CHAPTER

2

Fundamentals and data collection

2.1 Introduction

Crashes are very complex and multidimensional events. Although, in police reports, a single factor may have been identified as the primary cause of the crash, usually several interrelated factors could have contributed to a crash. For example, we can have a scenario where an 18-year old driver, who is traveling late at night during very windy conditions in a 20-year old pickup truck, starts dosing off, then runs off the road in a horizontal curve with low-friction pavement due to lack of maintenance and a radius that is below the design standards (but approved via a design exemption), and then hits a tree located within the designated clear zone. In this scenario, the primary factor may have been identified as "falling asleep behind the wheel," but if we remove any other factors, the crash could have been avoided (e.g., adequate maintenance, no tree, no wind, on a tangent section). In addition to showing that contributing and interrelated factors can be related to the driver, the vehicle or the roadway, this scenario highlights that crashes are complex and probabilistic events (if they were deterministic events, we would be able to know when and where a crash would happen). Hence, all relevant events need to be analyzed with appropriate tools in order to account for the complexity and randomness of crash data.

This chapter describes the fundamental concepts related to the crash process and crash data analysis as well as the data collection procedures needed for conducting these analyses. The first section covers the crash process from the perspective of drivers, roadways, and vehicles. The second section describes the crash process from a theoretical and

Highway Safety Analytics and Modeling
https://doi.org/10.1016/B978-0-12-816818-9.00010-X

mathematical perspective. The third section provides important information about sources of data and data collection procedures. The fourth section describes how to assemble crash and other related data. The fifth section presents a four-step modeling procedure for developing models and analyzing crash data. The sixth section describes methods that can be used for evaluating models. The last section presents a heuristic method that allows the selection of models before models are fitted to the crash data.

2.2 Crash process: drivers, roadways, and vehicles

As explained earlier, the crash process is a very complex phenomenon that can involve a multitude of factors. These factors can generally be separated into three categories: drivers or the human element, roadways or the roadway element, and the vehicles. Over the years, studies have examined how risk factors associated with these three categories contribute to crashes. The most recent report from the National Highway Traffic Safety Administration (NHTSA), published in 2018, details in Table 2.1 the precrash caustion factors (i.e., the last event in the crash caustion chain) by proportion based on a survey of 5740 crashes throughout the United States. In this table, driver errors accounted for about 94% of all the crashes, whereas roadways and vehicles played a critical role in only 2% of all the crashes each (NHTSA, 2018). Unknown causation was estimated to be around 2% as well.

All driver errors can be further categorized as recognition (41%), decision (33%), and performance (11%). For vehicles, tires or wheel-related (35%), brake-related (22%), and steering/suspension/transmission-related (3%) were the most common failures or causational factors. It should be pointed out that in 40% of the cases where the vehicle was the primary causation factor, the exact mode of failure was unknown. For roadways,

TABLE 2.1 Critical reasons for crash occurrences (NHTSA, 2018).

Crash causation	Percentage (standard error)
Drivers	94% (2.2%)
Vehicles	2% (0.7%)
Roadways	2% (1.3%)
Unknown	2% (1.4%)
Total	100%

slick road (ice, debris, etc.) (50%), glare (17%), and view obstructions (11%) were the most common causation factors. Adverse weather accounted for about 8% of roadway-related factors.

Unfortunately, the NHTSA report does not cover the interaction between all three categories as if they were independent of each other. However, in practice, many crash events may not have independent precrash factors, as illustrated in the example presented in the introduction. As such, it is not uncommon for drivers, for example, to be confused by elements of the roadway environment (geometric design, regulatory and commercial signs, lane-occupancy, etc.) that could lead to critical driver errors (e.g., increase in perception-reaction times, getting more distracted), and eventually resulting into a vehicular crash. In a more distant study, Rumar (1985), for example, examined this interaction and found that up to 27% of the crashes included both the roadway and driver as the primary precrash contributing factors (see Fig. 2.1). Unfortunately, there has not been any more recent study that examined the interaction between vehicles, drivers, and roadways as part of the precrash causational effect. With the prevalent in-vehicle technology and distraction (i.e., talking on a cellphone, texting), such study should be done in short order.

The Haddon Matrix, originally developed in 1970 by Dr. William Haddon, was used to better understand the mechanism associated with nonintentional injuries and consequently develop strategies (i.e., countermeasures) for reducing the number and severities of these injuries. This matrix was more specifically proposed to help identify the relative importance of risk factors and to tailor interventions for the most important and identifiable factors. Initially developed as a three by three matrix (that is the three categories listed earlier), a fourth dimension that focuses on policy or social environment was eventually added to the matrix.

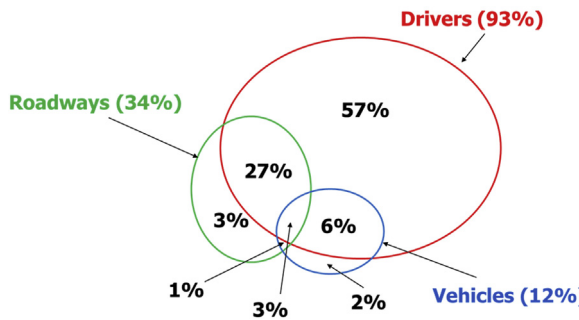

FIGURE 2.1 Precrash causation factors for roadways, drivers, and vehicles (Rumar, 1985).

TABLE 2.2 Haddon matrix for urban crashes (Herbel et al., 2010).

Event	Driver	Vehicle	Roadway	Social environment
Precrash	Poor vision, speeding	Failed brakes, worn out tire	Poorly timed traffic lights	Speeding culture, red light running
Crash	Failure to use seatbelt	Air bag failure	Poorly designed brake-away pole	Lack of vehicle regulations
Postcrash	Age (to sustain injury), alcohol	Poorly design fuel tank	Poor emergency communication	Lack of support for EMS trauma systems

In 1980, Haddon (1980) revised the matrix to address injuries caused by motor vehicle crashes. The matrix is often used for targeting specific crashes, such as pedestrians, urban or rural crashes, and driving while intoxicated (DWI) among others. In the context of the crash process, the matrix can be very useful for classifying precrash causation factors. The matrix has three levels or categories on the vertical axis: precrash, crash (during the crash), and postcrash. Moreover, it has four dimensions on the horizontal axis: the driver, the vehicle, the roadway, and the social environment. Table 2.2 illustrates an example of how the matrix can be used for crashes occurring in an urban environment.

2.3 Crash process: analytical framework

To develop and use appropriate analytical tools for analyzing crash data, which are described in the subsequent chapters, the crash process needs to be represented in a mathematical form or as an analytical framework. This framework helps address characteristics that are specifically related to the crash generation process, some of which will be covered in various chapters of the textbook.

As discussed in Lord et al. (2005) and Xie et al. (2019), a crash is, in theory, the result of a Bernoulli trial. Each time a vehicle (or any other road user) enters an intersection, a highway segment, or any other type of entity (a trial) on a given transportation network, it will either crash or not crash. For the purpose of consistency, a crash is termed a "success" while failure to crash is called a "failure." In a Bernoulli trial, a random variable, defined as X, can be generated with the following probability model: if the outcome w is a particular event outcome (e.g., a crash), then $X(w) = 1$

whereas if the outcome is a failure, then X (w) = 0. Thus, the probability model becomes.

X	1	0
P(x = X)	p	q

where p is the probability of success (a crash) and $q = (1 - p)$ is the probability of failure (no crash).

In general, if there are N independent trials (vehicles passing through an intersection, road segment, etc.) that give rise to a Bernoulli distribution, then it is natural to consider the random variable Z that records the number of successes out of the N trials. Under the assumption that all trials are characterized by the same failure process (this assumption is revisited below), the appropriate probability model that accounts for a series of Bernoulli trials is known as the binomial distribution, and is given as follows:

$$P(Z = n) = \binom{N}{n} p^n (1 - p)^{N-n} \tag{2.1}$$

where $n = 0,\ 1,\ 2, \ldots,\ N$. In Eq. (2.1), n is defined as the number of crashes or collisions (successes). The mean and variance of the binomial distribution are $E(Z) = Np$ and $VAR(Z) = Np(1 - p)$, respectively.

For typical motor vehicle (or another category of) crashes, these events have a very low probability of occurrence given a large number of trials (e.g., million entering vehicles, vehicle-miles-traveled), it can be shown that the binomial distribution can be approximated by a Poisson distribution. Under the Binomial distribution with parameters N and p, let $p = \lambda/N$, so that a large sample size N can be offset by the diminution of p to produce a constant mean number of events λ for all values of p. Then, as $N \to \infty$, it can be shown that (see Olkin et al., 1980)

$$P(Z = n) = \binom{N}{n} \left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n} \cong \frac{\lambda^n}{n!} e^{-\lambda} \tag{2.2}$$

where, $n = 0,\ 1,\ 2, \ldots,\ N$ and λ is the mean of the Poisson distribution.

The approximation illustrated in Eq. (2.2) works well when the mean λ and p are assumed to be constant. In practice, however, it is not reasonable to assume that crash probabilities across drivers and across road segments/intersections are constant. Specifically, each driver–vehicle combination is likely to have a probability p_i that is a function of the driver (e.g., driving experience, attentiveness, mental workload, risk adversity), the roadway (e.g., lane and should widths, deficiencies in design and operations, weather), and the vehicle (e.g., maintenance, safety features). All these factors (known and unknown) will affect to various degrees the

individual risk of a crash. Outcome probabilities that vary from trial to trial are known as Poisson trials (note: Poisson trials are not the summation of independent Poisson distributions; this term is used to designate Bernoulli trials with unequal probability of events). As discussed by Feller (1968), count data that arise from Poisson trials do not follow a standard distribution, but they are still considered a Poisson process. In this process, the variance of the process, $VAR(Z)$, is usually not equal to the mean of the process, $E(Z)$. In this regard, Neldelman and Wallenius (1986) have shown that the unequal outcome occurrence of independent probabilities usually leads to overdispersed data (they referred to this characteristic as a convex relationship between mean and variance). They examined 24 datasets and, 23 of those showed a convex relationship. The same characteristic has been observed with crash datasets (Abbess et al., 1981; Poch and Mannering, 1996; Hauer, 1997).

The main characteristics of Eq. (2.2) dictate that all the crash-frequency and crash-severity models described in the next two chapters are used to approximate the Poisson process with an unequal probability of events and overdispersion in most cases. So far, the "real" process is not known to safety analysts or researchers. More details are provided in Lord et al. (2005).

2.4 Sources of data and data collection procedures

To quantify safety, that is estimating the safety performance of entities or measuring the safety effects of countermeasures, data need to be collected and analyzed. Unfortunately, in highway safety, the "best" sources of data usually involve collecting data from crashes that have already occurred (note: noncrash based data exist and are covered in this chapter and elsewhere in the textbook). This means that the data involved people who have been injured, sometimes fatally; and properties, such as vehicles or roadside objects, that have been damaged. In addition to the pain and suffering, traffic crashes often cause high direct and important societal costs, as covered in Chapter 1—*Introduction*. An example of such costs is illustrated in Table 2.3 for the United States. This table shows the latest comprehensive costs (2018) for people who may be injured in future highway-related projects (or, alternatively, the savings captured by reducing future crashes via improvements in highway design and operational characteristics or the implementation of countermeasures). These values are usually used for calculating the benefit–cost analysis of highway projects.[1]

[1] https://injuryfacts.nsc.org/all-injuries/costs/guide-to-calculating-costs/data-details/.

TABLE 2.3 Comprehensive crash costs (per person) (2018 dollars).

Severity of injuries	Costs
Fatal (K)	$10,855,000
Incapacitating (type A)	$1,187,000
Nonincapacitating (type B)	$327,000
Possible (type C)	$151,000
No injury (property damage only or PDO)	$50,000

Source: NSC.[1]

In this section, we will cover different sources of data used for analyzing safety. Section 2.4.1 describes traditional data that can be utilized for estimating the safety performance of highway entities and countermeasures. Section 2.4.2 covers relatively new sources of data that come from the perspective of a naturalistic driving environment. The last section describes data collected from disruptive technologies, such as those coming from smartphones.

2.4.1 Traditional data

Traditional data can generally be grouped into five broad categories: (1) crash data; (2) roadway data; (3) traffic flow data; (4) supplemental data, and (5) other safety-related data. Although other sources of safety-related data exist, such as citations or traffic conflicts (both briefly covered below), crash data remain the best source of information that can be used for better understanding the safety performance of a system. Ultimately, crashes are events that can truly measure the safety performance. To better understand crashes that occurred on the highway system, we also need to collect data about the characteristics of the highways under study. This includes obtaining information about the physical and operational characteristics of the highway and/or its users. Traffic flow data are primarily used for estimating the level of exposure in the system, which dictates that if no traffic is present, no crashes can occur. The supplemental data refer to data that are collected manually, based on site visits or via tools, such as Google Earth or Streetview, which are not routinely collected by transportation agencies.

2.4.1.1 Crash data

Crash data are the fundamental type of data needed for conducting safety studies. They are usually collected from police reports, although

some transportation or law enforcement agencies could collect self-reported reports from drivers involved in crashes that did not lead to any injuries. Over the last 20–25 years, most agencies have upgraded to providing the data in electronic format or databases (e.g., SAS, DBF, MS Excel). Essentially, these agencies have specifically trained staff who code the data from hard copies that have been filled out by police officers, most of whom were called in at the scene of a crash. Usually, these agencies have a validation process to ensure the data are properly coded. In the United States, state agencies generally collect the data based on the guidelines outlined in the Model Minimum Uniform Crash Criteria[2] to maintain consistency in the numerous variables collected. To be reported in official statistics, crashes need to meet a set of criteria, such as a minimum level of damage (usually around $1200 US, but could vary by state), include at least one injured person or, in some cases, one of the vehicles involved in the crash has to be towed away. A fatal injury is often defined as such if a vehicle occupant dies within 30 days after the crash (caution: this may not be true everywhere and should be validated by the safety analyst). Given the characteristics described earlier, it is important that the safety analyst becomes familiar with the characteristics of the database that will be used for safety analyses.

Table 2.4 lists some of the important crash-related variables that are relevant in analyzing the safety performance of highway entities or users. In the past, police officers would fill out the form manually either at the scene of the crash or at the police station at the end of the shift (based on the notes taken at the scene). In recent years, reports are usually filled out electronically inside the police vehicle.

The electronic crash databases usually include one line per crash, although some could include one line per vehicle involved. Each column contains information for each variable. From past experiences, it is not uncommon to have files that contain more than 100 variables that describe the characteristics of the crashes.

It is important to point out that crash data variables collected by law enforcement agencies are often primarily utilized for determining if the driver(s) involved in a crash will be cited or subjected to criminal charges, especially when one or more fatal injuries occurred or a criminal conduct happened before the crash such as DWI. In many cases, variables that are irrelevant for this goal may not be adequately gathered. Hence, based on the authors' own experience, transportation agencies should establish a very good line of communication with law enforcement agencies to ensure that important variables utilized for determining the safety performance of the highway network are properly collected.

[2] https://www.nhtsa.gov/mmucc-1.

TABLE 2.4 Important variables collected from crash data.

Variable	Description
Identification number	Each crash report should have its own identification number. This ensures that each crash is unique and can be easily traceable.
Location	The location can be identified using a linear system, such as control-section mile point on predefined maps maintained by the transportation agency. More recently, most agencies are now reliably coding crash data using geographic information system (GIS) technology.
Date and time	These two variables can be used to assign crashes for different seasons and whether the crash occurred during nighttime, dusk, dawn, or daytime conditions.
Severity	This is used to characterize the most severe injuries among all the occupants or vulnerable road users (pedestrians or bicyclists). For example, if a crash has three injures, one incapacitated (type A) and two possible injuries (type C), the crash will be classified as incapacitating injury (type A).
Collision type or manner of collision	This variable describes the characteristic of the crash, such as right-angle, side-swipe, or left-turn/through collision.
Direction of travel	This variable explains the direction or trajectory of each vehicle or road user involved in the crash.
Alcohol or drugs	This variable explains if any of the drivers or vulnerable road users was under the influence of alcohol or drugs. This variable will be often be updated in the report after the crash to account for the time needed to get laboratory results back.
Vehicle occupants	This variable describes the gender and age of each vehicle occupant or road user. It may include the legal driving and insurance statuses of drivers.
Vehicles involved	This one describes the characteristics of each vehicle. This variable defines the crash as being a single-vehicle or multivehicle event.

Continued

TABLE 2.4 Important variables collected from crash data.—cont'd

Variable	Description
Narratives	This section of the report is usually not coded electronically. However, the narrative is very important, as it provides information about the crash process (based on the testimony of witnesses and the visual assessment of the officer). It is usually accompanied by one or more figures or sketches that help explain what happened. Based on the authors' experience, many research projects involved the review of these narratives for validating the electronic databases. This is a very time consuming and costly process.

2.4.1.2 *Roadway data*

Roadway data provide information about the design and operational characteristics of highway segments and intersections. In this day and age, transportation agencies also maintain these kinds of data electronically. Table 2.5 lists common variables that are found in these databases.

2.4.1.3 *Traffic flow data*

Traffic data provide information about users traveling on the facilities under study. For segments, the traffic flow represents the number of vehicles, bicyclists, or pedestrians that travel on the segment. For intersections, traffic flow represents the number of vehicles, bicyclists, or pedestrians that enters the intersection. Usually, transportation agencies collect the traffic flow data from manual and automatic counters on their highway network. These data are also available electronically and are usually separated by year (i.e., each year has its own file). It is important

TABLE 2.5 Common variables found in roadway data.

Location (control-section mile point or geographical coordinates)	Highway classification (freeway, arterials, etc.)
Segment length	Type of lane and width
Type of pavement	Type of shoulder and width
Traffic control at intersections	Type of median and width
Speed limit	Number of lanes
Road alignment (tangent, curve)	Divided/undivided
Road surface condition	Lighting
Right-of-way width	
Parking	

TABLE 2.6 Traffic flow data.

Location (control-section mile point or geographical coordinates)
Annual average daily traffic/AADT (vehicles/day)
Average daily traffic/ADT (vehicles/day)
Traffic mix (heavy vehicles, motorcycles, passenger cars, etc.)
Speed distribution
Short counts (hourly volumes, 15-min values, etc.)
Vehicle occupancy (on instrumented urban freeway segments)
Traffic density (on instrumented urban freeway segments)
Turning movements at intersections

to note that many traffic flow data are actually estimated values that are extrapolated from expansion factors based on when and where the traffic counts were performed on the network such as the day of the week and the month of the year. Table 2.6 summarizes key traffic flow variables.

As discussed earlier, for intersections, entering flows are usually used as input variables for the crash-frequency and crash-severity models described in the subsequent chapters. Fig. 2.2 describes how the traffic flows are assigned for each (undivided or single traveled way) leg of a 4-legged intersection (note: 3-legged intersections would work in the same manner). This figure shows values for the annual average daily traffic

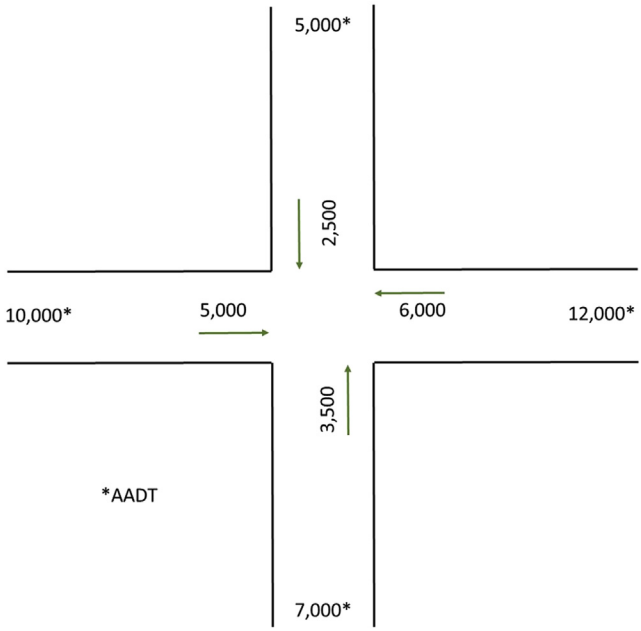

FIGURE 2.2 Entering flows in vehicles per day (AADT).

(AADT) for each street and the entering flow in vehicles per day for each leg. It is basically the AADT divided by two.

2.4.1.4 *Supplemental data*

Although transportation agencies collect a wide amount of data, researchers and safety analysts often need to collect (traditional) data that are not routinely available. On other occasions, supplemental data could be collected for validating the data provided by these agencies. Traffic flow (collected by the analysts themselves) and speed limits are such variables that have been collected for validation purposes. Other supplemental data could include for example:

- Number and types of driveways located in urban or rural areas
- Pedestrian and bicycle traffic flow counts
- Side slope along rural two-lane highways
- Superelevation on horizontal curves
- Deflection angles on horizontal curves
- Pavement friction
- Length of clear zones
- Location of roadside devices (e.g., longitudinal barriers)

Below are useful methods or tools that have been used to provide supplemental data.

Site visits: In many safety-related projects, site visits are commonly performed to collect supplemental data. They could include those achieved for collecting pedestrian and bicycle counts at urban or suburban intersections. Site visits can also include the use of specialty-equipped vehicles that can collect on- and off-road data. Fig. 2.3 shows, for example, a screenshot of the Dewetron data collection system. This system can measure the lateral clearance measured by LiDAR (top right corner of the screen). It also includes a video signal recorded by the cameras as shown in the middle of the screen (forward camera on the left and side camera on the right) and allows the recording of the roadway profile created by the GPS signal, as shown at the bottom right.

Google Earth: This is a powerful mapping service software program. This program offers satellite views of basically any location around the earth. It can provide aerial views for several years, distance measurements, horizontal curve radius measurements, and the location of railway lines among others. The program can be used to collect variables such as driveway densities, validate lane and shoulder widths, turning radii at intersections, etc. Fig. 2.4 shows a typical satellite image view in Google Earth.

Street View: This tool is attached to Google Earth and allows the user to see the highway from the driver's perspective. The tool can be used to record the location and type of traffic control devices (such as signal type

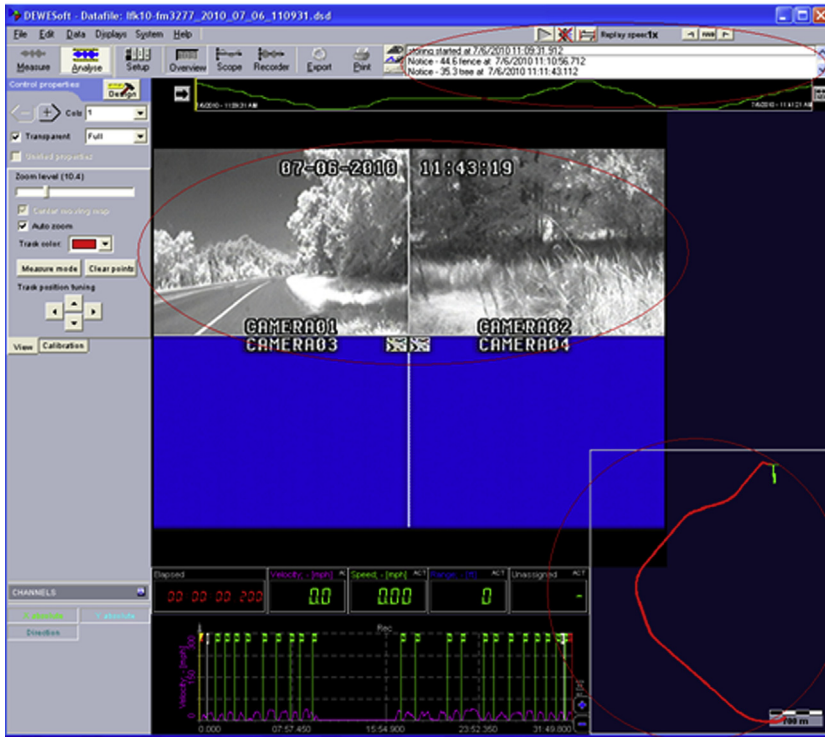

FIGURE 2.3 Screenshot of the Dewetron data collection system (Lord et al., 2011).

FIGURE 2.4 Satellite view of a divided rural arterial and the location of driveways. *Image Credit: Google Earth Mapping Service.*

FIGURE 2.5 Image similar to Google's street view

or rumble strip presence), the location of roadside objects, and severity of sideslopes, etc. Streetview has been useful for collecting supplemental data without having to conduct site visits. Fig. 2.5 shows an image from the driver's perspective similar to typical images available in Street View (Google does not allow showing their Street View images in textbooks).

Video Recording and Processing: Over the last 5–10 years, video recording has become increasingly useful for collecting safety data. Video recording has usually been utilized for collecting traffic conflicts (see Chapter 11—*Surrogate Safety Measures*) and driver or pedestrian behavior in urban environments (say crossing paths at intersections) as well as on fully instrumented urban freeways (e.g., closed circuit TV). Usually, hours of videos are recorded and then these videos are manually processed in a laboratory or back at the analysis center. Fig. 2.6 shows cyclist motion patterns at two intersections in Montreal, Canada, which were collected from a video recording process.

2.4.1.5 Other safety-related data and relevant databases

As part of traditional data, other categories of data are used for analyzing safety. Some of these data include:

- Citation records
- Hospital data (note: in the United States, there are some privacy issues that impede on obtaining these kinds of data)
- Driver data (from governmental licensing agencies)
- Land-use
- Demographics and population statistics
- Traffic conflicts (discussed in Chapter 11)

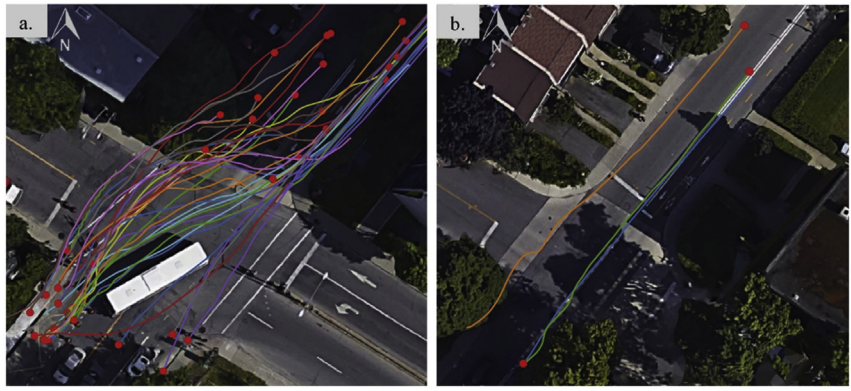

FIGURE 2.6 Cyclist motion patterns at two intersections in Montreal, Canada (Niakia et al., 2019).

- Microsimulation output (conflicts, jerk, deceleration rate, etc.)
- Precipitation data (from National Oceanic and Atmospheric Administration)
- Pavement friction data (from Pavement Management Information System)
- Vehicle registration and driver records (from state department of motor vehicles).

The data described earlier are not directly based on crashes that occurred on the system, with the exception of hospital data to some degree. However, these kinds of data have been used in the past to measure, for instance, the safety risk of road users when they are combined with crash data (e.g., regional-based crash-frequency models).

Table 2.7 shows a list of potential databases that can be used for collecting data and conducting safety analyzes. Some of these databases are available to the public, while others are only available for governmental employees, researchers or may require special permission to access the data. In many cases, the person requesting the data needs to fill out forms before accessing them. The list, presented in Table 2.7, includes some of the publically available safety databases, the name of the agency and region or country. Although somewhat old right now, Montella et al. (2012) have evaluated different crash databases around the world.

2.4.2 Naturalistic driving data

Naturalistic driving data are data that come from instrumented vehicles in which drivers are given no special instructions about how they should drive nor are outside observers present when they travel in the

TABLE 2.7 Sample of national and regional public databases.

Database	Agency	Region
Highway Safety Information System (HSIS) http://www.hsisinfo.org/	FHWA	USA
Fatality Analysis Reporting System (FARS) https://www-fars.nhtsa.dot.gov/Main/index.aspx	NHTSA	USA
National Automotive Sampling System (NASS) https://www.nhtsa.gov/national-automotive-sampling-system-nass/nass-general-estimates-system	NHTSA	USA
General Estimates System (GES) https://www.nhtsa.gov/national-automotive-sampling-system-nass/nass-general-estimates-system	NHTSA	USA
Crashworthiness Data System (CDS) https://www.nhtsa.gov/national-automotive-sampling-system/crashworthiness-data-system	NHTSA	USA
Crash Outcome Evaluation System (CODES) https://www.nhtsa.gov/crash-data-systems/crash-outcome-data-evaluation-system-codes	NHTSA	USA
Bureau of Transportation Statistics https://www.bts.gov/content/motor-vehicle-safety-data	BTS	USA
Mobility and Transport—Road Safety https://ec.europa.eu/transport/road_safety/specialist/statistics_en	European Commission	EU
Statistics Norway https://www.ssb.no/en/transport-og-reiseliv/statistikker/vtu	Gov Norway	NO
Transport Analysis https://www.trafa.se/en/road-traffic/road-traffic-injuries/	Gov Sweden	SE

TABLE 2.7 Sample of national and regional public databases.—cont'd

Database	Agency	Region
CBS Open Data Online (The Netherlands) https://opendata.cbs.nl/statline/portal.html?_la=en&_catalog=CBS&tableId=81452ENG&_theme=1160	CBS	NL
Open Data Portal https://www.data.qld.gov.au/dataset/crash-data-from-queensland-roads https://www.webcrash.transport.qld.gov.au/webcrash2	Gov Queensland	AU
National Collision Database Online https://wwwapps2.tc.gc.ca/Saf-Sec-Sur/7/NCDB-BNDC/p.aspx?l=en&l=en	Transport Canada	CA
Road Accidents—OECD https://data.oecd.org/transport/road-accidents.htm	OCED	EU
Road Safety Statistics http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/	DGT	ES
Statistics of Traffic Accidents in Kaohsiung City https://data.gov.tw/dataset/127489	Government of Taiwan	TW
A2 Road Traffic Accident in New Taipei City https://data.gov.tw/dataset/125657	Government of Taiwan	TW

vehicle. This means that the data collection procedure is considered unobtrusive (Dingus et al., 2006). The goal is to collect data in a "natural" environment with the hope that the instrumented vehicles do not influence the behavior of drivers. In most cases, the owner of the vehicle agreed to have their vehicle equipped with all sorts of sensors that measure forces that act of the vehicle, cameras looking at both inside and outside the vehicle, LiDAR(s) that can measure distances and the relative speed with other vehicles and fixed objects, and a GPS unit that locates the vehicle on the highway system (after the vehicle moves away from home for privacy reasons).

The first study on this topic was known as the 100-car naturalistic study that examined drivers located in the Northern Virginia–Washington, D.C. metro area in the early 2000s (Dingus et al., 2006). The data were collected over a year-and-a-half time period. Approximately 2,000,000 vehicle-miles of driving and about 43,000 h of data were recorded. This study was initially designed as a pilot program to evaluate the data collection procedure, study design methodologies, and potential tools for analyzing the data for a future larger-scale study.

The 100-car study eventually led to the much larger Strategic Highway Research Program (known as SHRP 2[3]) Naturalistic Driving Study. As stated by the National Academy of Sciences (NAS[4]):

"The central goal of the SHRP 2 Safety research program was to address the role of driver performance and behavior in traffic safety. This included developing an understanding of how the driver interacts with and adapts to the vehicle, traffic environment, roadway characteristics, traffic control devices, and the environment. It also included assessing the changes in collision risk associated with each of these factors and interactions. This information will support the development of new and improved countermeasures with greater effectiveness."

The study involved more than 3000 vehicles located in six cities across the United States. The data file contained approximately 35 million vehicle miles, 5.4 million trips, 2705 near-crashes, 1541 crashes, and more than one million hours of video (NAS, 2014). That study also included detailed roadway data collected on 12,538 centerline miles of highways in and around the study sites, which could be matched with the drivers who traveled on these segments.

Other naturalistic driving research that has been performed across the world includes the UMTRI naturalistic driving study,[5] the UDRIVE European naturalistic driving study,[6] and the Australian Naturalistic Driving Study.[7] Some of the data can be available to the public and researchers, but the users may need special permissions such as the approval by a researcher's Institutional Review Board.

Naturalistic driving studies provide unique data that can help study driving behavior in a natural environment. Unfortunately, the amount of

[3] https://www.shrp2nds.us/index.html.

[4] http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/SHRP2DataSafetyAbout.aspx.

[5] http://www.umtri.umich.edu/our-focus/naturalistic-driving-data.

[6] https://results.udrive.eu/.

[7] http://www.ands.unsw.edu.au/.

data collected can be extremely large (these types of data are also known as "Big Data"). For example, the total collected data for SHRP-2 have required approximately 1.5 Petabyte (PB) of archival storage, 700 Terabyte (TB) of parametric data (sensors, etc.), and over 1.2 PB of video storage (NAS, 2014). The entire database contains millions of files. Analyzing such datasets can be very challenging as traditional analytical tools, such as traditional crash-frequency or crash-severity models, are not usually adequate. Alternative tools are therefore needed for extracting information from extremely large datasets. Chapter 12—*Data Mining and Machine Learning Techniques* provides data-driven tools that can be used for properly analyzing datasets categorized as Big Data.

2.4.3 Disruptive technological and crowdsourcing data

In general terms, disruptive technology refers to an innovation that significantly alters the way that consumers, industries, or businesses operate, which usually has much better attributes than the older technology (Smith, 2020). Smartphones are such technology (as compared to regular cell phones) (Appiah et al., 2019). Smartphones can provide a large amount of data that could be used for safety analyzes. They include deceleration rates, locations of crashes, where and when a driver was texting and driving, and traveled speed of a vehicle on the highway network among others (see, e.g., Bejani and Ghatee, 2018; Kanarachosa et al., 2018; Stipancica et al., 2018). Such data are collected by third-party companies or vendors (who buy data directly from cellphone service providers) and sell them to private and public agencies as well as private citizens/researchers. *Streetlight Data* and *Safe 2 Save LLC* are examples of such companies.

Crowdsourcing is defined as obtaining data or information from a large group of people. Over the last few years, some have used crowdsourced data for analyzing highway safety. Data can be collected from social media platforms such as Twitter, Reddit, Waze, Instagram, or Facebook. Similar to the disruptive technological data, third-party vendors extract data from these platforms (often with a fee) and sell them to public and private agencies or anyone else who request the data. These data could be used to validate police reports data or identify crash data not commonly collected by transportation agencies because they do not meet the minimum reportability criteria (see, e.g., Flynn et al., 2018; Goodall and Lee, 2019; Li et al., 2020).

2.4.4 Data issues

Traditional safety data offer unique characteristics that are not found in other types of data, such as those related to crime (Levine, 2008) or power failures caused by hurricanes (Guikema et al., 2014). Some of these issues are attributed to the huge cost associated with collecting crash and other related data (Lord and Bonneson, 2005; Lord and Mannering, 2010). Safety analysts should be made aware of these issues as they could negatively influence the performance of crash-frequency and crash-severity models. For a full description of these issues, the reader is referred to Chapter 6—*Cross-Sectional and Panel Data in Safety* as well as Lord and Mannering (2010) and Savolainen et al. (2011).

2.5 Assembling data

After a crash is reported, the law enforcement agencies investigate and complete the crash report with all factual information. If some information is unknown (such as driver distraction), law enforcement officials use their best judgment and record their considered opinions based on their investigation. As they collect an extensive list of variables, the data are usually stored in different electronic files such as crash, vehicle, driver/passenger, citation/adjudication, and EMS/Injury Surveillance. Each of these data electronic files contains a unique identifier (such as crash ID or number) with which they can be combined.

As described earlier, crash data may not necessarily provide a complete picture about the safety performance of an entity. It must be therefore combined with other data sources (such as traffic data, road inventory data, and vehicle registrations) for further investigation and analysis. For example, the main reason for an abundance of roadway departure crashes could be attributed to a narrow shoulder, a sharp curve, or fixed objects located within the clear zone, and it may not be possible to know this unless the crash data are combined with roadway and roadside data. The process of combining or merging multiple data sources is often called data assembly or integration. Deterministic and probabilistic integrations are the two most common procedures used when assembling traffic safety data.

Most of the databases provided by the state agencies use linear referencing, which allows roadway and traffic attributes to be stored individually and to be defined by the route, along with the start and end reference markers (such as mileposts). The primary advantage of linear

referencing is that it allows for a very detailed delineation of features along a route without breaking the route into very small segments. Crashes are also referenced using the same system. Therefore, crashes can be merged using location code to exactly match records belonging to the same point with the roadway. This kind of data assembly is called deterministic integration because it relies on common elements shared among the datasets to make exact matches. Many agencies have already started identifying different data elements in a geographic information system (GIS) environment (using latitude and longitudes). This makes the integration of databases attractive as map-based interface can effectively present and interpret data.

In some instances where an exact match cannot be conducted using unique identifiers, probabilistic integration is often used that relies on similar elements and values shared among the datasets to make matches. For instance, crash records are often required to be linked to data collected by emergency medical services, hospital emergency department, and hospital admission and discharge files. All these files are subjected to strict confidentiality rules and regulations, which prevents merging databases by the name of the vehicle occupants who were injured and hospitalized for example. The probabilistic linkage uses common fields between databases such as incident longitude/latitude, date, age/date of birth (if available), time of admission at the hospital, and seat position among others.

2.6 4-stage modeling framework

This section describes a general 4-step modeling process that can be used for developing statistical models for crash data analyses. These steps are applicable for crash-frequency models (*Chapter 3*) and crash-severity models (*Chapter 4*). In the safety literature, statistical models have often been defined or called safety performance functions (SPFs) and crash prediction models (CPMs). The former is used in the AASHTO's Highway Safety Manual (HSM) (AASHTO, 2010) and publications that are associated with the manual. In this textbook, we refer to all the models as statistical models, either crash-frequency or crash-severity models, unless the methodology is specifically tied to the HSM. In this case, the model may be referred to as an SPF or a CPM.

2.6.1 Determine modeling objective matrix

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, etc. as described in Chapter 3—*Crash-Frequency Modeling*), 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. The latter characteristics are described in greater detail in Chapter 5—*Exploratory Analyses of Safety Data*.

Table 2.8 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. For ramps, only crash-frequency models will be developed.

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. For instance, it is important to have a 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) to be assigned to each observational unit. These units may be associated with the highway network, but could also be related to the analyses of drivers, vehicle occupants, or pedestrians. The range of traffic flows can be used as another example. There is a need to determine the range of flows and geometric characteristics of interest to this study (e.g., AADT = 200−20,000) and make sure commensurable data can be obtained and enough data can be collected. The time unit of analysis (i.e., number of crashes per unit of time) is another critical element when developing statistical models. 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. As discussed in Chapter 3—*Crash-Frequency Modeling in Safety* (and in Lord and Geedipally, 2018), using a small-time or space scale could "artificially" increase the proportion of zero reponses in the dataset. This could lead to erroneously selecting an inappropriate statistical model for analyzing such datasets.

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

2.6.2 Establish appropriate process to develop models

This step is used to ensure the best possible statistical models be developed to achieve the modeling objectives identified in the previous step. This includes ensuring that (1) data sources and limitations, sampling design, and statistical, functional, and logical assumptions are clearly spelled out; (2) supporting theories are properly defended and/or cited (i.e., goodness-of-logic, as discussed by Miaou and Lord, 2003); (3) models are systematically developed and tested; and (4) modeling results are properly interpreted.

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.

2.6.3 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). As will be discussed in the next section, more detailed inferential goals will require more sophisticated computational methods to fully capture the sampling variations in producing estimates and predictions.

Fig. 2.7 shows an example of a posterior distribution generated by the WinBUGS software (Lunn et al., 2000) for the inverse dispersion parameter of an NB model developed from 868 signalized 4-legged intersections in Toronto, ON using the Bayesian estimation method. The posterior mean, standard deviation, and median of the distribution were 7.12, 0.67, and 7.03, respectively. The inverse dispersion parameter (point estimate) that was originally estimated using the MLE was 7.15 (0.63) for the same dataset (Lord, 2000).

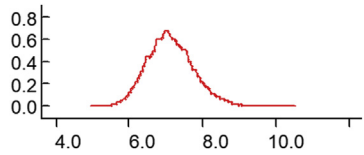

FIGURE 2.7 Posterior distribution for the inverse dispersion parameter.

2.6.4 Select computational techniques and tools

This is the process where Frequentist (analysts who use the likelihood-based method or MLE) (McCullagh and Nelder, 1989), and the Bayesian method (Carlin and Louis, 2008; Gelman et al., 2013) 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. By being able to take advantage of the unprecedented computing power available today, simulation-based methods, including various bootstraps and Markov Chain Monte Carlo (MCMC) methods (Gilks et al., 1996), have been particularly popular in the statistical community over the last 20 years, regardless of whether the MLE or Bayesian estimating method is considered. The characteristics of the likelihood-based and the Bayes methods are described later. Note that in the highway safety literature, crash-frequency and crash-severity models estimated using the Bayes method are often called a "Full" Bayes (FB) model (Miaou and Lord, 2003). The terminology is used to distinguish models estimated using the Bayes method from techniques that employed the empirical Bayes (EB) method. The EB method is covered in Chapter 7—*Before-After Studies in Safety* and Chapter 8—*Identification of Hazardous Sites*.

2.6.4.1 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. For example, when crash data are assumed to follow an NB distribution, the most popular

model used in highway safety (Lord and Mannering, 2010) and details can be found in Appendix A, the likelihood is as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} NB(\mathbf{y}_i; \boldsymbol{\beta}) \tag{2.3}$$

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 parameters can be obtained by maximizing the likelihood by using the Newton−Raphson or other search techniques, such as the Fisher scoring (McCullagh and Nelder, 1989). Generally, the likelihood function is not directly optimized for efficiency reasons. Instead, its logarithm, called log-likelihood (LL) is preferred. As the likelihood function, under independent and identically distributed (i.i.d) assumption, is the product of the sampling distributions of individual data points (as shown in Eq. 2.3), taking the logarithms converts the product into sums of the log densities. Sum of logs is numerically more stable than the log of the products.

The Hessian of the LL obtained at the point of convergence is often used to report the standard errors of the parameters and perform model selection. Appropriate distributional assumptions have to be satisfied, such as the asymptotic normality of the test statistics, for performing valid hypothesis testing on the parameters estimated by the MLE method.

In the above simplified model, besides the crash history y_i, no other data are used. However, typically, crash counts from several observation periods (e.g., years) are collected over presumably static roadway, intersection, or other entity conditions. It is fairly trivial to incorporate such additional meta-data into the modeling process, by assuming that the mean response depends on that additional meta-data. That is, the conditional mean is considered an unknown function of the covariates, as given in the following equation:

$$E(y_i|\boldsymbol{\beta}) = \mu_i = f(\mathbf{x}_i'\boldsymbol{\beta}) \tag{2.4}$$

The application of the MLE method is now commonly implemented in all commercially available statistical programs, such as SAS, Genstat, R, Python, and STATA. The theory associated with the generalized linear models, the foundation for crash-frequency models, is also well covered in seminal textbooks on this topic (Cameron and Tridevi, 2013; Hilbe, 2011, 2014). For crash-severity models, the theory is also well covered in

Train (2009). Appendix A shows how the characteristics of the MLE is used for estimating NB models.

If panel or longitudinal data are used (data collected over time as discussed in Chapter 6—*Cross-Sectional and Panel Studies in Safety*), there is a strong probability that the observations will be correlated in time. In other words, the same observation is measured at different points in time (note that some researchers have labeled such datasets as data with repeated measurements). It has been shown that this kind of dataset will most likely create a temporal correlation that will influence the inferences associated with the parameter estimates (Diggle et al., 1994). To handle the temporal or serial correlation, the generalized estimating equations (GEE) has been proposed for handling panel data. The GEE is also a likelihood-based method (more specifically, quasi-likelihood, which does not require the assumption of normality), but is designed specifically to handle the temporal correlation. Generally, the safety analyst needs to assess the covariance matrix, which can be either defined as Independent, Exchangeable, Auto-Regressive Order One or Unstructured. The models are estimated using quasi-likelihood estimators via an iterative process (note: quasi-likelihood equations are also called generalized estimating equations, hence the name GEE). Not accounting for the serial correlation often underestimates the standard errors of the parameters (Lord and Persaud, 2000; Hardin and Hilbe, 2013).

2.6.4.2 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

$$\begin{aligned} P(\mu|y) &\propto P(\mu)P(y|\mu) \\ &\equiv \frac{P(y)}{P(\mu)P(y|\mu)} \end{aligned} \tag{2.5}$$

where $P(\mu)$ is the prior distribution of the parameters, $P(y \mid \mu)$ is the likelihood function, and $P(\mu \mid y)$ is the posterior distribution. The normalization constant $P(y)$, which ensures that the posterior has a valid density, is only a function of the data. Another way to think about the Bayesian paradigm is that the prior belief is updated in light of evidence (collected

via the data) resulting in the posterior belief. It is the posterior belief that is of interest for inference. Inference is typically carried out by generating approximate samples from the posterior density using MCMC techniques. When only point estimates are sufficient or computational time is of concern, it is not uncommon to use Variational Inference or MAP estimates. MAP is estimated as the MLE analogy in the Bayesian setting, where the posterior distribution is maximized instead of the likelihood.

Models elicited under the Bayesian paradigm are actually 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:

$$\text{(i)}\quad y_i \big| \omega_i \sim \text{Poisson}(\omega_i) \rightarrow y_i \big| \omega_i \sim \text{Poisson}(\mu_i e^{\varepsilon_i}) \tag{2.6a}$$

$$\text{(ii)}\quad e^{\varepsilon_i} \big| \eta \sim \pi_\varepsilon(\eta) \tag{2.6b}$$

$$\text{(iii)}\quad \eta \sim \pi_\eta(\cdot) \tag{2.6c}$$

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

Various prior choices can be considered for modeling the parameters e^{ε_i} and η. Depending on the specification of the priors $\pi_\varepsilon(\cdot)$ and $\pi_\eta(\cdot)$, different alternative hierarchical models can be defined. For the hierarchical NB (HNB) model, we specify a gamma prior on e^{ε_i} with shape (a) and scale (b) parameters to be equal. This leads to the following error function:

$$e^{\varepsilon_i} | \varphi \tilde{} gamma(\varphi, \varphi) \text{ and } \varphi \tilde{} gamma(a, b) \tag{2.7}$$

Instead of assuming a gamma distribution as a prior distribution for e^{ε_i}, the lognormal distribution can be used as an alternative function. With this prior choice, the hierarchical Poisson-lognormal model, another model very popular in highway safety, is derived by assuming a proper hyper-prior for the parameter σ^2 such that (Lord and Miranda-Moreno, 2008):

$$\varepsilon_i = \log(e^{\varepsilon_i}) \big| \sigma^2 \tilde{} Normal(0, \sigma^2) \text{ and } \sigma^{-2} \tilde{} gamma(a, b) \tag{2.8}$$

The choice of prior for the parameter σ^2 relies on the fact that a conjugate distribution of the Normal distribution is the Inverse-gamma.

Convenient priors are conjugate distributions that produce full conditional posteriors of the same form. Furthermore, the hyper-prior parameters a and b have fixed values and must be specified by a safety analyst.

Different from the MLE, which subscribes to the Frequentist paradigm, inference in the Bayesian setup is easy to interpret, as every quantity of interest is a probability statement—case in point being credible intervals versus confidence intervals that quantify the uncertainty of the model parameters. As the complete joint distribution of all model parameters is available, any question concerned with the parameters, expressed as a functional of the parameters, can be routinely obtained. However, care must be taken in both model elicitation, of which prior specification is a big component, and performing extensive checks on the inference technique. For example, when a vague or noninformative hyper-prior is for defining the model's parameters, the posterior estimates (MAP) will be similar to the estimates provided by the MLE. Another advantage of the Bayes method is that information extracted from previous studies can be used to refine the hyper-priors (see, e.g., Heydari et al., 2013). Appendix A documents the characteristics for estimating HNB using the Full Bayes method.

For the development of crash-frequency models, the Bayes method is preferred when crash data are characterized by low sample mean values and small sample size (see Chapter 3—*Crash-Frequency Modeling*) (Lord and Miranda-Moreno, 2008). In such instances, informative hyper-priors based on prior knowledge, as obtained from the previous studies, can be used. Some studies have developed FB models by accounting for temporal and/or spatial correlations (Huang et al., 2009; Jiang et al., 2013). Fawcett et al. (2017) proposed a novel FB hierarchical model that incorporated crash counts from multiple past periods rather than from a single before period in the identification of hazardous locations. These authors used a discrete-time indicator in the model to account for the effects of a temporal trend in crashes.

2.7 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). More details about this topic are covered in the Chapter 3 - Crash-Frequency Models and Chapter 6 - Cross-Sectional Studies.

The GOF methods can be classified into two general groups. One group of methods relates to likelihood statistics, while the other group assesses the model performance based on the model's errors. It is suggested to use several GOF methods from both groups to assess the performance of different models. It should be pointed out that likelihood-based methods should compare models that are estimated using the same dataset.

2.7.1 Likelihood-based methods

The methods presented in this section describes how well the model maximizes the likelihood function, with different parametrizations or with different penalty functions. Most of these likelihood-based methods can be used either for crash-frequency and crash-severity models. Although the basic equations are described here, all these methods can be calculated automatically (predefined functions/modules or written codes) in statistical software programs.

2.7.1.1 Maximum likelihood estimate

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. Simultaneously solve all these partial derivative equations to find the parameters and put all them back in the log-likelihood function. Algorithms, such as the Newton–Raphson search algorithm (second-order conditions), need to be used to solve these equations. Fortunately, all statistical computer programs can now automatically estimate the maximum log-likelihood or MLE estimate. The MLE is given as follows:

$$\text{MLE} = -2 \times LL \tag{2.9}$$

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. Other methods below can overcome this problem. Appendix A shows how to calculate the MLE for the NB model.

2.7.1.2 Likelihood ratio test

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

$$LRT = -2[LL(\boldsymbol{\beta}_U) - LL(\boldsymbol{\beta}_R)] \tag{2.10}$$

where $LL(\boldsymbol{\beta}_R)$ is the log-likelihood at the convergence of the "restricted" model and $LL(\boldsymbol{\beta}_U)$ is the log-likelihood at the convergence of the "unrestricted" model. Larger values indicate a better fit.

2.7.1.3 Likelihood ratio index

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 ρ^2 statistic or sometimes just, ρ. The estimation of potentially insignificant parameters is accounted for by estimating a corrected ρ^2 as where p is the number of parameters estimated in the model. The formulation is:

$$\rho^2 = 1 - \frac{LL\left(\widehat{\boldsymbol{\beta}}\right)}{LL(0)} \tag{2.11a}$$

$$\rho^2_{corrected} = 1 - \frac{LL\left(\widehat{\boldsymbol{\beta}}\right) - p}{LL(0)} \tag{2.11b}$$

Where $LL(\widehat{\boldsymbol{\beta}})$ is the log-likelihood function at the estimated parameter $\widehat{\boldsymbol{\beta}}$; and, and $LL(0)$ is the log-likelihood function where the parameters are set to zero. Therefore, this index ranges from zero (when the estimated parameters are no better than zero, not optimal estimates) to one (when the estimated parameters perfectly predict the outcome of the sampled observations). The name ρ^2 is somewhat similar to the R^2 statistic. However, while R^2 indicates the percentage of the variation in the dependent variable that can be "explained" by the estimated model, ρ^2 is the actual percentage increase in the log-likelihood function above the value taken at the zero parameter.

Unfortunately, the meaning of such an increase is unclear in terms of the power of the model explanation. When comparing two models estimated using the same set of data with the same set of alternatives (the premise for model comparison so that $LL(0)$ is the same for both models), the model with the higher ρ^2 fits the data better. This is the equivalent of saying that the model with a higher value of the likelihood function is preferable.

2.7.1.4 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. The AIC is computed as follows:

$$AIC = -2 \times LL + 2p \tag{2.12}$$

where p is the number of unknown parameters included in the model (this also includes the dispersion or shape parameters of models, such as the inverse dispersion parameter of the NB model or the random spatial effect) LL. Smaller values indicate better model fitting.

2.7.1.5 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. It is computed the following way:

$$AIC = -2 \times LL + p \ln n \tag{2.13}$$

Like the AIC, smaller values indicate better model fitting.

2.7.1.6 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. The DIC is defined as follows:

$$DIC = \widehat{D} + 2\left(\overline{D} - \widehat{D}\right) \tag{2.14}$$

where $\overline{D}$ is the average of the deviance ($-2 \times LL$) over the posterior distribution, and $\widehat{D}$ is the deviance calculated at the posterior mean parameters. As with the AIC and BIC, the DIC uses $p_D = \overline{D} - \widehat{D}$ (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 (Spiegelhalter et al., 2002).

2.7.1.7 Widely applicable information criterion

The widely applicable information criterion (WAIC) (Watanabe, 2010) is a measure that is similar to the DIC (i.e., adds a penalty term for minimizing overfitting), but incorporates the variance of individual terms (the D s in Eq. 2.12). 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.

2.7.1.8 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:

$$\log\left\{p(\mathbf{y}|M)\right\} \approx \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\{|\mathbf{H}^*|\} + \log\{f(\mathbf{y}|\boldsymbol{\beta}^*)\}\Big\} + \log\{\pi(\boldsymbol{\beta}^*)\} \tag{2.15}$$

where p is the number of parameters, $\log\{f(\mathbf{y}|\boldsymbol{\beta}^*)\}$ is the log-likelihood of data at $\boldsymbol{\beta}^*$, and $\log\{\pi(\boldsymbol{\beta}^*)\}$ is the log-likelihood of prior distribution at $\boldsymbol{\beta}^*$. One way for estimating $\boldsymbol{\beta}^*$ is to find the value of $\boldsymbol{\beta}$ at which $\log\{f(\mathbf{y}|\boldsymbol{\beta}^*)\} + \log\{\pi(\boldsymbol{\beta}^*)\}$ achieves its maximum from the posterior simulation output. $|\mathbf{H}^*|$ is the determinant of the variance–covariance matrix estimated from the Hessian at the posterior mode, and is asymptotically equal to the posterior variance–covariance matrix. This can be estimated from the sample variance–covariance matrix of the posterior simulation output. Assuming that the prior probabilities for the competing models are equal, B_{12} is expressed as follows:

$$\log\{B_{12}\} = \log\{p(\mathbf{y}|M_1)\} - \log\{p(\mathbf{y}|M_2)\} \tag{2.16}$$

According to Kass and Raftery (1995), values between 20 and 150 strongly support the selection of Model 1 over Model 2.

2.7.1.9 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:

$$D(\mathbf{y}, \mathbf{u}) = 2\left\{LL(\mathbf{y}) - LL\left(\widehat{\boldsymbol{\mu}}\right)\right\} \tag{2.17}$$

Smaller values mean that the model fits the data better. This GOF measure applies only to models with a defined likelihood function. As opposed to the DIC, the measure does not include a penalizing function.

2.7.2 Error-based methods

There are methods for estimating how well the model fits the data that are based on minimizing the model's errors (i.e., the difference between the observed and estimated values). The following methods can be

applied for the entire dataset after the model is fitted to the data or when the dataset is split in different proportions (say the model is first estimated with 70% of the data and applied to the rest of the data). The first four methods have been proposed by Oh et al. (2003) to evaluate the fit of crash-frequency models. In addition, most of statistical software programs have modules available for calculating the measures described in the following.

2.7.2.1 Mean prediction bias

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

$$MPB = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - y_i) \tag{2.18}$$

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

2.7.2.2 Mean absolute deviation

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

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\mu_i - y_i| \tag{2.19}$$

Smaller values are better.

2.7.2.3 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:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - y_i)^2 \tag{2.20}$$

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

2.7.2.4 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:

$$MSE = \frac{1}{n-p}\sum_{i=1}^{n}(\mu_i - y_i)^2 \tag{2.21}$$

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.

2.7.2.5 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 measured as a percentage. The MAPE is calculated using this equation:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \times 100 \tag{2.22}$$

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.

2.7.2.6 Pearson Chi-square

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

$$Pearson - \chi^2 = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{VAR(y_i)} \tag{2.23}$$

If the mean and the variance are properly specified, then $E\left[\sum_{i=1}^{n}(y_i - \mu_i)^2 \Big/ VAR(y_i)\right] = n$ (Cameron and Tridevi, 2013). Values closer to n (the sample size) show a better fit.

2.7.2.7 Coefficient of determination R_α^2

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

$$R_\alpha^2 = 1 - \frac{\alpha}{\alpha_{null}} \tag{2.24}$$

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

2.7.2.8 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 (Hauer and Bamfo, 1997). 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. 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 (Hauer and Bamfo, 1997) for this purpose:

$$\sigma_i^2 = \sigma^2(n_i) \times \left(1 - \sigma^2(n_i) \,/\, \sigma^2(N)\right) \tag{2.25}$$

where σ_i^2 is the variance at observation/rank i; $\sigma^2(n_i)$ is the cumulative variance at the residual i; and $\sigma^2(N)$ is the cumulative variance for the last observation in the dataset. The last part of the equation incorporates the proportion of the cumulative residual. For the last observation, the variance is equal to zero. The 95% CI can be calculated as follows (1.96~2.0) at observation i:

$$\pm 2 \times \sqrt{\sigma_i^2} \tag{2.26}$$

Although some statistical programs provide CURE plots, they can also be created in a spreadsheet. Table 2.9 presents an example describing how the CURE plot can be calculated in a spreadsheet. The example uses a traffic flow variable (in vehicles per day). The dataset contains 215 observations ranked from the smallest flow to the largest flow. The last two columns apply Eqs. (2.18) and (2.19).

Fig. 2.8 shows the CURE plot using the data shown in Table 2.7. The columns "Cumulative Residual," "Upper CI" and "Lower CI" were used for the figure.

In most cases, the CURE plot does not start nor end at 0 (zero), which may make it difficult to compare different models. To help with the comparison, the plot can be adjusted by proportionally changing the values along the curve to ensure it starts and end at 0.

Using the example shown in Table 2.7, Fig. 2.9 illustrates how the adjustment can be accomplished. The CURE plot starts at −12.4 (for the flow 1542 veh/day) and ends at 71.7 (for the flow 45,685 veh/day). The rate of the red (gray in printed version) line is 0.0019/veh/day. The goal is to adjust the cumulative residual (add or subtract) by the proportion shown inside the triangle. Between 1542 and 8057 veh/day, you add the value to cumulative residual and between 8057 and 45,685 veh/day, you subtract the value. For the first flow value, the adjusted cumulative residual will be $-12.4 + 12.4 = 0$. In addition, this can be accomplished in a spreadsheet.

2.8 Heuristic methods for model selection

The methods described in the previous section are only applicable after the model is applied and fitted to the data. This approach to selecting a model over another could be very time consuming, especially if complex models are evaluated. Lord et al. (2019) have proposed a heuristics

TABLE 2.9 CURE plot calculations.

Rank	Flow	Residuals	Cumulative residuals	Squared residuals	Cumulative squared residuals	Upper CI	Lower CI
1	1,542	−12.4	−12.4	152.6	152.6	24.7	−24.7
2	7,793	−30.0	−42.4	902.1	1,054.7	64.9	−64.9
3	8,425	−29.4	−71.8	864.1	1,918.8	87.6	−87.6
4	9,142	−53.2	−124.9	2,826.6	4,745.4	137.6	−137.6
5	9,474	74.1	−50.9	5,489.3	10,234.7	201.7	−201.7
6	9,856	−37.6	−88.4	1,412.9	11,647.6	215.1	−215.1
…	…	…	…	…	…	…	…
215	45,685	258.3	71.7	66,733.8	1,660,753.9	0.0	0.0

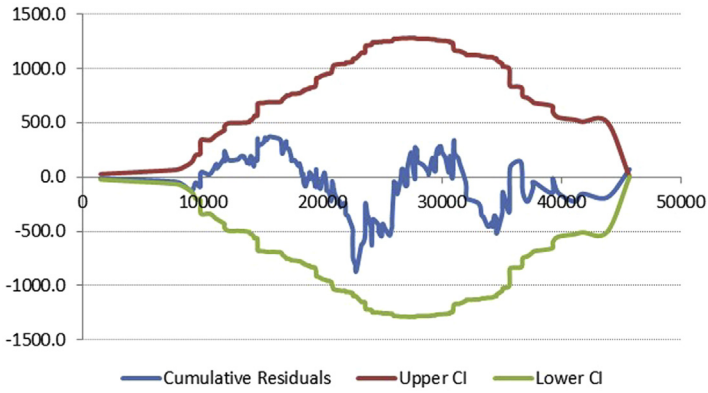

FIGURE 2.8 Cumulative residuals for the data shown in Table 2.7.

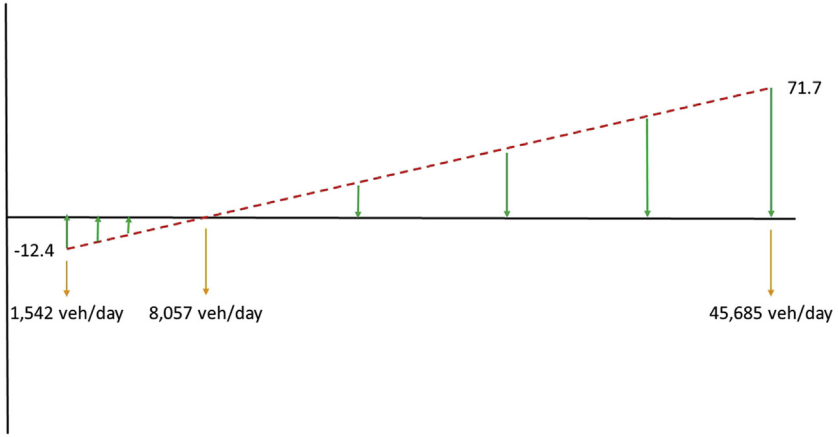

FIGURE 2.9 Adjustment procedure for the cumulative residuals.

method that could be used before the models are fitted and evaluated. This method relies on simulating many datasets from competitive distributions and recording key summary statistics for each dataset. Then, run a Machine Learning classifier (such as Decision Tree or Random Forest) to distinguish one distribution from another. Once the classifier is trained, the descriptive statistics for new datasets could be used to select one distribution over the other (see Shirazi et al., 2017; Shirazi and Lord, 2019). Table 2.10 provides a list of descriptive statistics that could be used for comparing distributions, most of which are described in Chapter 5—*Exploratory Analyses of Safety Data*.

Lord et al. (2019) have already compared the NB with the NB-Lindley (NB-L) and the NB with the Poisson-Lognormal (PLN). The results are

shown in Figs. 2.10 and 2.11, respectively. In Fig. 2.10, if the skewness of the data is greater than 1.92, the NB-L should be selected over the NB. For the NB versus PLN comparison, the safety analyst just needs to follow the tree-diagram for the statistics "percentage of zeros" and "Kurtosis." All these models are described in the next chapter.

TABLE 2.10 Descriptive statistics needed for the heuristics methods.

Descriptive statistics
Coefficient-of-variation
Interquantile (10%–90%, in increments of 10%)
Kurtosis
Mean
Percentage of zeros
Quantile (10%–90%, in increments of 10%)
Range
Skewness
Standard deviation
Variance
Variance-to-mean ratio

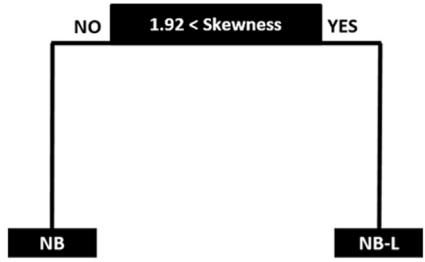

FIGURE 2.10 Heuristics to select a model between the NB and NB-L distributions.

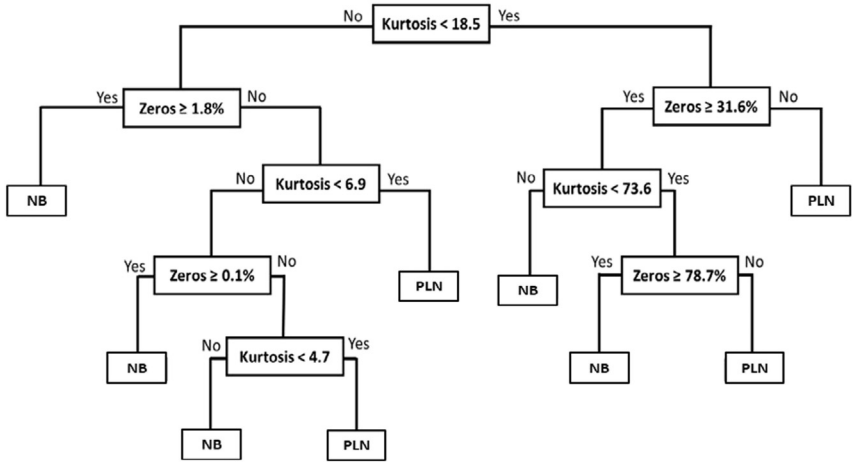

FIGURE 2.11 Heuristics to select a model between the NB and PLN distributions.

References

AASHTO, 2010. Highway Safety Manual, first ed. American Association of State Highway Transportation Officials, Washington, D.C.

Abbess, C., Jarett, D., Wright, C.C., 1981. Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the "Regression-to-Mean" effect. Traffic Eng. Contr. 22 (10), 535–542.

Appiah, D., Ozuem, W., Howell, K., 2019. Disruptive technology in the smartphones industry. In: Book: Leveraging Computer-Mediated Marketing Environments, pp. 351–371. https://doi.org/10.4018/978-1-5225-7344-9.ch017.

Bejani, M.M., Ghatee, M., 2018. A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data. Transport. Res. Part C 89, 303–320.

Cameron, A.C., Tridevi, P.K., 2013. Regression Analysis of Count Data, second ed. Cambridge University Press, Cambridge, U.K.

Carlin, B.P., Louis, T.A., 2008. Bayesian Methods for Data Analysis, third ed. Chapman and Hall/CRC, London, U.K.

Diggle, P.J., Liang, K.-Y., Zeger., S.L., 1994. Analysis of Longitudinal Data. Clarendon Press, Oxford, U.K.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment. In: DOT HS, 810. National Traffic Highway Safety Agency, Washington, D.C.

Fawcett, L., Thorpe, N., Matthews, J., Kremer, K., 2017. A novel Bayesian hierarchical model for road safety hotspot prediction. Accid. Anal. Prev. 99 (Pt A), 262–271.

Feller, W., 1968. An Introduction to Probability Theory and its Application, 3rd Ed., 1. John Wiley, New York, New York.

Flynn, D.F.B., Gilmore, M.M., Sudderth, E.A., 2018. Estimating Traffic Crash Counts Using Crowdsourced Data Pilot Analysis of 2017 Waze Data and Police Accident Reports in Maryland. DOT-VNTSC-BTS-19-01. U.S. DOT, Volpe National Transportation Systems Center, Cambridge, MA.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. Chapman & Hall/CRC Press, Boca Raton FL.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24, 997–1016.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. London: Chapman and Hall.

Goodall, N., Lee, E., 2019. Comparison of Waze crash and disabled vehicle records with video ground truth. Transp. Res. Interdiscip. Perspect. 1, 100019.

Guikema, S.D., Nateghi, R., Quiring, S.M., Staid, A., Reilly, A.C., Gao, M., 2014. Predicting hurricane power outages to support storm response planning. IEEE Access 2, 1364–1373.

Haddon, W., 1980. Options for the prevention of motor vehicle crash injury. Isr. J. Med. Sci. 16 (1), 45–65. https://slideplayer.com/slide/7780000/.

Hardin, J.W., Hilbe, J.M., 2013. Generalized Estimating Equations, second ed. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Hauer, E., 1997. Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd, Oxford.

Hauer, E., Bamfo, J., 1997. Two tools for finding what function links the dependent variable to the explanatory variables. In: Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

Herbel, S., Laing, L., McGovern, C., 2010. Highway Safety Improvement Program Manual. Report No. FHWA-SA-09-029. Federal Highway Administration, Washington, D.C. https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/index.cfm#toc. (Accessed 4 June 2020).

Heydari, S., Miranda-Moreno, L.F., Fu, L., Lord, D., 2013. How to specify priors for full Bayes road safety studies?. In: 4th International Conference on Road Safety and Simulation, Rome, Oct. 23rd–25th, 2013.

Hilbe, J.M., 2011. Negative Binomial Regression, second ed. Cambridge University Press, Cambridge, U.K.

Hilbe, J.M., 2014. Modelling Count Data. Cambridge University Press, Cambridge, U.K.

Huang, H., Chin, H.C., Haque, M.M., 2009. Empirical evaluation of alternative approaches in identifying crash hot spots. Transport. Res. Rec. 2103, 32–41.

Jiang, X., Huang, B., Zaretzki, R.L., Richards, S., Yan, X., 2013. Estimating safety effects of pavement management factors utilizing Bayesian random effect models. Traffic Inj. Prev. 14 (7), 766–775.

Kanarachosa, S., Christopoulosa, S.-R.G., Chroneos, A., 2018. Smartphones as an integrated platform for monitoring driver behaviour: the role of sensor fusion and connectivity. Transport. Res. Part C 95, 867–882.

Kass, R.E., Raftery, A.E., 1995. Bayes factors and model uncertainty. J. Am. Stat. Assoc. 90, 773–795.

Levine, N., 2008. In: Shekhar, S., Xiong, H. (Eds.), CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents. Encyclopedia of Geographic Information Science. Springer, pp. 187–193.

Lewis, S.M., Raftery, A.E., 1997. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. J. Am. Stat. Assoc. 92, 648–655.

Li, X., Dadashova, B., Turner, S., Goldberg, D., 2020. Rethinking Highway Safety Analysis by Leveraging Crowdsourced Waze Data. Presented at the 99th TRB Annual Meeting. Washington, DC.

Lord, D., 2000. The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models. Ph.D. Dissertation. Department of Civil Engineering, University of Toronto, Toronto, Ontario.

Lord, D., Bonneson, J.A., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transp. Res. Rec. 1908, 88–95.

Lord, D., Geedipally, S.R., 2018. Safety prediction with datasets characterised with excess zero responses and long tails. In: Lord, D., Washington, S. (Eds.), Safe Mobility: Challenges, Methodology and Solutions (Transport and Sustainability, vol. 11. Emerald Publishing Limited, pp. 297–323.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transport. Res. Part A 44 (5), 291–305.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Saf. Sci. 46 (5), 751–770.

Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations (GEE) procedure. Transp. Res. Rec. 1717, 102–108.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Lord, D., Brewer, M.A., Fitzpatrick, K., Geedipally, S.R., Peng, Y., 2011. Analysis of Roadway Departure Crashes on Two-Lane Rural Highways in Texas. Report No. FHWA/TX-11/0-6031-1. Texas A&M Transportation Institute, College Station, TX.

Lord, D., Geedipally, S.R., Guo, F., Jahangiri, A., Shirazi, M., Mao, H., Deng, X., 2019. Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data. Report No. 01-001, Safe-D UTC. U.S. Department of Transportation, Washington, D.C.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10, 325–337.

McCullagh, P., 1989. Generalized Linear Models, second ed. Chapman and Hall, Ltd, London, U.K.
Miaou, S.-P., 1996. Measuring the Goodness-of-fit of Accident Prediction Models. FHWA-RD-96-040, Final Report. Federal Highway Administration, McLean, VA.
Miaou, S.-P., Lord, D., 2003. Modeling traffic-flow relationships at signalized intersections: dispersion parameter, functional form and Bayes vs empirical Bayes. Transport. Res. Rec. 1840, 31–40.
Montella, A., Andreassen, D., Tarko, A.P., Turner, S., Mauriello, F., Imbriani, L.L., Romero, M.A., Singh, R., 2012. Critical review of the international crash databases and proposals for improvement of the Italian national database. Procedia Soc. & Behav. Sci. 53, 49–61.
Nabavi Niakia, M.S., Saunier, N., Miranda-Moreno, L.F., 2019. Is that move safe? Case study of cyclist movements at intersections with cycling discontinuities. Accid. Anal. Prev. 131, 239–247.
National Academies of Sciences, Engineering, and Medicine, 2014. Naturalistic Driving Study: Technical Coordination and Quality Control. The National Academies Press, Washington, DC. https://doi.org/10.17226/22362.
Neldelman, J., Wallenius, T., 1986. Bernoulli trials, Poisson trials, surprising variance, and Jensen's inequality. Am. Statistician 40 (4), 286–289.
NHTSA, 2018. Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. Traffic Safety Facts. National Highway Traffic Safety Administration, Washington, D.C.
Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transport. Res. Rec. 1840, 41–49.
Olkin, I., Gleser, L.J., Derman, C., 1980. Probability Models and Applications. MacMillan Publishing Co., Inc, New York, N.Y.
Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection-accident frequency. J. Transport. Eng. 122 (No. 2), 105–113.
Rao, J.N.K., 2003. Small Area Estimation. Wiley, Hoboken, New Jersey.
Rumar, K., 1985. The role of perceptual and cognitive filters in observed behavior. In: Evans, L., Schwing, R. (Eds.), Human Behavior in Traffic Safety. Plenum Press.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.
Shirazi, M., Lord, D., 2019. Characteristics based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. Transportmetrica Transport. Sci. 15 (2), 1791–1803.
Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on characteristics of data: application to investigate when the negative binomial lindley (NB-L) is preferred over the negative binomial (NB). Accid. Anal. Prev. 107, 186–194.
Smith, T., 2020. Disruptive Technology. Investopedia. https://www.investopedia.com/terms/d/disruptive-technology.asp. (Accessed 12 June 2020).
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B 64, 583–639.
Stipancica, J., Miranda-Moreno, L., Saunierc, N., 2018. Vehicle manoeuvers as surrogate safety measures: extracting data from the gps-enabled smartphones of regular drivers. Accid. Anal. Prev. 115, 160–169.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, U.K.
Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.
Xie, K., Ozbay, K., Yang, H., Yang, D., 2019. A new methodology for before–after safety assessment using survival analysis and longitudinal data. Risk Anal. 39 (6), 1342–1357.