Mortality Mysteries – How Are Young People Dying in America?

*Bayesian Analysis of Annual Causes of Death in the US for Ages 25-34*

# Introduction

This report analyzes the top 5 annual causes of death in the United States for individuals aged 25-34, over a five-year period from 2018-2022. The analysis aims to provide insights into the patterns of death, and how well a hypothetical 6th cause of death could be predicted. We use Bayesian statistical methods (explained later) for the analysis.

# Summary of Findings

* Accidents are the leading cause of death for 25–34-year-olds in the US. The estimated mean is 24,894 annual deaths, significantly higher than other causes.
* Suicide and homicide are the next most significant causes, followed by heart disease and malignant neoplasms, though there is considerable uncertainty in these estimates.
* The analysis reveals uncertainty in the estimates. This is likely due to the limited dataset of five years.
* The hierarchical model (explained later) provides the most balanced view. It estimates an overall mean of 11,873 annual deaths across all causes for this age group.

# Methodology

If you are not familiar with Bayesian inference, it’s a method of updating our beliefs about the number of deaths from each cause. We start with some initial guesses (*priors*) and then use the data being analyzed to update these guesses. It’s like guessing the number of M&Ms in a jar. We make an informed decision based on the jar’s size (our prior belief). As we get more info – like comparing it to other jars we’ve seen or how many bags of M&Ms we think we could cram in there, we adjust our guess.

For this analysis we chose our priors from the known causes of death of U.S. citizens aged 24-35 from the years 1999-2017[[1]](#endnote-1). We chose an informed prior where we had some historical information available, to potentially lead to more accurate estimates.

For our analysis we are using three types of models to make the predictions:

*separate*, *pooled*, and *hierarchical*.

* The separate model makes individual guesses for each cause of death.
* The pooled model makes one guess for all causes of death.
* The hierarchical model is a hybrid that balances between these approaches.

By running many simulations (*samples*), we get a range of plausible values for each cause. This helps us understand the typical number of deaths and how uncertain we are about the estimates.

For those familiar with Bayesian Inference: we use a standard Bayesian workflow. A custom Gibbs sampler is used. The use of the multiple models allows a comparison across the models and highlights the differences between pooling and separating the data.

The full code implementation and modeling can be found in the accompanying Jupyter notebook.

# Model Approach

Each model represents a different approach to understanding the underlying patterns in the data:

**Separate Model**

* Approach: Treats each cause of death independently.
* Prior: We start with a guess (noninformative prior) about both the average number of deaths (mean) for each cause and how much that number might vary (standard deviation) from year to year, allowing the actual data to strongly shape our final estimates.
* Posterior: After analyzing the data, we create separate updated estimates (posterior) for each cause of death. This shows how likely different numbers of annual deaths are for each cause of death.

**Pooled Model**

* Approach: Assumes all causes share the same underlying distribution.
* Prior: Similar to the separate model, but with a single mean and standard deviation for all causes.
* Posterior: Produces a single posterior distribution representing all causes.

**Hierarchical Model**

* Approach: Balances between separate and pooled approaches.
* Implementation: Uses a custom Gibbs sampler.
* Posterior: Generates posterior distributions for each cause, influenced by both individual data and the overall distribution.

**Hypothetical 6th Cause**

To predict the 6th cause of death we use posterior samples from the existing models. By using the posterior means and standard deviations of the five existing causes, we incorporate the patterns and trends we have already observed into the data. This provides a more reasonable estimate for the hypothetical cause, since no direct data is available.

The informed priors help constrain the estimates to something we are likely to see. For example, if the mean number of deaths is in an existing range, using that as a starting point helps ensure our estimate isn’t way outside what we’ve seen before.

The analysis includes *posterior distribution* plots for visualizing the results.

**Posterior Distribution**

Using Bayesian inference, we update our prior (original) beliefs with the observed data. This process updates our beliefs based on the new information.

The result is a *posterior distribution*, which gives us a range of likely outcomes.

To understand the accuracy of our estimates, we use credibility intervals to express the uncertainty in our results.

*Note: A credibility interval gives us a range of values within which we believe the true average lies, with 94% certainty. Using a 94% credibility interval is standard practice. It provides a good balance between precision and practicality, capturing the most likely values while excluding extreme outliers. It is widely accepted and easy to interpret, whereas using 100% would include all possible values, making the interval too wide and less informative. This approach ensures that the results are both reliable and useful for decision-making.*

# Separate Model

The separate model gives a detailed view of each cause of death individually. This is valuable if we want to make predictions about specific causes in isolation.

It’s like a having a different expert focusing on each cause, rather than a single generalized expert looking at all of the causes. By having an expert look at each cause independently, they don’t influence each other. If one expert looked at all areas, there might be some influence that could distort the results.

Figure 1 shows the posterior distributions of the mean for each cause of death individually.

A graph of a normal distribution

Description automatically generated with medium confidence

A graph of a normal distribution

Description automatically generatedA diagram of a normal heart disease

Description automatically generated

A diagram of a normal distribution

Description automatically generated

**Figure 1:** The posterior distribution of the mean of annual deaths of the 5 causes

Our separate model reveals for each of the causes the mean most likely to be seen is as follows:

A screenshot of a phone

Description automatically generated

A red line graph on a white background

Description automatically generated

**Figure 2:** The predictive distribution for a hypothetical (6th) cause of death for the separate model.

The posterior of the mean of the annual deaths for the hypothetical cause for the separate model is 10,361 (95% CI: (2101, 30768))

Figure 2 shows the updated belief about the possible values for the mean annual deaths of our hypothetical 6th cause. It’s based on our prior knowledge and the data from the five other causes.

This is a bimodal distribution, with two distinct peaks. This suggests there are two possibilities for the 6th cause – it’s more like the less common causes of death (left peak), or it’s more similar to the more common causes of death (right peak).

Because the peaks are wide, we are not certain about the estimate. It could be anywhere from 0 to 40,000 annual deaths. It’s most likely to fall into one of the two areas.

What we infer from the model is, based on the other five causes, when considering them separately, if there were a sixth cause the death rate would likely be in one of these two areas.

# Pooled Model

In the pooled model all the data is treated as one big distribution, ignoring that there were different causes to begin with. There is no distinction between causes.

The pooled model doesn’t analyze each cause separately. This can work well to get an overall picture. The downside is it may miss specific aspects of individual causes, losing the distinctions between them.

A single distribution (large sample) with overall mean and variance is used to estimate the posterior distribution and make a prediction.

A diagram of a normal distribution

Description automatically generated

**Figure 3:** The predictive distribution for a hypothetical (6th) cause of death for the pooled model (red-dotted), layered with the posterior distribution of the mean of annual deaths of the five other causes (blue-solid).

The blue line represents the known causes. The narrowness suggests certainty about this estimate. This makes sense, since we can visually look at the small set of input data and get an idea of what the range of values could be without doing any math.

The red dashed line is the hypothetical 6th cause. The red line is much wider and flatter than the blue line, indicating more uncertainty. The peak is roughly aligned with the blue line, suggesting the model predicts this new cause might be similar in magnitude to the average of known causes.

The posterior of the mean of the annual deaths for the hypothetical cause for the pooled model is 10,326 (95% CI: (-9310, 29967)). The wide spread of the hypothetical distribution indicates the model allows for a broad range of possibilities, from very few deaths to potentially more than any known cause.

Our CI is well below zero. This indicates we aren’t certain about the estimates and wouldn’t put much confidence in it.

Compared to the separate model which showed two distinct possibilities, the pooled model suggests a single range which is quite broad. This approach assumes there is a relationship between the different causes of death.

The prediction of the hypothetical cause is heavily influenced by the average trend of the combined causes. This is a feature of the pooled model and might be beneficial if we believed there to be underlying similarities across all the causes of death.

# Hierarchical Model

Hierarchical models are designed to capture complex relationships and variations within and between the causes of death. They recognize that not only are there multiple causes of death, but the models assume the causes are related in some way.

In our model there are individual cause of death parameters, such as accidents might have an average of 30,000 deaths per year. Each cause has its own mean and variance.

Then the population parameters describe the overall patterns found in the causes of death. Such as – the average number of deaths per year, and how much each cause varies from the average.

These models allow for *partial pooling* of information across the causes, which can lead to more nuanced and sometimes multimodal distributions.

The hierarchical model uses Gibbs sampling, which iteratively samples from the conditional distributions of each parameter.

A graph of a number of individuals

Description automatically generated with medium confidence

**Figure 4:** The posterior distribution of the mean of annual deaths of the 5 causes for the hierarchical model, layered with the predictive distribution for a hypothetical 6th cause of death.

A screenshot of a phone

Description automatically generated

The posterior of the mean of the annual deaths for the hypothetical cause for the hierarchical model is 11,880 (95% CI: (10488.11, 13464.09)

This posterior distribution reveals some interesting patterns:

* Some causes (homicide, heart disease, malignant neoplasms), show distributions with multiple peaks. Since the data spans years, it’s likely there were periods with different average death rates.
* The hypothetical cause (red dashed line) shows a multimodal distribution that doesn't align closely with any single existing cause. This could represent a complex new cause of death with characteristics shared across multiple existing causes.
* Accidents has a very wide, flat distribution. This can indicate high variability or uncertainty.
* Suicide has a very narrow, and peaked distribution. This suggests more predictable rates.
* Some distributions (accidents) are right-skewed. This indicates occasional years or events with higher deaths that pulled the distribution to the right.

**Hypothetical 6th cause**

The observations around the 6th cause in the hierarchical model are:

* The hypothetical 6th cause is bimodal (multiple peaks), with a mean of 11,880. This suggests the model predicts two possible patterns for this cause.
* The range is between 10,488 and 13,464 annual deaths. This suggests it’s a moderate cause of death.
* The wider distribution compared to other causes indicates the uncertainty (wider spread) of this cause compared to some others.
* The magnitude is predicted to be greater than homicide and suicide, but less than other causes.

The prediction reveals the complex interactions between the different causes and the data.

# Suitability of Methods

The Bayesian methods employed are well-suited for this type of analysis for several reasons:

Handling of uncertainty – Bayesian methods excel when uncertainty is involved. This is a crucial feature given the small data size.

Use of priors – Prior knowledge is helpful when dealing with health-related studies. We used historical data for our priors and a common variance to lead to more reliable estimates.

Hierarchical modeling – We can balance information about the individual causes of death as well as overall patterns across all the causes of death. This leads to more reliable estimates, especially when we have small data sets.

However, the suitability is somewhat limited by the small sample size. This is shown by the wide credible intervals and potentially unreliable estimates. Five years of data may not do the best job of capturing long-term trends or patterns.

The size of the data is also sensitive to outliers. For example, between 2019 and 2020 there was a large spike in accidents and homicides. Are these due to random chance or are they tied to an event like Covid? More data would help answer this question.

The choice of priors in the models can influence the shape of posterior distributions, sometimes leading to multimodality or skewness.

The original data is organized by decreasing mortality rates, in descending order per cause. This relationship was not picked up in the hierarchical model, which estimated a hypothetical 6th cause to have a mean of 11,873. If we consider a 6th cause should have a lower average than the 5th cause of death, the mean should have been lower than the 5th cause. The hierarchical model we used doesn't inherently understand or preserve this descending order relationship. It treats each cause as a related but separate entity without assuming any specific ordering between them. If ordinality were important, we would need to revisit the analysis.

# Conclusions

Accidents are the leading cause of death across all models. This category has a significantly higher mean than other categories. The hierarchical model estimates a mean of 11,880 annual deaths from accidents.

After accidents the mean cause of death drops considerably.

The wide credibility intervals indicate there is considerable uncertainty in the estimates. For example, the 95% CI for accidents in the hierarchical model is 15,589 to 29,544.

The estimates in the hierarchical model are a balance between the extremes of a fully pooled model and all separate models.

The hierarchical model is generally the most trustworthy. This model provides the best representation since it includes both the differences in individual causes and the overall trends of the data.

However, because of the uncertainty mentioned earlier, the findings should be compared against other sources of information, or the models recomputed with larger datasets.

Future research would benefit from refining the hierarchical model to better capture relationships within the data. Some potential areas to include are adding additional data, refining the priors, and considering the ordinality of the data.

# Appendix

The statistical program used for analysis of the separate and pooled models was PyMC. Credibility intervals were computed using the ArviZ coding library. This information is provided for those interested in reproducing the results or understanding the technical details of the analysis.

Each model parameter estimated was monitored to ensure that convergence was established for the posterior distributions. The statistical tool Markov Chain Monte Carlo (MCMC) was used to help monitor and evaluate chain convergence in the separate and pooled models.

Our custom Gibbs sampler was implemented using Python and the NumPy library. The sampler runs for 10,000 iterations with a burn-in period of 1,000 iterations.

Visual inspection of trace plots was also used to detect non-convergence, as can be found in the accompanying notebook. Look for the 'fuzzy caterpillars' in the trace plots!

# References

1. CDC Underlying Causes of Death <https://wonder.cdc.gov/ucd-icd10.html> [↑](#endnote-ref-1)