# Initial population

The SHIELD model begins in 1940. However, since census data on age, race, and sex is only available back to 2010, we need to select a baseline year before 2010 to set the population composition. Without data to inform earlier years, modeling dynamic changes would not add value here.

**Fixed Strata Year**: Birth data begins in 2007, and deaths are reported in two time frames, modeled as static parameters. This makes 2007 a suitable year for setting initial strata sizes. Thus, we start the model in 1940 but keep the population size and demographic composition fixed at 2007 levels until we introduce dynamic changes over time.

**Initial Population**: We model three subgroups based on the intersection of sex and HIV acquisition risk, including females, heterosexual males, and males who have sex with males (MSM). Census data provides the population sizes for females and males. To further segment the male population, we estimate the proportion of males who are MSM.

**Determining MSM Population Size**:

**Proportion of MSM in the Male Population**: We have two key data sources for estimating the MSM population size:

* **County-Level Data from Emory**: This provides estimates of the total proportion of males who are MSM, available at the county level in the Surveillance Manager. The basic approach is to compute a weighted average of these county-level proportions based on each county’s population size. This would give us an MSA-level estimate that we could apply uniformly across age and race groups. However, this method overlooks potential racial heterogeneity.
* **MSA-Level Data from BRFSS**: The Behavioral Risk Factor Surveillance System (BRFSS) provides race-specific proportions of males reporting same-sex behavior at the MSA level.

**Race-Specific MSM Risk Adjustment:** Using the BRFSS data, we first apply race-specific MSM proportions to the male population by race in each county to estimate an initial count of MSM males.

We then compare this initial estimate to the county-level MSM proportion from Emory and adjust the race-specific proportions to better fit the Emory data. This method ensures that the final estimates align with both the overall MSM proportion from Emory and the racial distribution from BRFSS, capturing both racial heterogeneity and county-level variation.

**Adjusting Race Categories**: There's a slight complication in using BRFSS data because the race categories differ from those in the Census. Here’s how we handle the discrepancies:

* **Flattening Race and Ethnicity**: We combine race and ethnicity into a single race category, resulting in five groups:

1. Hispanic
2. American Indian or Alaska Native
3. Asian or Pacific Islander
4. Black
5. White

This structure aligns more closely with Census data, where Hispanic is treated as an ethnicity, and we create a single "Asian or Pacific Islander" category to match the available Census groupings.

* **Combining Asian & Pacific Islander Groups**: In the BRFSS, Asians and Native Hawaiians/Other Pacific Islanders are reported separately. Instead of taking a simple average, which assumes equal population sizes, we apply a weighted combination based on an expected population ratio of 9:1 (Asian to Pacific Islander). This gives:

*’Asian or Pacific Islander’=0.9×(proportion MSM for Asian)+0.1×(proportion MSM for Native Hawaiian/Other Pacific Islander)*

This approach provides a more accurate reflection of the population composition in the "Asian or Pacific Islander" group.

# DEMOGRAPHIC MODELING

We model births + immigration into and deaths + emigration out of each “compartment”

* Migration (immigration and emigration) doesn’t depend on disease state, so it doesn’t change the disease prevalence in each compartment

## Births:

Birth rate= number of births/ population size

Fertility rate= number of births/number of female in childbearing ages

* JHEEM has used “birth rate” (although it’s called fertility rate in the code).
* SHIELD will use “fertility rate” to capture vertical transmissions more accurately

## Deaths

There are two components: general mortality and disease-specific mortality rates

Jheem uses state specific death rates, called metro-deaths, these are not reported directly.

To estimate the national-level or MSA-level mortality rate, we extract the counties within each MSA, map these counties to their corresponding states, and then take a weighted average of the state-level death rates to approximate the MSA-level mortality rate

Steps:

1. Extract county level death rates
2. Extract county level population count
3. Compute county weights as proportion of their population relative to the state or national population
4. Compute weighted average death rate

## Aging

In a compartmental model, the aging process represents individuals transitioning from one age group to the next as they age. The rate at which individuals "age out" of one compartment and enter the next is governed by the aging rate, which is often calculated as the inverse of the duration individuals typically remain in that age group. When individuals in a particular compartment reach the upper age limit, they transition to the next age group at this rate. For example, for each age group i (e.g., 0-15), the aging rate ai​ can be calculated by:

ai=1/duration in age group > 1/15 per year

**Issue**: Fix aging rates don’t allow for the calibration of models to historical or future trends in aging. These models don’t account for shifts in life expectancy, fertility, or mortality, and could result in inaccurate projections of the age distribution.

**Solution:** To address the limitations of fixed aging rates, you can incorporate **differential aging rates** that **vary by age group and time**

### Time-Dependent Aging Rates:

Define aging rates ai​(t) for each age group that can change over time. Rather than assuming a constant rate, aging rates can be modeled as functions of time to reflect evolving demographic factors like healthcare advancements or lifestyle changes.

1. **Number of parameters**? We need to determine the number of parameters required to model differential aging rates effectively. In our model:
   1. We have 11 age groups, requiring 10 independent aging rates. The model also considers 3 races and 3 sexes, totaling 90 parameters if we include interactions.
2. **Dynamic Aging Rate Modeling?** Should changes be linear, or should we use a spline-based approach? Generally, we use splines with knots in 2010 and 2020:
   1. 2010 is our base year for population calibration.
   2. 2020 serves as a recent benchmark to fine-tune projections for the baseline year (2024).
3. **Differentiating by Infection Status?** Whether to differentiate aging rates for infected versus uninfected populations depends on data availability.
   1. Uninfected Population: Census data on population age structure helps us calibrate aging among uninfected individuals.
   2. HIV-Positive Population: Estimated HIV prevalence by age allows us to capture aging patterns among the infected population.
   3. Syphilis: No prevalence estimates are available, so we lack the data to differentiate aging rates for syphilis.

**Aging Rates in Shield Model**: In the Shield model, we do not differentiate aging rates by infection type. To provide most flexibility, we include 90 parameters for all compartments \* 2knots=180 independent parameters.

The general theory behind calibrating stochastic models using a likelihood function involves adjusting model parameters to best match observed data, so that the model’s predictions are as close as possible to real-world observations. Calibration in this context refers to the process of finding the set of model parameters that maximizes the likelihood of observing the data given the model

 **Definition**: The likelihood function L(θ) quantifies the probability of observing the given data y (or a set of observations) for a specific set of model parameters θ.

L(θ∣y)=P(y∣θ)

where y represents the observed data, and θ are the parameters of the model that we want to estimate.

In the context of stochastic models, the likelihood function is constructed based on the probability distributions of the model’s outputs given the input parameters. The goal is to determine the values of θ (parameters) that make the observed data most likely under the model.

* **y** represents the **observed data** (what we actually measure or collect).
* **y′** represents the **true values** (the underlying, true data or outcome we are trying to infer, which may not be directly observable).
* **M** represents the **simulated outcome** under the model with parameters θθ, which gives us predictions based on the model’s assumptions and parameters.

To put this into context for constructing a likelihood function, we can consider the relationship between these components and how the likelihood function is built around the data, the model, and the parameters.

1. **1. Observed Data (y) vs. True Outcome (y′y)**

In most modeling contexts, the true outcome y′ is unobservable directly, and we rely on observed data y to infer it. There are often several sources of error that can explain the differences between the true values y′ and the observed data y. In infectious disease epidemiology, for example, the outputs are typically represented by the observed number of events—such as new diagnoses or the number of patients reaching viral suppression—over a given time period (e.g., one year).

* **Potential Sources of Error:** Sources of error in the observed data can include:
* **Sampling Variability**: Data may come from different sample populations or regions, which may not perfectly represent the larger population.
* **Reporting Inconsistencies**: Data from different sources (e.g., national vs. local health departments) may not always match due to reporting standards or practices.
* **Measurement Error**: There may be errors in how data are collected or recorded, such as errors in diagnosing HIV or in classifying the stage of the disease.
* **Modeling Assumptions**: The assumptions made in calculating or projecting the data can introduce error, such as assumptions about disease progression.

**How to estimate the error:** Sometimes, data are published with uncertainty ranges (e.g., confidence intervals), reflecting inherent variability or unknowns in the data collection or reporting process. These intervals give an indication of how much uncertainty is present in the reported data. In some cases, data is available from multiple sources that don't necessarily match. For example, HIV diagnosis data might be reported by both the CDC and local health jurisdictions, but these figures may differ.

**Characterizing and Modeling Uncertainty:** To address these discrepancies and better quantify the uncertainty around y, we can make assumptions about the distribution of the error. For example, we can model the observed data y as being drawn from a distribution around the true value y′, with some error term σ2 representing this uncertainty:

Y∣y′=f(y′,σ2)

Here, f is the likelihood function representing an appropriate distribution (e.g., normal, Poisson), and σ2 is our estimate of the error or variance.

1. **Example: HIV Diagnosis Data**

For instance, HIV diagnosis data might be available from the CDC’s Atlas website for a given year. Similarly, local jurisdictions and health departments may report similar data, but the numbers may not always match. In such cases, we can assess the likelihood that the data from each source reflects the "true" number of diagnoses. If we assume that the CDC’s figures are more reliable, we can use the differences between the various estimates as a measure of error.

This difference between estimates can help characterize the uncertainty around the true number of diagnoses, and we can use this information in our model to better represent the distribution of error. For example, if we assume the CDC estimate is the most accurate, we could model the observed data from other sources as being drawn from a distribution centered around the CDC's value, with σ2 capturing the variability or error between these sources?

Linking to likelihood:

To build a likelihood function, we aim to describe the probability of observing the data (i.e., the truth) based on simulated outcomes. In the case of count data, we typically assume the underlying distribution is Poisson.

Y′∣M=Poisson(M)

where M is both the mean and the variance of the distribution, and σ2 (the variance) is equal to M.

Poisson distribution inherently weights deviations from the mean based on its value, which is important when interpreting errors. Let’s consider two examples to illustrate this:

**Example 1:** If the target number of diagnosed cases is 6 over a given period, we have:

* M=6
* σ2=6 (since in Poisson, variance equals the mean)
* σ ​≈2

This gives an error ratio (the relative error) of: σ/M​=33%

**Example 2:** Now, suppose the target number of diagnosed cases is much higher, say M=600. Here, the variance would be σ2=600, so: σ≈24.5

The error ratio is: σ/M​≈4%

1. **Error in Poisson and Likelihood:**

As you can see, the relative error decreases as the number of events (or cases) increases, which is a natural property of the Poisson distribution. When the target value M is large, the **relative error** is smaller relative to the mean, even though the **absolute error** (in terms of the standard deviation) increases. This reflects the higher confidence in estimating larger numbers of events versus smaller numbers. This translates to the likelihood function in the following way:

1. **Small Counts** (e.g., M=6): When the true value is low, the relative error (σ/M) is larger, meaning there is greater uncertainty in our estimate. The likelihood function will be more sensitive to deviations between the observed count and the true count.
2. **Large Counts** (e.g., M=600): When the true value is high, the relative error (σ/M​) is smaller, indicating less uncertainty. The likelihood function will be less sensitive to deviations between the observed count and the true count.
3. **Approximating Poisson with Normal Distribution:**

For large values of M (the mean), the Poisson distribution becomes highly skewed and heavy-tailed. This leads to computational inefficiencies and challenges when calculating probabilities or likelihoods, especially for large MM. These difficulties are particularly problematic when using the Poisson probability mass function (PMF) or when incorporating Poisson-based likelihoods in models.

**Central Limit Theorem and Normal Approximation:**

As the mean MM increases, the Poisson distribution starts to resemble a normal distribution with the same mean and variance. This result is well-known in probability theory and is useful for simplifying computations when MM is large.

According to the **Central Limit Theorem (CLT)**:

* As MM increases, the Poisson distribution approaches a normal distribution with mean MM and variance MM.
* In mathematical terms:Y′∼Poisson(M)≈N(M,M)Y′∼Poisson(M)≈N(M,M)

**Implications:**

* **Mean**: The mean of the normal distribution is MM (the same as the Poisson mean).
* **Variance**: The variance of the normal distribution is also MM (since, in the Poisson distribution, the variance equals the mean).

Thus, the Poisson distribution Pois(M)Pois(M) can be approximated by a normal distribution N(M,M)N(M,M) when MM is sufficiently large.

**When is Normal Approximation Appropriate?**

The normal approximation to the Poisson distribution works well when MM is sufficiently large. A common rule of thumb is that the approximation is accurate when:

M≥10M≥10

For smaller values of MM, the Poisson distribution remains more skewed, and the normal approximation may not accurately reflect the characteristics of the data. Therefore, for smaller MM, it's better to use the Poisson distribution directly rather than the normal approximation.

Putting it al together:

Y|y’=N(y’,sigma2)

Y’|M=N(M,M)

Y|M ~ N(M, sigma2+M)

# Calibration of Stochastic Models Using Likelihood Functions

**Introduction:** In statistical modeling, one of the primary challenges is calibrating a model to accurately represent real-world phenomena, especially when the true outcomes are not directly observable. In stochastic models, this task often involves estimating the model parameters that maximize the likelihood of observing the data, given the model's assumptions. The process of **model calibration** ensures that the model’s predictions align as closely as possible with observed data, enabling more accurate inferences and predictions.

This chapter explores the general theory behind constructing likelihood functions in stochastic modeling. Specifically, we will focus on how likelihood functions can be used to bridge the gap between observed data, true but unobservable outcomes, and the model's simulated predictions. By accounting for errors and uncertainties in the data, we can develop a robust framework that helps to quantify and incorporate these uncertainties into the calibration process.

**Motivating Example: Disease Dynamics Model**

In this example, we are evaluating the performance of a disease dynamics model by compring projected number of diagnosed cases against the observed data. To assess the model's accuracy, we will consider two different scenarios as follow:

* In **Scenario 1**, the target is to match the number of diagnosed disease cases over a year, where the actual observed number is 6 cases, and the model predicts 4 cases. The absolute error in this case is 6 - 4 = 2 cases, and the relative error is calculated as 2/6 = 33%. At first glance, this error might not seem terribly large, given that the model is missing only 2 cases. However, when considering the relative error of 33%, we see that this is a substantial portion of the target value. This prompts the question of whether the model’s performance is good enough, or whether the error is too large to accept the model as an accurate representation of reality.
* In **Scenario 2**, the target number of diagnosed cases is much higher: 600 cases, while the model predicts 500 cases. In this case, the absolute error is 100 cases, which seems larger than in Scenario 1. However, the relative error is 100/600 = 16.7%, which is much smaller than the 33% error in the first scenario. This suggests that, despite the larger absolute error, the model’s prediction is more accurate relative to the target when we take into account the larger scale of the numbers involved. Thus, the model appears to perform better in this scenario in relative terms, even though the absolute error is larger.

To make a more informed decision about the model's acceptability, we must also consider how the **uncertainty** in the **model projections** and **observed data** could influence our judgment.

1. **Uncertainty in observed data** can arise from several sources, such as poor data quality, reporting inconsistencies, or measurement error. We will now explore two assumptions about the measurement error in the observed data: 10% and 50%, to see how they affect our evaluation of the model’s predictions.

* In **Scenario 1**, if we assume that the observed 6 cases have a **10% measurement error**, this means the true number of cases could realistically fall within a range of 5.4 to 6.6 (since 10% of 6 is 0.6). In this case, the model’s prediction of 4 cases is not within the uncertainty range of the observed data, which is 5.4 to 6.6. The absolute error of 2 cases might seem small, but given the narrow range of uncertainty (±0.6), this difference could be considered too large to accept the model’s performance. Therefore, in this scenario, we might reject the model or at least consider improving it.
* However, if we assume a **50% measurement error** in the observed 6 cases, the true number of cases could realistically fall between 3 and 9 (since 50% of 6 is 3). In this case, the model’s prediction of 4 cases is comfortably within the range of 3 to 9. Given the wide uncertainty range, the absolute error of 2 cases is much more acceptable. Therefore, with 50% measurement error, the model’s prediction would likely be considered acceptable.

1. **Uncertainty in model projections** stems from the stochastic nature of simulation models, where random seeds are used to generate a different array of events in each run. This variability introduces random fluctuations in the model’s outcomes. To quantify and assess this stochastic variation, we can run multiple simulations (replicates) of the model and compute summary statistics such as the mean and standard deviation. If the simulated outcomes are approximately normally distributed, the mean can be estimated by averaging all simulated outcomes, and the stochastic variance can be estimated as the average of the squared differences between each outcome and the mean. In our example, we will consider two scenarios where the standard deviation (sigma) is evaluated at 10% and 50% of the predicted mean.

* In **Scenario 1**, if we assume a standard deviation of 10% around the model's predicted mean of 4 diagnosed cases, we can quantify the uncertainty in the model’s output. f the simulated outcomes follow a normal distribution, we can use this standard deviation to calculate a confidence interval for the model's predictions. For example, assuming a 95% confidence level, the range of likely outcomes would be within two standard deviations of the mean. This would give a confidence interval of 4 ± 0.8 cases, or between 3.2 and 4.8 diagnosed cases. Since the target of 6 diagnosed cases falls outside the model's 95% confidence interval (3.2 to 4.8), it indicates that the model's prediction doesn't adequately capture the observed data. While the model is close to the target, its uncertainty range suggests it's not a good fit, as the observed value is higher than the predicted range.
* However, if we assume a standard deviation of 60% around the model's predicted mean of 4 diagnosed cases, the 95% confidence level will range from 1.6 and 6.4 diagnosed cases. In this case, the target of 6 diagnosed cases falls at the upper end of the model's 95% confidence interval. The uncertainty in the model's prediction (as indicated by the wider confidence interval) suggests that, while the model may not be an exact match, it is a more acceptable fit given the larger variability in the predictions.
* However, if we assume a **50% measurement error** in the observed 6 cases, the true number of cases could realistically fall between 3 and 9 (since 50% of 6 is 3). In this case, the model’s prediction of 4 cases is comfortably within the range of 3 to 9. Given the wide uncertainty range, the absolute error of 2 cases is much more acceptable. Therefore, with 50% measurement error, the model’s prediction would likely be considered acceptable.

In conclusion, the decision to accept or reject a model based on its prediction accuracy is not solely determined by the absolute or relative error but must also take into account the uncertainty in the observed data. Likelihood functions provide a generalizable framework to incorporate these uncertainties by quantifying the probability of observing the given data under different model assumptions and parameter values. This approach ensures that models are evaluated not just on their fit to observed data but also on their ability to reflect the true underlying dynamics while accounting for the uncertainty that naturally arises from real-world data collection.

**1. Definition of the Likelihood Function**

The likelihood function L(θ∣y)L(θ∣y) quantifies the probability of observing the given data yy for a specific set of model parameters θθ. Mathematically, it is expressed as:

L(θ∣y)=P(y∣θ)L(θ∣y)=P(y∣θ)

where:

* yy represents the observed data (what we actually measure or collect),
* θθ are the parameters of the model that we aim to estimate.

In stochastic models, the likelihood function is built based on the probability distributions of the model’s outputs, given the input parameters. The objective is to determine the values of θθ that maximize the likelihood of observing the data under the model.

**2. Observed Data yy vs. True Outcome y′y′**

In most modeling contexts, the true outcome y′y′ is not directly observable. Instead, we rely on observed data yy to infer y′y′. Differences between the true values y′y′ and the observed data yy can arise from various sources of error.

In the context of infectious disease epidemiology, for example, observed data typically represent counts of events (such as new disease diagnoses or the number of patients reaching viral suppression) over a specific time period (e.g., one year).

**Potential Sources of Error:**

1. **Sampling Variability**: Data may come from different sample populations or regions, which may not perfectly represent the entire population.
2. **Reporting Inconsistencies**: Data from different sources (e.g., national vs. local health departments) may not always match due to differences in reporting standards or practices.
3. **Measurement Error**: Errors may occur in data collection or recording (e.g., diagnosing the disease or classifying its stage).
4. **Model Assumptions**: Assumptions made in projecting or calculating the data (e.g., assumptions about disease progression) can also introduce error.

**3. Estimating the Error and Characterizing Uncertainty**

Data are sometimes published with uncertainty ranges (e.g., confidence intervals), reflecting inherent variability or unknowns in the data collection process. These ranges help quantify the uncertainty surrounding the observed data. In cases where data are available from multiple sources, these sources may not match exactly. For instance, HIV diagnosis data may be reported by both the CDC and local health departments, but the numbers may differ.

To account for this uncertainty, we can model the observed data yy as being drawn from a distribution centered around the true value y′y′, with an error term σ2σ2 capturing this uncertainty. The relationship between the observed data and the true outcome is:

y∣y′=f(y′,σ2)y∣y′=f(y′,σ2)

where:

* ff is the likelihood function representing an appropriate distribution (e.g., Poisson, normal),
* σ2σ2 is the variance or error term.

**4. Example: HIV Diagnosis Data**

For example, HIV diagnosis data might be available from the CDC's Atlas website, while local health departments may report similar data. In such cases, we can assess the likelihood that the data from each source reflects the "true" number of diagnoses. If we assume that the CDC data is more reliable, we can model the observed data from other sources as being drawn from a distribution centered around the CDC's estimate, with σ2σ2 capturing the variability or error between the sources.

This discrepancy between different data sources provides a measure of error, which we can incorporate into our model to more accurately represent the distribution of uncertainty.

**5. Linking to the Likelihood Function**

To construct a likelihood function, we aim to describe the probability of observing the data based on simulated outcomes. In the case of count data, we typically assume the underlying distribution is Poisson. The relationship between the true outcome and the simulated outcome is given by:

y′∣M=Poisson(M)y′∣M=Poisson(M)

where MM is both the mean and variance of the Poisson distribution.

The Poisson distribution inherently weights deviations from the mean based on its value. This is important when interpreting errors. Let’s consider two examples to illustrate this:

* **Example 1**: If the target number of diagnosed cases is M=6M=6, then:
  + M=6M=6
  + σ2=6σ2=6
  + σ≈2σ≈2

The relative error (error ratio) is σM=26≈33%Mσ​=62​≈33%.

* **Example 2**: If the target number of diagnosed cases is M=600M=600, then:
  + M=600M=600
  + σ2=600σ2=600
  + σ≈24.5σ≈24.5

The relative error is σM=24.5600≈4%Mσ​=60024.5​≈4%.

**6. Error in Poisson and Likelihood**

As observed, the relative error decreases as the number of events (or cases) increases. This reflects the natural property of the Poisson distribution: the uncertainty in estimating larger numbers of events is smaller compared to smaller numbers. This can be interpreted in terms of the likelihood function as follows:

* **Small Counts (e.g., M=6M=6)**: When the true value is low, the relative error σMMσ​ is larger, meaning there is greater uncertainty in the estimate. The likelihood function will be more sensitive to deviations between the observed count and the true count.
* **Large Counts (e.g., M=600M=600)**: When the true value is high, the relative error σMMσ​ is smaller, indicating less uncertainty. The likelihood function will be less sensitive to deviations between the observed count and the true count.

**7. Approximating Poisson with Normal Distribution**

For large values of MM, the Poisson distribution becomes highly skewed and heavy-tailed, making it computationally inefficient and difficult to calculate probabilities or likelihoods. This is particularly problematic when using the Poisson probability mass function (PMF) for large values of MM.

**Central Limit Theorem and Normal Approximation**

As the mean MM increases, the Poisson distribution becomes more symmetric and resembles a normal distribution. This is a well-known result from probability theory and is useful for simplifying computations when MM is large.

According to the **Central Limit Theorem (CLT)**:

Y′∼Poisson(M)≈N(M,M)Y′∼Poisson(M)≈N(M,M)

where N(M,M)N(M,M) denotes a normal distribution with mean MM and variance MM. This approximation becomes more accurate as MM increases.

**Implications of the Normal Approximation**

* **Mean**: The mean of the normal distribution is MM, which is the same as the Poisson mean.
* **Variance**: The variance of the normal distribution is also MM, since in the Poisson distribution, the variance equals the mean.

Thus, the Poisson distribution Poisson(M)Poisson(M) can be approximated by a normal distribution N(M,M)N(M,M)when MM is sufficiently large.

**8. When is the Normal Approximation Appropriate?**

The normal approximation to the Poisson distribution is generally valid when MM is sufficiently large. A common rule of thumb is that the approximation works well when:

M≥10M≥10

For smaller values of MM, the Poisson distribution remains more skewed, and the normal approximation may not accurately reflect the characteristics of the data. In these cases, it is better to use the Poisson distribution directly.

**9. Putting It All Together**

Combining the concepts of error in the observed data and likelihood construction, we can summarize the key relationships:

1. **For observed data yy and true values y′y′**:

y∣y′∼N(y′,σ2)y∣y′∼N(y′,σ2)

1. **For simulated data MM under a Poisson distribution**:

y′∣M∼N(M,M)y′∣M∼N(M,M)

1. **The likelihood function given MM and observed data yy**:

y∣M∼N(M,σ2+M)y∣M∼N(M,σ2+M)

This framework allows for constructing likelihood functions that capture the relationship between the true values, the simulated data, and observed data, accounting for errors and uncertainties in the process.

# ESTIMATING LIKELIHOOD ERRORS

Overall, the likelihood is represented by a joint likelihood with seven components by data type. Each likelihood follows a multivariate normal distribution (as an approximation to a binomial distribution) where the observed data is centered at some true value (unknown) with some level of measurement error; and the true, unknown value is centered at the model-generated estimate with some level of model error.

* Measurement error is determined based on the source of the data type . Model error assumes a normal approximation to a binomial distribution, where the variance of the model estimates equals the mean estimated value. Both measurement error/bias and model estimates are allowed to be correlated: we allow measurement error/bias to be correlated over time (e.g., the error in reported incidence for 2009 is correlated to the error in reported incidence for 2010), and model estimates to be correlated due to overlapping strata
* Correlations are represented using an estimate of standard deviation and a correlation matrix with either an auto regressive or compound symmetry structure. An auto regressive structure assumes correlations between one year apart are stronger than those 10 years apart, for example, and is used for data over longer periods of time. A compound symmetry structure assumes the correlation between one year apart is the same as the correlation between 10 years apart and is used for data over shorter periods of time.”

## Theory

The two main assumptions allow us to model the uncertainty in both the reported data and the simulated outputs.

**Assumption 1: Measurement Error in Reported Data**

his assumption states that the reported values, ​, are an approximation of the "true" values, , with some measurement error. This is modeled as a multivariate normal (MVN) distribution:

Where

* is the reported value (e.g., CDC reported number of diagnosis),
* is the “true” underlying value
* is the measurement error covariance matrix

However, the complication is that data is often not reported at the finest granularity across all dimensions (e.g., age and race combinations), but instead aggregated over certain dimensions (e.g., total counts by age or by race, but not both). This creates the need for a transformation to account for the aggregated reporting.

Example: Consider four groups:

* Young Black (YB)
* Old Black (OB)
* Young White (YW)
* Old White (OW)

The true number of diagnoses is represented as while the reported data is aggregated along the margins (age and race separately), so: where  and are aggregated over age (young and old), and ​ and  are aggregated over race (black and white).

To transform the true data into the reported data format, we use a transformation matrix M, which maps the more detailed data  into the coarser, aggregated data ​. The matrix MM would look like:

**Summary of Assumption 1:**Thus, the reported data ​ is modeled as: Where M is the transformation matrix that maps the true, detailed data to the reported, aggregated data .

**Assumption 2: Simulation Error in Model Output**

The second assumption deals with the simulated values from the model. We assume that the simulated values, , approximate the true underlying values data , but with some simulation error . This is also modeled as a multivariate normal distribution: 

Where:

* : The simulated value from the model.
* : The simulation error covariance matrix.

**Combining Assumptions 1 and 2**

When we combine these two assumptions, we aim to model the reported data ​ as a function of the simulated values , considering both the measurement error and the simulation error

This equation tells us that the reported data ​ follows a multivariate normal distribution with a mean  (the simulated values transformed into the reported data format) and a combined error covariance , which accounts for both measurement error and simulation error.

## Components Required for Full Error Estimation:

To estimate the total error between the simulated data  and the reported data ​​, we need three components:

### Transformation Matrix M:

This maps the detailed simulated data into the aggregated form of the reported data. The matrix M is constructed based on the comparison of dimensions between the simulated and reported data (as mentioned, this is estimated using ontology codes or other metadata).

### Simulation Error Γ:

This error represents the uncertainty in the simulated values. For frequency outcomes (like counts of new diagnoses), the assumption is that the simulation error follows a Poisson distribution, where the variance equals the mean.

This reflects the fact that the variance for each outcome is proportional to the mean value of the simulation for that outcome.

**Why Poisson?** The Poisson distribution is typically used to model the number of occurrences of an event within a fixed period or population size. A key property of the Poisson distribution is that the mean and variance are equal: Var(y)=E(y)

This is useful when dealing with frequency data, especially when there is a mean-variance relationship. For example, if the simulated number of new diagnoses for a particular group is expected to be , then the standard deviation is also  ​, which matches the behavior of real-world count data where the variability grows with the magnitude of the counts.

The Poisson distribution assumes that successive draws (e.g., counts of new diagnoses from one year to the next) are independent of each other. This assumption is reasonable in many epidemiological models where the occurrence of new cases in one year or group doesn't directly affect the number of cases in another year or group.

## Measurement Error Σ:

This captures the uncertainty in the reported data, such as inaccuracies in CDC-reported diagnoses. Σ is typically estimated from external data or through expert judgment.

It’s easier to break covariance matrix into its two main components—**standard deviations** and the **correlation matrix**

### Component1: Standard Deviation:

The  for each dimension gives us an idea of the error for that dimension

We can estimate SD in multiple ways:

**Option1) Use of External Data (Direct Estimation):** Some external datasets or validation studies report measurement errors and we can translate that directly to SD

For example, the Census reports a 3% error in their population size estimates

Should we assume , and therefore ? The problem with this approach is that if or , the error remains the same

We could assume ? This translate to , which is the coefficient of variation.

We could also assume , and say if q= 0.5, this means:

Todd has tried these combinations and concluded that CV was the best one for population count.

**Option 2) Use of historical data from the same source:** this is true for situation where historical data is corrected in a future publication and we have 2 versions of the same data. For example, CDC published new HIV diagnosis counts each year but the values that are later reported in the HIV atlas are different from the original reports.

See calculating\_error\_terms\_for\_ehe\_likelihoods.R for an example

### Component2) The correlation matrix

Correlations are derived from the covariance matrix by normalizing each covariance by the product of the standard deviations of the related variables. This provides a standardized measure of the relationship between errors across different dimensions.

The correlation captures how measurement errors in one dimension are related to one another. For example:

* If the reported values are inaccurate in **year 1**, how likely are they to also be inaccurate in **year 2** and **year 3**?
* If the values are incorrect for the **Black race**, to what extent are they also incorrect for **White** or **Hispanic** populations?

In practice, we typically assume that measurement error correlations exist **over time** (i.e., errors in different years are correlated), but not across other strata such as **age, race, or sex**.

There are 2 typical formats for correlations that are applicable here:

1. **Autoregressive (AR) Correlation:** In an AR(1) process, correlation between consecutive years (or time points) decreases exponentially as the time gap increases.

Where ρ is the correlation coefficient between successive time points.

1. **Compound Symmetry (CS):** All pairs of observations within the same group (e.g., race, age group) have the same correlation. This is often used for longitudinal data or repeated measures within a group. Constant correlation across all pairs of observations in the same group.

Where ρ is the constant correlation between all pairs of observations.

In reality, a hybrid approach is often needed for modeling error correlations, depending on the data structure and available information. For example, when we have **limited data points** (e.g., 10 years of observations), a **compound symmetry (CS)** model is frequently used. This assumes that the error is similarly distributed across all data points. In such cases, if we know that the reported values were off by 10% in 2000, it would be reasonable to assume that they were also off by around 10% in 2010. This simplifies the correlation structure, making it easier to model with limited data.

However, when working with a **larger dataset** spanning a **longer time horizon**, an **autoregressive (AR)** model may be more appropriate. This model assumes that errors in consecutive time periods are more strongly correlated, but this correlation weakens over time. For example, if the new diagnosis data in 2000 were off by 10%, an AR model would suggest that errors in the following years (e.g., 2001, 2002) are similarly off, but the impact would diminish over longer periods (e.g., by 2010).

# CALIBRATION TARGETS

## Demographic model

* **Basic Likelihood:** This applies when data at the desired location level is available.
* **Handling Data Gaps:** At times, calibration data may not be available for a specific location of interest. For example, if we need to calibrate the proportion of awareness in Baltimore but lack local data, we could use data from Maryland and incorporate some uncertainty to account for differences between these locations.

**Targets for the Demographic Model:**

1. **Population Size:** Available from 2010 to 2020, broken down by age, sex, and race.
2. **Deaths:** Total count available for two periods—2001 to 2010 and 2011 to 2020.
3. **Births:** Available from 2007 to 2023, broken down by the age and race of mothers.

# SEXUAL MIXING

## AGE Mixing Model Overview

The model for mixing considers three key components: age, sex, and race.

In general, we model contacts between the following groups:

* MSM (men who have sex with men) with other MSM, heterosexual men, or women.
* Heterosexual men with MSM, other heterosexual men, or women.
* Women with MSM or heterosexual men.

## Age Mixing Model

For the age component, Todd utilized data from a study in Australia that reported the ages of each pair of partners. Based on these data, the following model was assumed to represent differences in partner ages:

Age difference=N(μ,σ) where

μ=B0+B1×a

σ=L0+L1×a

This gives us:

Diff(a)=N(B0+B1×a,L0+L1×a)

Here, a represents the age of the individual, and B0,B1, L0, and L1 are coefficients estimated from the data.

Using this model, we estimate three separate age models for females, heterosexual men, and MSM. These are stored in PAIRING.INPUT.MANAGER$sex.age.models.

### Age Mixing Matrices

The next step is to compute the age mixing matrices, which represent the *proportion of contacts that occur between different age groups*. Since we have categorical age groups, we need to determine the proportion of the population that falls into specific age years. For example, within the age group 25-34, we calculate the proportion of individuals who are 25, 26, 27, and so on.

In a simplified case, we can assume a uniform distribution across ages within an age group, meaning that each year represents 1 tenth of the total population for that group. We can then estimate the proportion of contacts from a "mixture normal distribution."

For example, to estimate the proportion of contacts for women aged 25-34 that occur with individuals in the 13-24 age group:

P(13−24)=0.1×pnorm(13−24, μ25, σ25)+0.1×pnorm(13−24,μ26,σ26)+…+0.1×pnorm(13−24,μ34,σ34)P(13−24)

In JHEEM, Todd used a more sofisticated method for maping popualtion proportions based on census in each location

* get.heterosexual.male.single.year.age.counts()
* get.female.single.year.age.counts()
* get.msm.single.year.age.counts()

Using this approach, we estimate three age mixing matrices: one for females, one for heterosexual males, and one for MSM.

### Age of Sexual Debut and Availability

Additionally, we must model the reduction in sexual availability for the youngest and oldest age groups. This is handled using the get.sexual.availability() function, which maps changes in sexual availability across ages. The model reflects an increase in sexual activity starting from age 13, reaching 100% at ages 20 to 64, and gradually tapering off until age 85, the final age group.

### Calibration

All the parameters introduced so far are estimated from data and remain fixed. However, we include one additional parameter specifically for calibration—a multiplier applied to the standard deviation in the age model. This calibration parameter adjusts the variability in age assortativity (age.mixing.sd.mult)

* Larger values of the multiplier increase the variability in age differences, resulting in less age assortativity (i.e., individuals tend to partner with others from a wider range of ages).
* Smaller values decrease the variability in age differences, leading to greater age assortativity (i.e., individuals tend to partner with others closer to their own age).

This allows us to fine-tune the model to reflect observed patterns in age mixing.

## Sex Mixing Model

We aim to construct a 3x3 matrix representing the proportion of partnerships between females, heterosexual males, and MSM (men who have sex with men). In this model, only female-female partnerships are excluded, while all other pairings can have a positive value.

**Logic**

Consider the case for females: if there is no sex assortativity, the proportion of female partners who are MSM or heterosexual males is proportional to their population distribution in a given location. For example, if 20% of men in Baltimore are MSM, then females would be expected to have 20% MSM and 80% heterosexual male partners. This implies that the **observed-to-expected (OE) ratio** for MSM partnerships would be equal to 1. However, when there is assortativity (i.e., a preference for partnering within specific groups), the OE ratio will deviate from 1—being either greater than or less than one, depending on the degree of assortativity.

### Estimating proportion of females contacts with msm and male hetrosexuals

We estimate the prior value for the OE ratio from a single study:

name=′oe.female.pairings.with.msm′, value=0.0895(Pathela 2006)

Using this value, we can estimate the proportion of female partnerships that are with MSM or heterosexual males in each location as follows:

* Pmsm=0.089×prop.males.msm / (0.089×prop.males.msm+prop.males.not.msm)
* Phet.male=prop.males.not.msm/ (0.089×prop.males.msm+prop.males.not.msm)

Since these are the only two options, the total must satisfy:

* Pmsm+Phet.male=1

This approach allows us to estimate the proportions of MSM and heterosexual male partnerships for females across different locations.

The oe.female.pairings.with.msm is also used for calibration

### Estimating proportion of male contacts with msm and male heterosexuals

What fraction of hetrosexual males have contact with other men: fraction.heterosexual.male.pairings.with.male, value =0.004

What fraction of msm

## Race mixing model

Similar to sex mixing, this relies on observed to expected proportion of contact between difference racial groups

# SYPHILIS TESTING

#probability of having received a HIV test in the last year (used to approximate syphilis screening rate)

#needs a functional form

#assuming X% of people with genital ulcer/rash will seek care, now the quesiton is what proportion of people develop genital ulcer/rash

# 40% have ulcer, 90% seek care = 36% of people with syphilis will seek care

# duration 4 monts -> rate: 36% = 1- e-rt

# look for time to diagnosis for people who are diagnosed, and the look for proportion not diagnosed

Improving efficiency and equity

* Drone operation: medical supply delivery

US department of transportation, NSF

* Deployment telehealth kiosk

Focus on rural

Network design modeling

Identify the objectives and constraints

* Changes in demand distribution in relation with existing facilities

Hospitals and health centers are clustered at more populated areas

Rural areas lack access

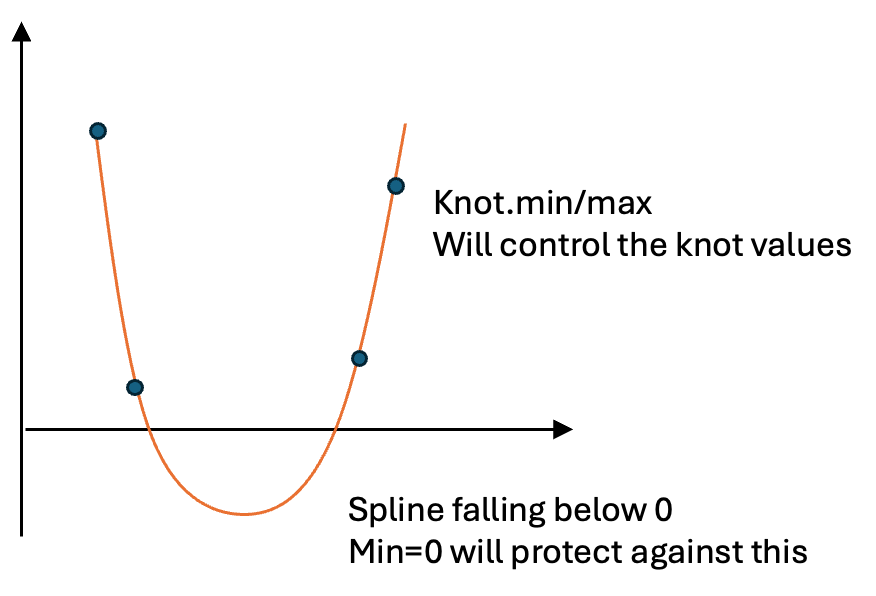
Depending on the current state of the system, we can estimate the demand: any change in system’s condition can change the demand for the future

# SETTING UP PARAMETERS FUNCTIONAL FORMS

## Natural Spline function()

**what are min/max, vs knot.min/knot.max?** Consider an example when we are setting a functional form for a rate parameter. by definition, rate can not fall below 0. There can be a situation where you have 4 knot values greater than zero. If we use an identity link, it will draw a linear projections between the knots and as long as all knots are greater than 0, we don’t have an issue. But if we do use a “log” link, the projected spline between knits could potentially fall below 0 .

* **min/max:** protect against spline values falling outside of range. The minimum and maximum values this functional form can create. The default (NA) sets to the min/max for the link specified (eg, 0-Inf for a log link, 0-1 for a logistic link)
* **knot.min/knot.max:** The min and max values that KNOTS can take before being splined (note, the min and max values the functional form can take, after splining, are determined by parameters min and max)
* **link:** The name of a transformation to the scale at which the SPLINE should apply. The knots are transformed to this scale, the spline is applied, and then the splined values are back-transformed. One of 'identity', 'log', or 'logistic'
* **knot.link:** The name of a transformation to the scale at which the knot values have alphas ADDED. The knots are back-transformed from this scale prior to the spline scale being applied



Example:

get.syphilis.to.hiv.testing.functional.form = function(specification.metadata){

create.natural.spline.functional.form(knot.times = c("1980"=1980, "1990"=1990, "2000"=2000, "2010"=2010,"2020"=2020),

knot.values=list("1980"=0.8, "1990"=0.8, "2000"=0.8, "2010"=0.8,"2020"=0.8),

knot.min = 0, #knot values can not exceed this range

knot.max = 1,#knot values can not exceed this range

min=0, #projected spline values should remain within this range

max=1, #projected spline values should remain within this range

knots.are.on.transformed.scale = F,

knot.link = "log",

link = "identity" #it's safer to use linear to avoid exponential growth,

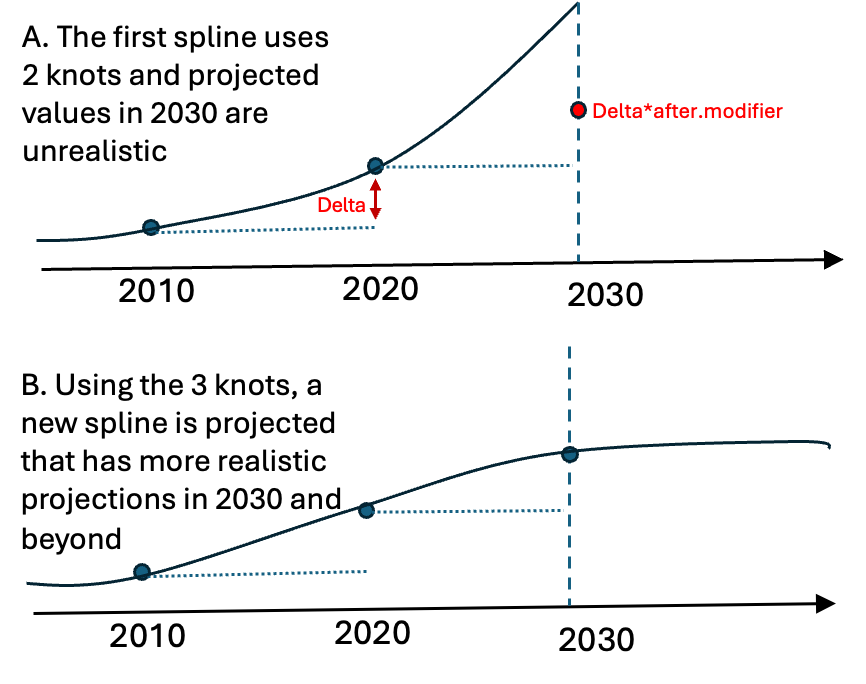
)

**What are Before.time/After.time. and Before.modifier/After.modifier?**

These parameters are used to manage spline projections for periods occurring before the first knot or after the last knot, ensuring that the projections remain realistic.

For example, consider a spline model with two knots at 2010 and 2020. Without adjustments, the projected values for 2030 may be unrealistic and excessively large. To address this, we set the after.time parameter to 2030 and apply an after.modifier of 10% to adjust future projections. Using these parameters, the model creates a new virtual knot in 2030 (red dot) and assigns it a value based on the change observed between 2010 and 2020 (delta), scaled by the modifier value.

The model then fits a new spline using all three knots, resulting in more realistic projections for the period after 2030 (Panel B).



Example:

get.emigration.rates.functional.form <- function(location, specification.metadata, population.years=DEFAULT.MIGRATION.YEAR){

rates = get.emigration.rates(location=location,

specification.metadata = specification.metadata)

create.natural.spline.functional.form(

knot.times = c(time.1 = 2010,time.2 = 2020),

knot.values = list(time.1 = rates, time.2 = rates),

link = "identity",

knot.link = 'log',

min = 0,

after.time = 2030, #this adds a new knot in 2030

after.modifier = 0.1,#sets the value of 2030 knot to 0.1\*changes between 2010-2020

knots.are.on.transformed.scale = F)

}