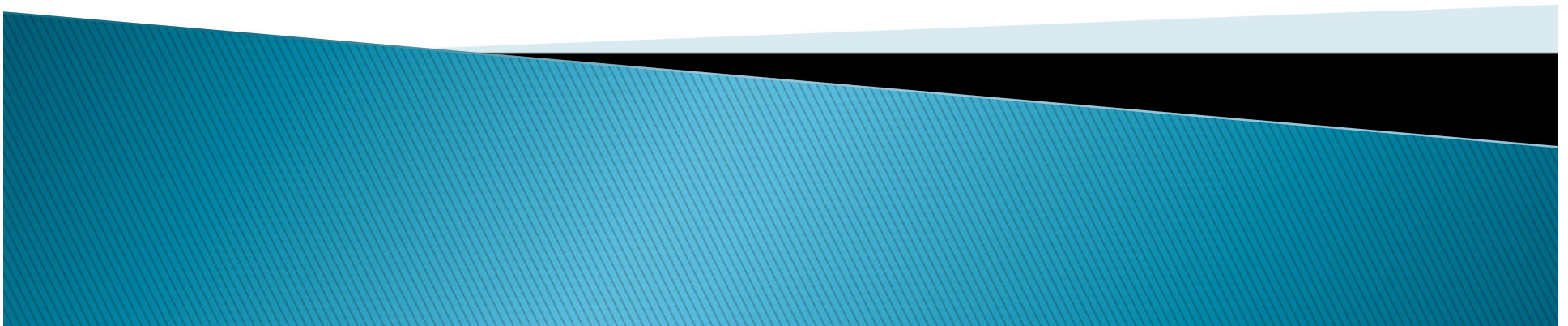


# Crash Modeling Fundamentals

Fall 2021



# Why use Statistical Models?

- ▶ Crashes are “independent” and “random” events (probabilistic events)
- ▶ Estimate a relationship between crashes and covariates (or explanatory variables)
- ▶ Determine the long-term average of crash occurrences for transportation facilities
- ▶ Have a wide variation of applications in safety analyses:
  - Prediction
  - Variable screening
  - Risk factors
  - Before-after study



# Why use Statistical Models?

- ▶ ***Understanding the System:*** The first application consists of developing models with the objective of learning something about the system from which the data are taken. Examining the sign of a coefficient is an example of such application.
- ▶ ***Screening Variables:*** The second application consists of developing models for screening purposes, where the objective is to determine which covariates have specific or significant effects on the risk of collisions. For this application, most of the modeling effort is devoted to the analysis of the covariates of the statistical models.
- ▶ ***Predictive Tool:*** The third application aims at developing models for prediction purposes. In this application, the goal is to develop models with the best predictive capabilities. These models are usually estimated using one dataset, but are applied or evaluated using a completely new dataset.



# Statistical Models For Crash Data

## Modeling Process

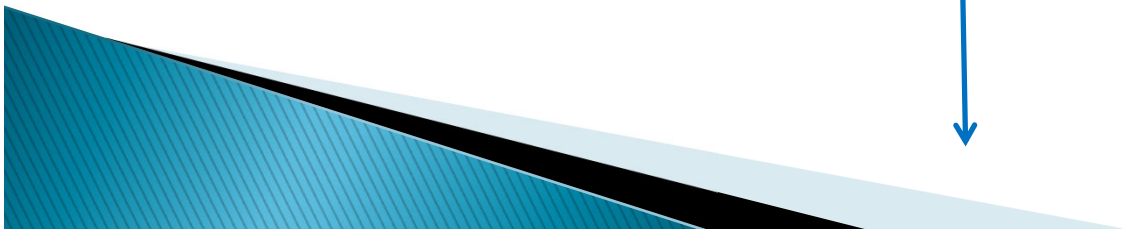
### 1. Determine Modeling Objectives

- Definition (Intersections, Pedestrians, etc.)
- Data availability
- Unit Scales (Crashes/year; Severity; etc.)



### 2. Establish Appropriate Process

- Sampling Models
- Observational Models
- Process/System State Models
- Parameter Models (Bayesian Models Only)



# Statistical Models For Crash Data

## Modeling Process

### 3. Determine Inferential Goals

- Point estimate (Value + Standard Error)
- Distribution (Bayesian Models)
- Percentiles (2.5%, 85%, etc.; Bayesian Models)



### 4. Select Computation Techniques

- Frequentist (MLE)
- Bayesian (via simulation)
- Empirical Bayes



### 5. Evaluate Models

- Goodness-of-Fit
- Prediction
- Confidence Intervals

# 1. Determine Modeling Objectives

The first step in developing statistical models is to layout the objectives of the modeling effort.

The main considerations, in this step, include application needs (e.g., prediction, screening variables), project requirements, data availability, logical scales both spatial and temporal scales of modeling units and their definitions, and range, definition, and unit of key input and output variables.



# 1. Determine Modeling Objectives

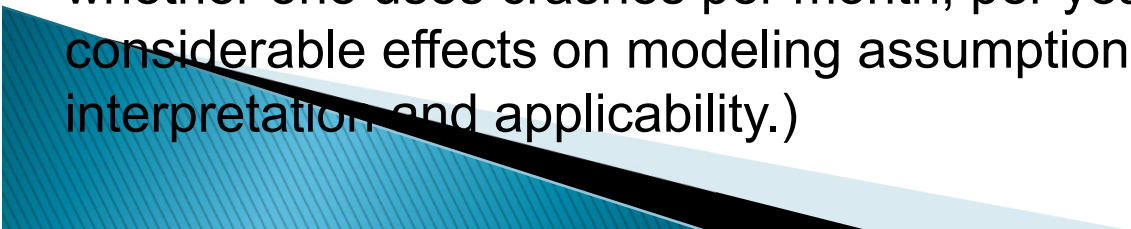
The table below lists an example of a matrix describing the modeling objectives. This table shows how the highway network is divided into segments and intersections, and the outcome of potential models. For this hypothetical project, crash-frequency and crash-severity and statistical models by collision type will be estimated, but crash cost will not be included in the analysis for segments and intersections.

TABLE 2.8 Modeling objective matrix.

Highway segments	Crash frequency	Crash severity (KABCO)	Crash frequency by collision type	Crash cost
Intersections	Y	Y	Y	N
Segments	Y	Y	Y	N
Ramps	Y	N	N	N

# 1. Determine Modeling Objectives

It is critically important in this step to determine the logical scales of modeling units and their definitions, as well as range, unit, and definition of key input and output variables. Example:

- Define spatial and physical definition of intersections and segments and the exact types of traffic crashes (e.g., intersection, intersection-related, pedestrian involved, or animal-involved crashes)
  - The range of traffic flows (e.g., AADT = 200 - 20,000)
  - Make sure commensurable data can be obtained and enough data can be collected (discuss later in the course).
  - The time unit of analysis (i.e., number of crashes per unit of time). (Note: whether one uses crashes per month, per year, per 3-year, etc., will have considerable effects on modeling assumptions and consequently on model interpretation and applicability.)
- 



## 2. Establish Appropriate Process

Typical modeling procedures employed in developing statistical models can be grouped into five major processes:

- (1) Establish a sampling model (such as those used in surveys with weight factors or stratified data),
- (2) Choose an observational model (or conditional model) (note: most crash-frequency and crash-severity models fall into this category),
- (3) Develop a process/state/system model (e.g., hierarchical/random effects models, etc.),
- (4) Develop a parameter model (for the Bayesian method and, to some degree, random-parameters models), and
- (5) Construct model and interrogation methods (e.g., interrogating theoretical models), including model comparison, sensitivity or robustness analysis, and specification test, among others.



## 2. Establish Appropriate Process

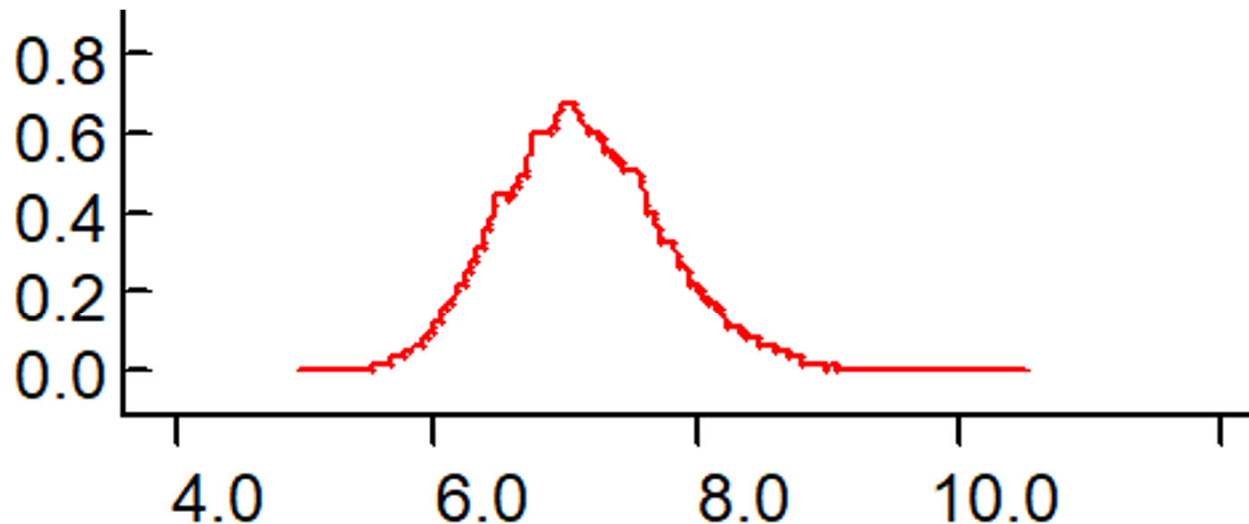
Typical modeling procedures employed in developing statistical models can be grouped into five major processes:

- (1) Establish a sampling model (such as those used in surveys with weight factors or stratified data),
- (2) Choose an observational model (or conditional model) (note: most crash-frequency and crash-severity models fall into this category),
- (3) Develop a process/state/system model (e.g., hierarchical/random effects models, etc.),
- (4) Develop a parameter model (for the Bayesian method and, to some degree, random-parameters models), and
- (5) Construct model and interrogation methods (e.g., interrogating theoretical models), including model comparison, sensitivity or robustness analysis, and specification test, among others.



# 3. Determine Inferential Goals

The inferential goals determine whether a point prediction combined with a simple estimate of its standard error (i.e., the maximum likelihood estimation method or MLE), an interval prediction (e.g., 2.5 and 97.5 percentile “credible” intervals using the Bayesian method), or a full probability distribution for the prediction is needed (also based on the Bayesian method).



**Fig. 2.7 Posterior Distribution for the Inverse Dispersion Parameter**

## 4. Select Computation Techniques

This is the process where Frequentist (analysts who use the likelihood-based method or MLE), and the Bayesian method are likely to differ in their estimating approaches and use of different “stochastic approximations” to reduce the computational burden.

Many statistical programs are now available for estimating the coefficients of statistical models for both the Bayesian and the MLE methods which fall under the exponential family of probability distributions (e.g., the Poisson model).

More difficult inferential goals will require additional sophisticated computational methods to fully capture the sampling variations in producing estimates and predictions.

# The likelihood-based method

Under this method, one estimates the parameters by maximizing the likelihood function. The likelihood function is nothing more than the joint distribution of the observed data under a specified model, but it is seen as a function of the parameters, with fixed data. Example:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^N NB(\mathbf{y}_i; \boldsymbol{\beta})$$

Where  $y_i$  is the response variable for observation  $i$ ;  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of estimable parameters;  $\mathbf{x}'_i$  is a vector of explanatory variables; and,  $p$  is the number of parameters in the model.

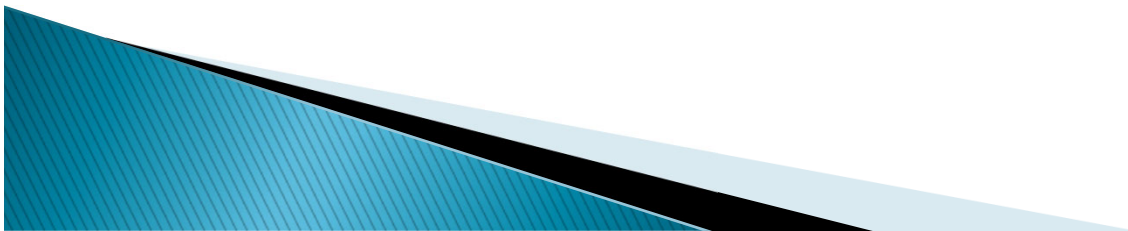


# The likelihood-based method

With the MLE, the conditional mean is considered an unknown function of the covariates, as given in the following equation:

$$E(y_i|\boldsymbol{\beta}) = \mu_i = f(\mathbf{x}'_i\boldsymbol{\beta})$$

All statistical programs have functions or subroutines that can be used for estimating the model's parameters using the MLE.



# The Bayesian method

Under the MLE method, the likelihood function is solely responsible for encoding the knowledge about the model. However, in many cases, a safety analyst may know something about the problem, even before collecting the data, often dubbed as prior knowledge or expert knowledge. The Bayesian paradigm formally combines the prior knowledge and the likelihood via the Bayes rule: we can say that posterior belief is proportional to the product of the prior belief and the likelihood. It is expressed as

$$P(\mu|y) \propto P(\mu)P(y|\mu)$$
$$\equiv \frac{P(y)}{P(\mu)P(y|\mu)}$$

Inference is typically carried out by generating approximate samples from the posterior density using MCMC techniques.

# The Bayesian method

Models elicited under the Bayesian paradigm are framed as a hierarchical or multilevel model. In highway safety, they are often defined as a hierarchical Poisson-mixed model (for crash-frequency models) or simply as an FB model, as explained earlier. Such a hierarchical modeling framework can be defined as follows:

$$(i) \quad y_i | \omega_i \sim \text{Poisson}(\omega_i) \rightarrow y_i | \omega_i \sim \text{Poisson}(\mu_i e^{\varepsilon_i})$$

$$(ii) \quad e^{\varepsilon_i} | \eta \sim \pi_{\varepsilon}(\eta)$$

$$(iii) \quad \eta \sim \pi_{\eta}(\cdot)$$

where  $\omega_i$  is the Poisson mean for observation  $i$ ;  $\pi_{\varepsilon}$  is the prior distribution assumed on the unobserved model error ( $e^{\varepsilon_i}$ ), which depends on hyper-parameter  $\eta$ , with hyper-prior  $\pi_{\eta}$ . Moreover, parameters  $\mu_i = f(\mathbf{x}'_i \boldsymbol{\beta})$  and  $\eta$  are assumed to be mutually independent (Rao, 2003).



# The Bayesian method

Models elicited under the Bayesian paradigm are framed as a hierarchical or multilevel model. In highway safety, they are often defined as a hierarchical Poisson-mixed model (for crash-frequency models) or simply as an FB model, as explained earlier. Such a hierarchical modeling framework can be defined as follows:

$$(i) \quad y_i | \omega_i \sim \text{Poisson}(\omega_i) \rightarrow y_i | \omega_i \sim \text{Poisson}(\mu_i e^{\varepsilon_i})$$

$$(ii) \quad e^{\varepsilon_i} | \eta \sim \pi_{\varepsilon}(\eta)$$

$$(iii) \quad \eta \sim \pi_{\eta}(\cdot)$$

where  $\omega_i$  is the Poisson mean for observation  $i$ ;  $\pi_{\varepsilon}$  is the prior distribution assumed on the unobserved model error ( $e^{\varepsilon_i}$ ), which depends on hyper-parameter  $\eta$ , with hyper-prior  $\pi_{\eta}$ . Moreover, parameters  $\mu_i = f(\mathbf{x}'_i \boldsymbol{\beta})$  and  $\eta$  are assumed to be mutually independent (Rao, 2003).

# The Bayesian method

Depending on the specification of the priors  $\eta$  and  $\pi_\eta(\bullet)$ , different alternative hierarchical models can be defined.

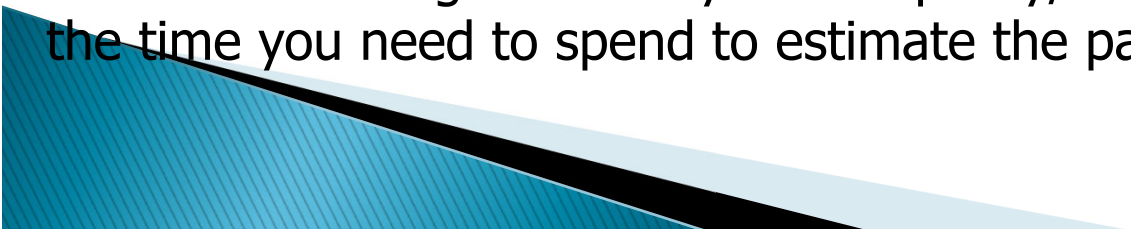
Poisson-gamma/NB:

$$e^{\varepsilon_i} | \varphi \sim \text{gamma}(\varphi, \varphi) \text{ and } \varphi \sim \text{gamma}(a, b)$$

Poisson-lognormal:

$$\varepsilon_i = \log(e^{\varepsilon_i}) | \sigma^2 \sim \text{Normal}(0, \sigma^2) \text{ and } \sigma^{-2} \sim \text{gamma}(a, b)$$

In addition to the criteria described above (step 3), the selection of MLE vs FB should be governed by the simplicity/complexity of the model and the time you need to spend to estimate the parameters.



# 5. Methods for evaluating model performance

This section describes different methods that can be used for evaluating the model performance of crash-frequency and crash-severity models. The methods are used to measure the “goodness-of-fit” (GOF) or how well the model fits the data.

Although evaluating the fit is an important measure in the assessment of models, it should not be the sole goal for selecting a model over another. It is also important to examine what is called the “goodness-of-logic” (Miaou and Lord, 2003).

There are two approaches: 1) Likelihood-based method and 2) Error-based methods



# Maximum Likelihood

As the name implies, the most basic method consists of maximizing the LL function. This is accomplished by first taking the log of the function. Then, take the partial derivatives (first-order conditions) of the LL for each model's parameter and make each one equal to zero.

The largest value indicates the best fit. The MLE is unfortunately not dependent on the number of parameters found in the model, which could potentially lead to an overfitted model.

$$MLE = -2 \times LL$$

See Appendix A of the textbook for how to calculate the LL.



# Log-Likelihood Test

The likelihood ratio test is used to select models by comparing the loglikelihood for the fitted model (restricted—**R below**—) with the log-likelihood for a model with fewer or no explanatory variables (unrestricted or less restricted model).

$$LRT = -2 \left[ LL(\boldsymbol{\beta}_R) - LL(\boldsymbol{\beta}_U) \right]$$

# Log-Likelihood Ratio

The likelihood ratio index statistic compares how well the model with estimated parameters performs with a model in which all the parameters are set to zero (or no model at all). This test is primarily used for assessing the GOF of crash-severity models. The index is more commonly called the McFadden  $R^2$ , the  $\rho^2$  statistic or sometimes just,  $\rho$ :

$$\rho^2 = 1 - \frac{LL(\hat{\boldsymbol{\beta}})}{LL(0)}$$

$$\rho_{corrected}^2 = 1 - \frac{LL(\hat{\boldsymbol{\beta}}) - p}{LL(0)}$$

## Akaike information criterion

The Akaike information criterion (AIC) is a measure of fit that can be used to assess models. This measure uses the log-likelihood, but add a penalizing term associated with the number of variables. It is well known that by adding variables, one can improve the fit of models. Thus, the AIC tries to balance the GOF versus the inclusion of variables in the model ( $p$  below).

$$AIC = -2 \times LL + 2p$$

## Bayes information criterion

Similar to the AIC, the Bayes information criterion (BIC) also employs a penalty term, but this term is associated with the number of parameters ( $p$ ) and the sample size ( $n$ ). This measure is also known as the Schwarz Information Criterion.

$$BIC = -2 \times LL + p \ln n$$



# Deviance information criterion

When the Bayesian estimation method is used, the deviance information criterion (DIC) is often used as a GOF measure instead of the AIC or BIC.

$$DIC = \hat{D} + 2\left(\bar{D} - \hat{D}\right)$$

where  $\bar{D}$  is the average of the deviance ( $-2 \times LL$ ) over the posterior distribution, and  $\hat{D}$  is the deviance calculated at the posterior mean parameters. As with the AIC and BIC, the DIC uses  $p_D = \bar{D} - \hat{D}$  (effective number of parameters) as a penalty term on the GOF. Differences in DIC from 5 to 10 indicate that one model is clearly better.

# Widely applicable information criterion

The widely applicable information criterion (WAIC) (Watanabe, 2010) is a measure that is like the DIC (i.e., adds a penalty term for minimizing overfitting), but incorporates the variance of individual terms (the  $D$  s in the equation above). According to Gelman et al. (2014), the “WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate” (p. 9), as it is done with the AIC and DIC. Because of this, the WAIC provides a better assessment of models estimated by the Bayesian method.



# Bayes factors

The Bayes factor is a powerful tool to assess different models using the same dataset when the Bayes estimating method is used. For example, the Bayes factor,  $B_{12}$ , compares model  $M_1$  to model  $M_2$  after observing the data (Lewis and Raftery, 1997). The Bayes factor is the ratio of the marginal likelihoods of the two models being compared  $B_{12} = p(\mathbf{y}|M_1)/p(\mathbf{y}|M_2)$  For calculating the marginal likelihood, the method developed by Lewis and Raftery (1997) can be used. The approximation of the marginal likelihood is carried out on the logarithmic scale such that:

$$\log\{p(\mathbf{y}|M)\} \approx \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\{|\mathbf{H}^*|\} + \log\{f(\mathbf{y}|\boldsymbol{\beta}^*)\} + \log\{\pi(\boldsymbol{\beta}^*)\}$$

Assuming that the prior probabilities for the competing models are equal,  $B_{12}$  is expressed as follows:

$$\log\{B_{12}\} = \log\{p(\mathbf{y}|M_1)\} - \log\{p(\mathbf{y}|M_2)\}$$

A difference between 20 and 150 strongly supports the selection of Model 1 over Model 2.

# Deviance

The deviance is a measure of GOF and is defined as twice the difference between the maximum likelihood achievable ( $y_i = \mu_i$ ) and the likelihood of the fitted model:

$$D(\mathbf{y}, \mathbf{u}) = 2 \{ LL(\mathbf{y}) - LL(\hat{\boldsymbol{\mu}}) \}$$

The deviance for the NB model can be calculated as follows:

$$D_{NB} = 2 \sum_{i=1}^n \left\{ y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i + \alpha^{-1}) \ln \left[ \frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}} \right] \right\}$$

## Mean prediction bias

The mean prediction bias (MPB) measures the magnitude and direction of the model bias. It is calculated using the following equation:

$$MPB = \frac{1}{n} \sum_{i=1}^n (\mu_i - y_i)$$

A positive value indicates the model over-estimate values, while a negative value shows the model under-predict values.



# Mean absolute deviation

The mean absolute deviation (MAD) calculates the absolute difference between the estimated and observed values:

$$MAD = \frac{1}{n} \sum_{i=1}^n |\mu_i - y_i|$$

Smaller values are better.

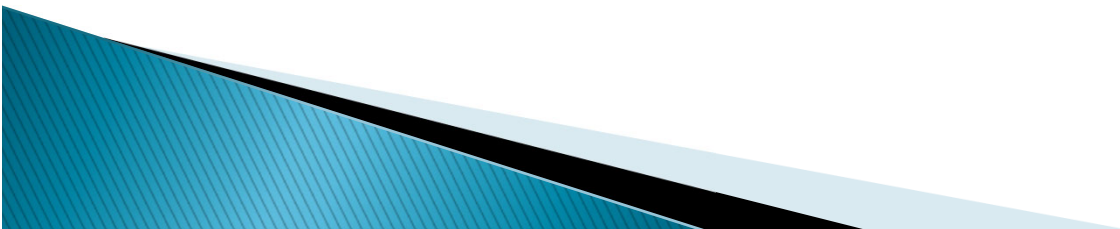


# Mean squared prediction error

The mean squared prediction error (MSPE) is a traditional indicator of error and calculates the difference between the estimated and observed values squared. The equation is as follows:

$$MPSE = \frac{1}{n} \sum_{i=1}^n (\mu_i - y_i)^2$$

A value closer to 1 means the model fits the data better.



## Mean squared error

The mean squared error (MSE) calculates the sum of the squared differences between the observed and estimated crash frequencies divided by the sample size minus the number of parameters in the model. The MSE is calculated as follows:

$$MSE = \frac{1}{n - p} \sum_{i=1}^n (\mu_i - y_i)^2$$

The MSE value can be compared to the MSPE. If the MSE value is larger than the MSPE value, then the model may overpredict crashes.



## Mean absolute percentage error

The mean absolute percentage error (MAPE) is a statistical technique that is used for assessing how well a model predicts values (in the future). It is measured as a percentage. The MAPE is calculated using this equation:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - P_i}{A_i} \right| \times 100$$

Where  $A_i$  is the actual value and  $P_i$  is the predicted value for site or observation  $i$ . It should be pointed out that the equation will not work if one or more actual values is 0. A smaller percentage indicates that a model is better at predicting values.

# Pearson Chi-square

Another useful likelihood statistic is the Pearson Chi-square and is defined as

$$\text{Pearson} - \chi^2 = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\text{VAR}(y_i)}$$

If the mean and the variance are properly specified, then

$$E \left[ \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\text{VAR}(y_i)} \right] = n \quad .$$

Values closer to  $n$  (the sample size) show a better fit.



# Coefficient of determination

Miaou (1996) has proposed using the dispersion-parameter-based coefficient of determination  $R_{\alpha}^2$  to evaluate the fit of an NB model when it is used for modeling crash data. It is computed as follows:

$$R_{\alpha}^2 = 1 - \frac{\alpha}{\alpha_{null}}$$

where  $\alpha$  is the dispersion parameter of the NB model that includes independent variables (i.e.,  $Var(Y) = \mu + \alpha\mu^2$ ); and,  $\alpha_{null}$  is the dispersion parameter of the NB model when no parameters are included in the model.

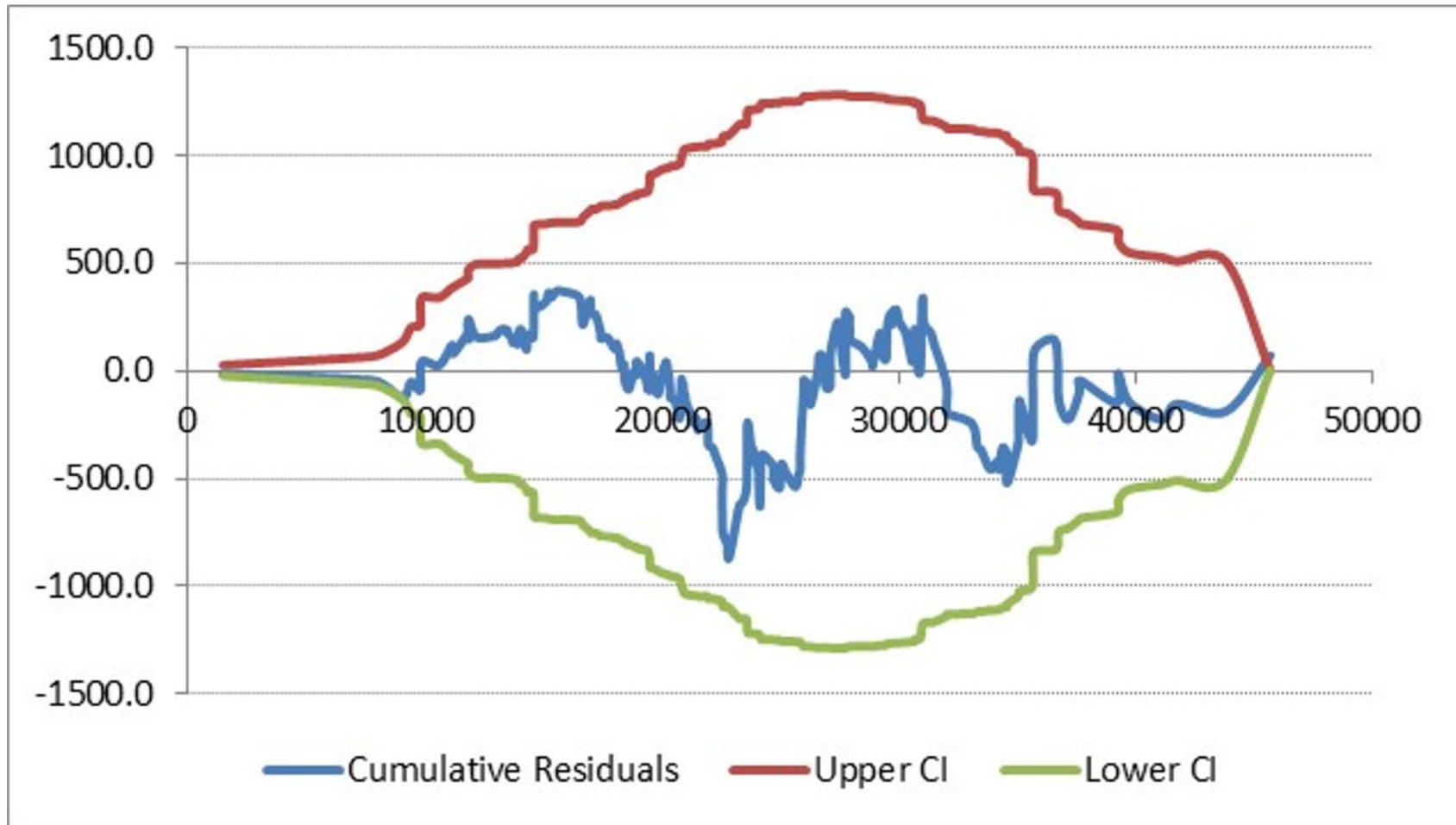
# Cumulative residuals

The cumulative residuals (CURE) consist of plotting the cumulative difference between the estimated and observed values ( $r_i = \mu_i - y_i$ , where  $r_i$  represents the residual for observation or rank  $i$ ) in the increasing order of the variable that is being analyzed.

The CURE plot allows the safety analyst to examine how the cumulative difference varies around the zero-line, which can help determine where, in the range of the variable examined, the model over- or underestimate the number of crashes.



# Cumulative residuals



**Figure 2.8 Cumulative Residuals for the Data Shown In Table 2.7**

## Cumulative residuals

To properly evaluate the fit, the 95%-percentile confidence interval (CI) needs to be calculated. The CI is calculated using the variance of the residual  $i$  (i.e.,  $r_i^2$ ) and then cumulating the variance for the increasing order of the

variable. The following equation can be used for this purpose:

$$\sigma_i^2 = \sigma^2(n_i) \times \left( 1 - \frac{\sigma^2(n_i)}{\sigma^2(N)} \right)$$

$$\pm 2 \times \sqrt{\sigma_i^2}$$

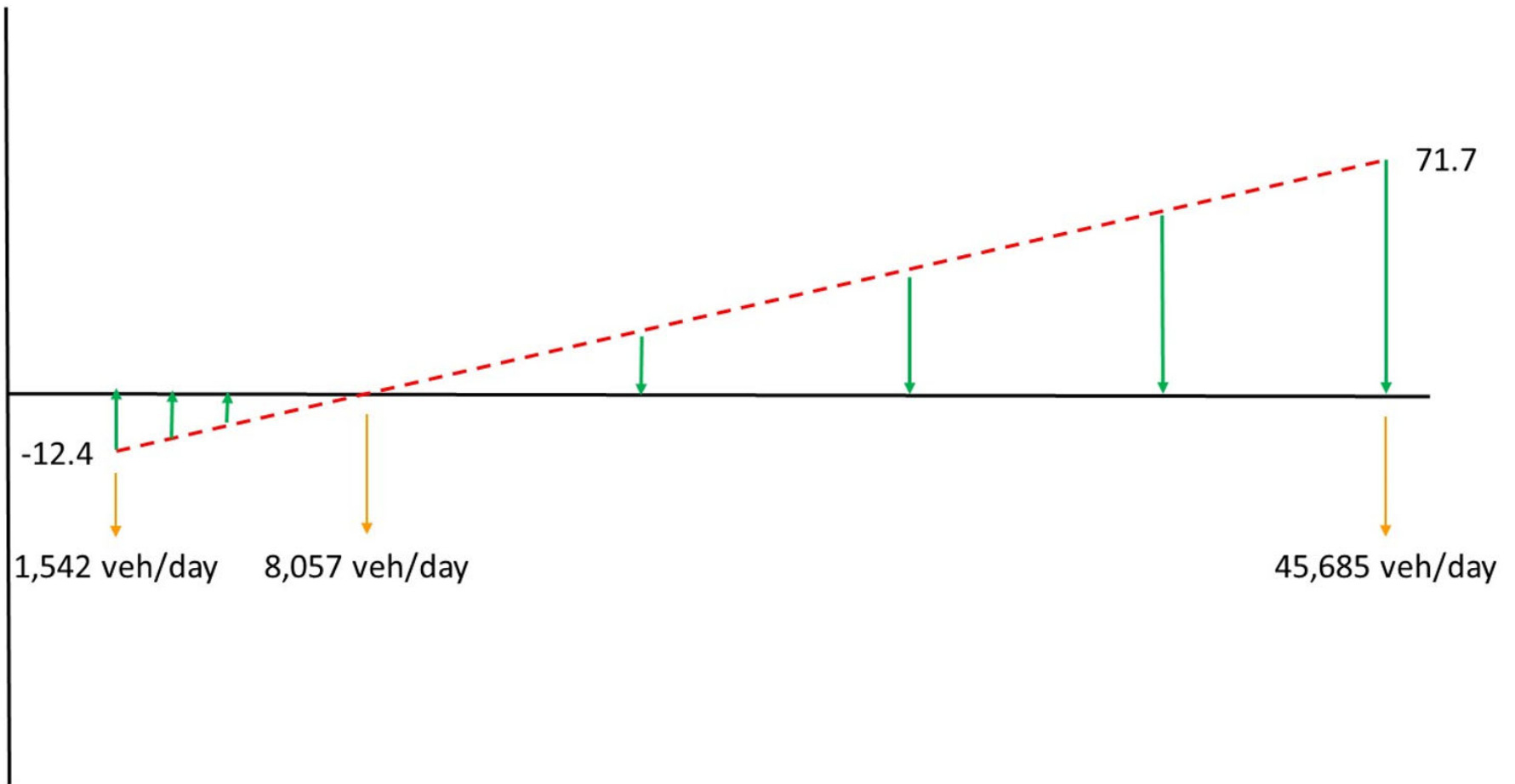
# Cumulative residuals

**Table 2.9 CURE Plot Calculations**

<b>Rank</b>	<b>Flow</b>	<b>Residuals</b>	<b>Cumulative Residuals</b>	<b>Squared Residuals</b>	<b>Cumulative Squared Residuals</b>	<b>Upper CI</b>	<b>Lower CI</b>
1	1542	-12.4	-12.4	152.6	152.6	24.7	-24.7
2	7793	-30.0	-42.4	902.1	1054.7	64.9	-64.9
3	8425	-29.4	-71.8	864.1	1918.8	87.6	-87.6
4	9142	-53.2	-124.9	2826.6	4745.4	137.6	-137.6
5	9474	74.1	-50.9	5489.3	10234.7	201.7	-201.7
6	9856	-37.6	-88.4	1412.9	11647.6	215.1	-215.1
...	...	...	...	...	...	...	...
215	45685	258.3	71.7	66733.8	1660753.9	0.0	0.0



# Cumulative residuals



**Fig. 2.9 Adjustment Procedure for the Cumulative Residuals**





# Heuristic methods for model selection



FIGURE 2.10 Heuristics to select a model between the NB and NB-L distributions.



# Heuristic methods for model selection

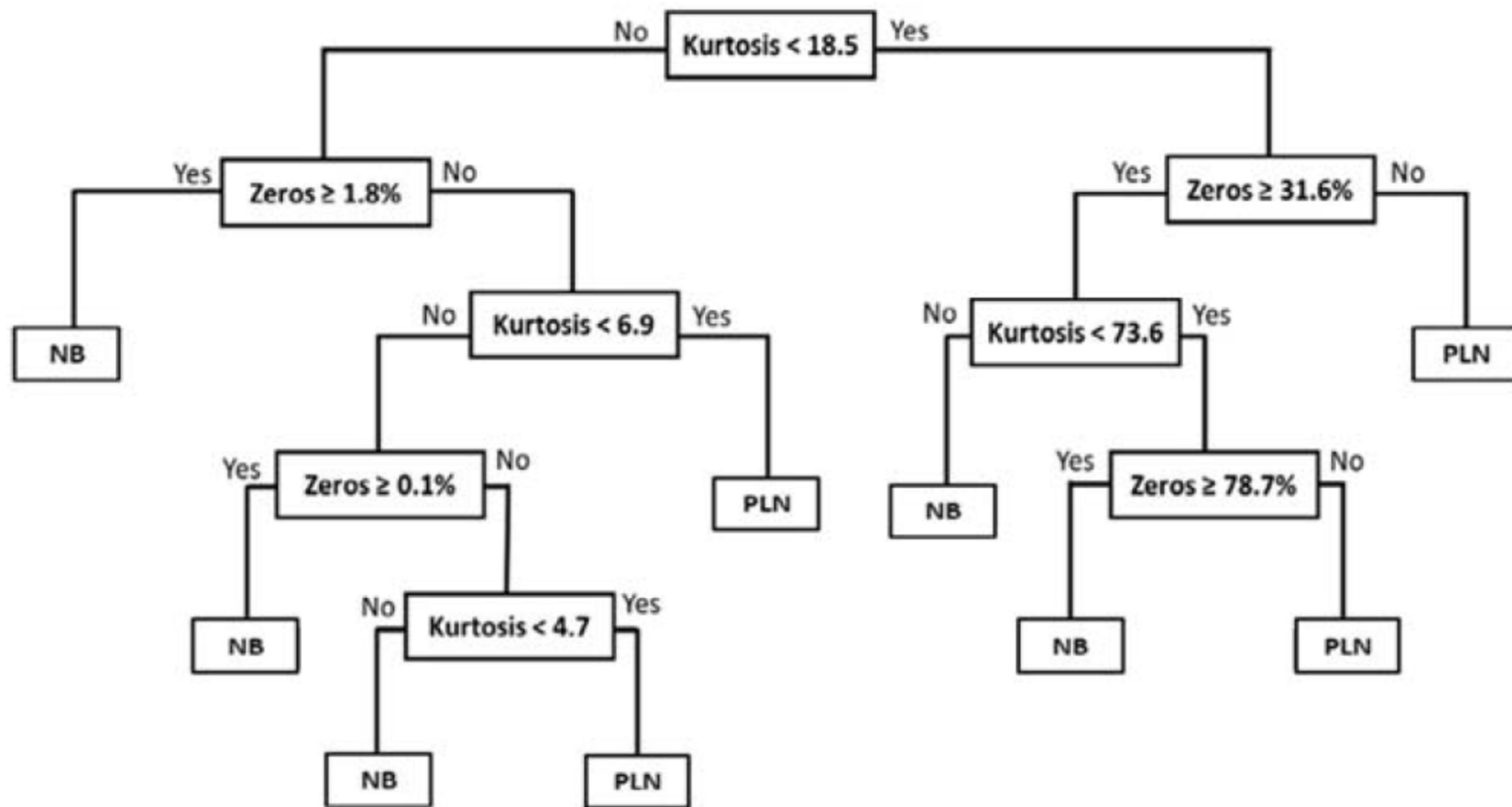


FIGURE 2.11 Heuristics to select a model between the NB and PLN distributions.