Estimating likelihood error

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).