyzed to assess the level of financial misinformation. By targeting young adults, we can gain insights into how well this demographic understands key financial concepts and identify areas where misinformation is prevalent.

Bayesian Model and Analysis Plan Using PyMC:
To analyze the collected data, we will employ a Bayesian statistical approach using PyMC. This method allows us to incorporate prior knowledge and update our beliefs based on the observed data, providing a robust framework for probabilistic inference.

Conceptual Steps for Defining the Bayesian Model:

Step 1: Define the Priors Misinformation (MM): We will use a Beta distribution as the prior for the level of misinformation. This choice is appropriate for modeling probabilities because the Beta distribution is defined on the interval [0, 1] and can represent various shapes based on its parameters.

Step 2: Define the Likelihood Test Scores (SS): Given the level of misinformation MM, the test scores will be modeled using a Gamma distribution. The shape parameter $k(M)$ of the Gamma distribution will be a function of the misinformation level, such as $k(M) = k_0(1 - M)$.

Step 3: Define the Posterior Posterior Distribution: By combining the prior (Beta distribution for MM) and the likelihood (Gamma distribution for SS), we can obtain the posterior distribution, which represents our updated beliefs about the level of misinformation given the observed test scores.

Model Components in PyMC:

1. Priors:
$M \sim \text{Beta}(\alpha, \beta)$
Constants for the Gamma distribution: k_0 and θ

2. Likelihood:
$S \mid M \sim \text{Gamma}(k_0(1 - M), \theta)$

Steps in PyMC (Conceptually):

1. Import PyMC:
Import the necessary modules from PyMC.

2. Define Priors:
Define M as a Beta distribution.

3. Define the Likelihood:
Define the Gamma distribution for S using $k(M)$ and θ.

4. Posterior Inference:
Use PyMC's sampling functions to draw samples from the posterior.

Conceptual Steps:

1. **Import Libraries:**
PyMC for Bayesian modeling.

2. **Define Constants:**
Define k_0, θ, α, and β.

3. **Define the Model:**
Use a context manager to define the model.
Specify the Beta distribution for M.
Define $k(M)$ as a deterministic variable.
Specify the Gamma distribution for S with the shape parameter $k(M)$.

4. **Sampling:**
Use Markov Chain Monte Carlo (MCMC) or other sampling methods to obtain the posterior distribution of M.

5. **Posterior Analysis:**
Analyze the samples to summarize the posterior distribution, including the mean and credible intervals.

By following these conceptual steps, we will set up a Bayesian model in PyMC to analyze the relationship between misinformation and test scores. This approach will allow us to quantify the extent of financial misinformation among young adults and identify the most significant sources of misinformation. The results of this analysis will inform targeted educational interventions aimed at improving financial literacy in this demographic.

1.3 Method

Let M denote the level of misinformation, and S denote the test scores. We assume the following model:

1. Prior Distribution

We assume a Beta prior distribution for M:

$$M \sim \text{Beta}(\alpha, \beta)$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters.

2. Likelihood Function

The likelihood function describes the relationship between the test scores S and the level of misinformation M:

$$S \mid M \sim \text{Gamma}(k_0(1 - M), \theta)$$

3. Posterior Distribution

Using Bayes' theorem, the posterior distribution of M given the observed test scores S is:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

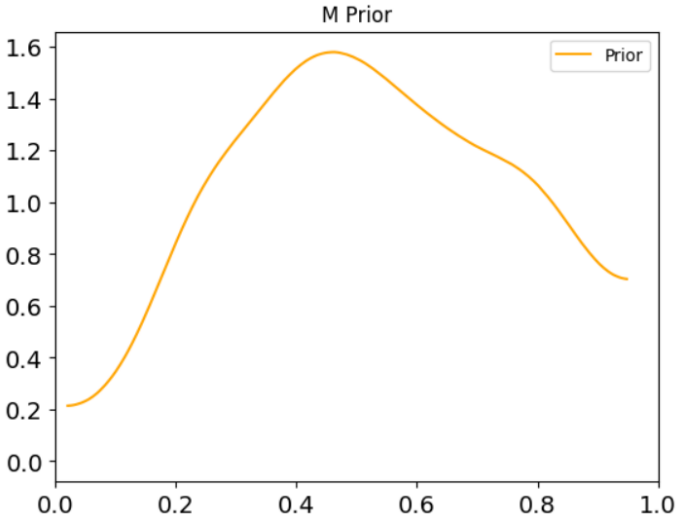
where:

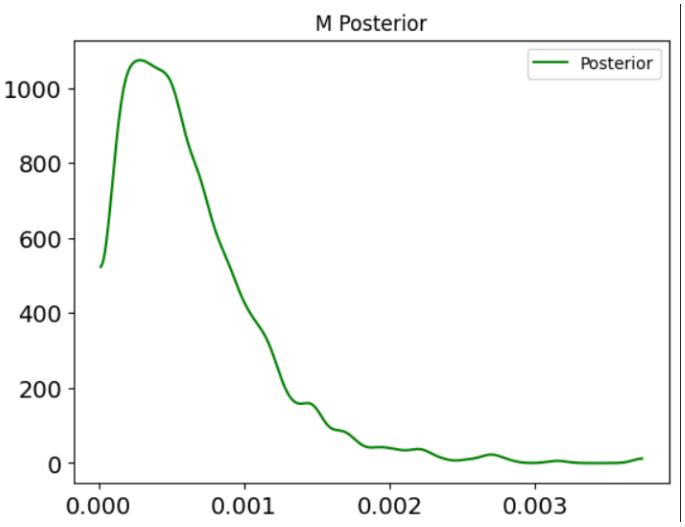
- $P(M)$ is the prior distribution of misinformation.
- $P(S \mid M)$ is the likelihood function.
- $P(S)$ is the marginal likelihood or evidence, computed as:

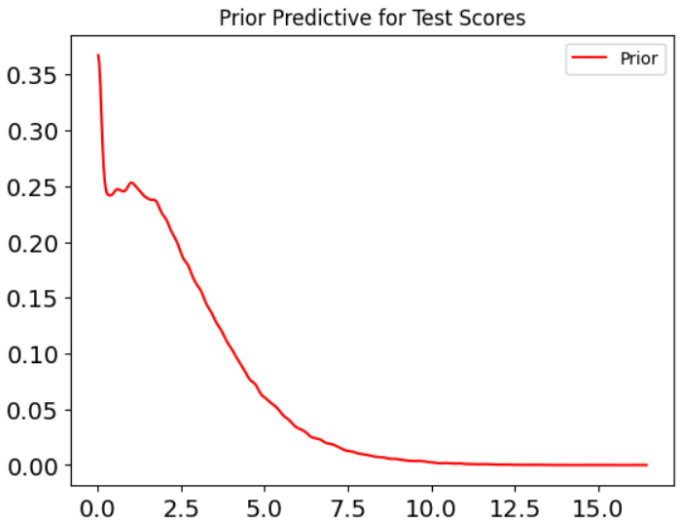
$$P(S) = \int_0^1 P(S \mid M)P(M)\, dM$$

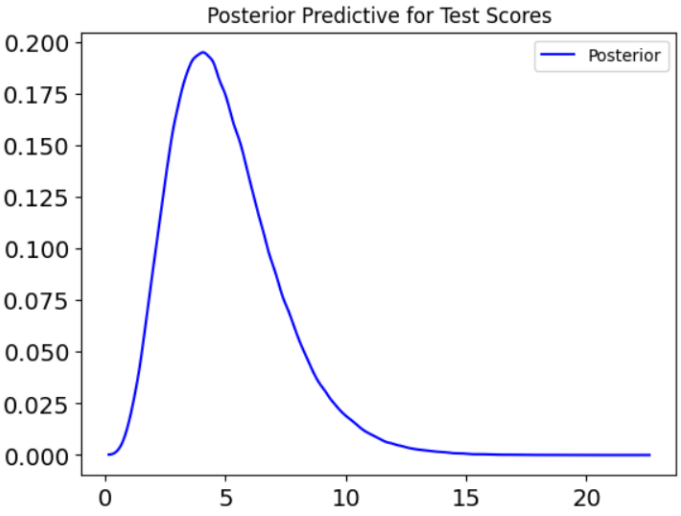
Model Parameters

- α: Shape parameter for the Beta distribution of M
- β: Shape parameter for the Beta distribution of M
- $k_0(1 - M)$: Shape parameter for the Gamma distribution of $S|M$
- θ: Rate parameter for the Gamma distribution of $S|M$


M Prior
Prior
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


M Posterior
Posterior
1000
800
600
400
200
0
0.000
0.001
0.002
0.003


Prior Predictive for Test Scores
Prior
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0


Posterior Predictive for Test Scores
Posterior
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20

Results

1. M Prior Distribution

- **Graph Description:** The first graph shows the prior distribution for the parameter M, representing our initial belief about the probability of misinformation before observing any data. The distribution appears to be a Beta distribution centered around 0.5, indicating that we initially believed the probability of misinformation was equally likely across the interval [0, 1].

2. M Posterior Distribution

- **Graph Description:** The second graph displays the posterior distribution for the parameter M, representing our updated belief about the probability of misinformation after observing the test scores. The distribution is much more concentrated and peaks around a very low value, indicating that the observed data has led to a strong belief that the probability of misinformation is quite low.

3. Prior Predictive Distribution for Test Scores

- **Graph Description:** The third graph shows the prior predictive distribution for test scores, which is generated based on the prior beliefs about M. The distribution, modeled using a Gamma distribution, covers a wide range of scores, indicating high uncertainty about the test scores before observing the actual data.

4. Posterior Predictive Distribution for Test Scores

- **Graph Description:** The fourth graph depicts the posterior predictive distribution for test scores, reflecting the expected distribution of test scores after updating our beliefs with the observed data. The distribution is more concentrated and peaks around a certain score range.

1. M Prior Interpretation

The prior distribution for the parameter M represents our initial belief about the probability of a certain event (e.g., the probability of misinformation) before observing any data. The Beta distribution with parameters $\alpha = 2$ and $\beta = 2$ is used, which gives a symmetric distribution centered around 0.5. This means we initially believe that the probability of misinformation is equally likely to be any value between 0 and 1, with no strong preference towards either extreme.

2. M Posterior Interpretation

The posterior distribution for the parameter M represents our updated belief about the probability of misinformation after observing the test scores. The Beta distribution parameters have been updated based on the observed data, resulting in a new distribution that reflects the influence of the data. The posterior distribution is more concentrated around a specific value compared to the prior, indicating increased certainty about the probability of misinformation after considering the data.

3. Prior Predictive for Test Scores Interpretation

The prior predictive distribution represents the expected distribution of test scores based on our prior beliefs about M (before seeing the actual data). Here, the test scores are modeled using a Gamma distribution, and the prior predictive distribution shows the range and likelihood of test scores we would expect to see if our prior beliefs about M were true. The distribution is fairly wide, indicating a high level of uncertainty about the test scores before seeing the data.

4. Posterior Predictive for Test Scores Interpretation

The posterior predictive distribution represents the expected distribution of test scores based on our posterior beliefs about M (after seeing the actual data). This distribution is also modeled using a Gamma distribution, but it has been updated to reflect the influence of the observed data. The posterior predictive distribution is more concentrated and possibly shifted compared to the prior predictive distribution, indicating that our expectations about the test scores have been refined based on the data.

Overall Analysis

Prior vs. Posterior for M: The prior distribution for M shows our initial, broad belief about the probability of misinformation. After observing the data,

the posterior distribution for M shows a more concentrated belief, indicating increased certainty about this probability.

Prior Predictive vs. Posterior Predictive for Test Scores: The prior predictive distribution shows a wide range of expected test scores based on our initial beliefs. After observing the data, the posterior predictive distribution is more focused, showing a refined range of expected test scores based on the data.

In summary, these graphs demonstrate how Bayesian updating works: starting with broad initial beliefs (priors), incorporating observed data, and resulting in more refined beliefs (posteriors). The prior predictive and posterior predictive distributions provide a way to visualize and compare our expectations before and after observing the data.