CHAPTER

5

Exploratory analyses of safety data

5.1 Introduction

Exploratory data analyses focus on presenting a variety of techniques for performing initial investigations on data with the help of summary statistics and graphical representations. They are used to accomplish the following objectives:

1. Understanding the data, mapping their underlying structure and identifying data issues such as errors and missing information,
2. Selecting the most important variables and identifying possible relationships in terms of direction and magnitude between independent and outcome variables,
3. Detecting outliers whose values are significantly different from the other observations in the dataset,
4. Testing hypotheses and developing associated confidence intervals or margins of error,
5. Examining underlying assumptions to know if the data follows a specific distribution, and
6. Choosing a preliminary model that fits the data appropriately.

This chapter describes different methods and techniques for exploring safety data. The exploratory data analyses are conducted using two different techniques: (1) quantitative techniques that involve the calculation of summary statistics, (2) and graphical techniques that employ charts to summarize the data. Additionally, exploratory data analyses can be divided into univariate or multivariate (typically bivariate) methods.

Highway Safety Analytics and Modeling
https://doi.org/10.1016/B978-0-12-816818-9.00015-9

Univariate methods look at one variable (independent or outcome variable) at a time, while multivariate methods look at two or more variables (several independent variables alone or with an outcome variable) simultaneously to explore relationships. It is always recommended to initially perform a univariate analysis for each variable in a multivariable dataset before performing a multivariate analysis. The first part of the chapter focuses on the quantitative techniques, while the second part summarizes the graphical techniques.

5.2 Quantitative techniques

This section describes five different quantitative techniques.

5.2.1 Measures of central tendency

Safety datasets are usually large with many variables. It is always useful to represent the variables using summary statistics. Central tendency is the most common statistic used to describe the "average" "middle" or "most common" value of a variable. The mean, median, and mode are the measures that are used to describe the central tendency. It is always suggested to compute and analyze the mean, median, and mode for a given dataset simultaneously as they elucidate different aspects of the given data. Considering them alone can lead to misrepresentations of the data due to outliers or extreme values.

5.2.1.1 Mean

The arithmetic mean, or simply called the mean is calculated by dividing the sum of all observations in the dataset by the total number of observations. The mean is significantly affected by the outliers, that is, extremely large or small values. The mean is also called as mathematical expectation, or average. The sample mean is denoted by $\overline{x}$ (pronounced as "x-bar") and is calculated using the following equation.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

where n is the total number of observations in the sample and $x_1, x_2, \ldots x_n$ are individual observations. As the sample mean changes from one sample to another, it is considered as a random variable. If the whole population is used, then $\overline{x}$ is replaced by the Greek symbol, μ and is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5.2}$$

where N is the total number of observations in the population. The population mean is always fixed and is a nonrandom variable.

5.2.1.2 Median

The median is a value that divides the dataset or a probability distribution into two halves. Sorting of the data in a particular order is important when calculating the median of a variable. The observations can be sorted in an ascending or descending order to calculate the median. If the total number of observations in a dataset is odd, then the median is simply the number in the middle of the list of all observations. If the total number of observations in a dataset is even, the average of the two middle values is the median. When the data contain outliers, the median is not affected, and so it is considered more robust than the mean.

5.2.1.3 Mode

The observation that has the highest number of occurrences in the dataset is called the mode. When two or more observations occur frequently, then we have more than one mode in the dataset. Similar to the mean and median, the mode is a measure of central tendency that has the highest probable outcome in the data sample. Unlike the mean and median, however, the mode can be applied to nonnumerical or qualitative data (i.e., data measured on the nominal and ordinal scale).

5.2.2 Measures of variability

The central tendency measures do not always provide adequate information related to the data. The information on the variability is required to understand the amount of spread or dispersion in the dataset. Low dispersion is indicated by the observations clustering tightly around the center. Alternatively, high dispersion means that the observations fall further away from the center.

5.2.2.1 Range

Range is the simplest measure that is used to calculate the amount of variability. The range is defined as the difference between the largest and smallest observations in the dataset. Although the range is simple to calculate and easy to understand, it is highly susceptible to outliers because its value is based on only the two most extreme observations in the dataset. When dealing with traffic crashes, in many situations, the value of the largest observation is unusually high, which affects the entire range.

Additionally, the range is significantly affected by the sample size of the dataset. For small samples with no possible outliers, the range is particularly suitable because the other measures cannot be calculated reliably. However, the chances of outliers increases as the sample size becomes larger. Consequently, if we draw multiple random samples with different sample sizes from the same population, then the range tends to increase with the increase in the sample size.

5.2.2.2 Quartiles and interquartile range

Quartiles separate the dataset into four equal parts after sorting in the ascending order. Quartiles use percentage points (or percentiles) to divide the data into quarters. A percentile is defined as a value below which lies a certain percentage of observations in the sample. The lowest or first quartile (Q1) is the 25th percentile value, the median or middle quartile (Q2) is the 50th percentile value, and the upper or third quartile (Q3) is the 75th percentile value. For setting up speed limits on highways, the 85th percentile speed is the commonly used measure. It suggests that 85% of sample driver speeds observed are lower than the 85th percentile speed.

The interquartile range (IQR) is the middle half of the data and is used to understand the data spread. The IQR is calculated as the difference between Q3 and Q1. The IQR includes 50% of observations that fall between Q1 and Q3. For skewed distributions, the IQR and median are the robust measures of variability and central tendency, respectively. Similar to the median, the IQR is not influenced significantly by outliers because it does not consider the extreme values.

5.2.2.3 Variance, standard deviation and standard error

The variance and standard deviation are the two most frequently used measures to calculate the dispersion in the data. Unlike the range and IQR, the variance and standard deviation consider all the observations in the calculation by comparing each observation to the mean. The variance is calculated using two equations, depending on whether we are interested in the sample variance or the variance for the entire population. As collecting the whole population is not always possible, the sample variance is commonly used as an estimate of the population variance. The sample variance changes from one iteration to the next so it is a random variable. The sample variance is calculated using the following equation:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{5.3}$$

where $\overline{x}$ is the sample mean, and n is the total number of observations in the sample. Similar to the sample mean, as the sample variance changes from one sample to another, it is considered a random variable. If the

whole population is used, then s^2 is replaced by the Greek symbol, σ^2 and the population variance is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{5.4}$$

where μ is the population mean and N is the total number of observations in the population. Unlike sample variance, the population variance is always fixed and is a nonrandom variable.

There are two reasons for using $n - 1$ in the denominator, instead of n for calculating the sample variance. First, as the sample mean $\overline{x}$ is used in the calculation, it loses 1 degree of freedom and there are $n - 1$ independent observations remaining to calculate the variance. Second, when the small sample size is used, the variance tends to be underestimated and so it is compensated by using $n - 1$, instead of n. However, when a large sample is considered, the difference in the variance calculation either with $n - 1$ or n becomes negligible.

The standard deviation is calculated using the square root of the variance. The standard deviation is defined as a standard difference between each observation and the mean. The standard deviation is small when the data points are grouped closer together. Alternatively, it is larger when the data points are spread out. The sample and population standard deviations can be calculated using the following equations, respectively:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{5.5}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{5.6}$$

The standard deviation is a more widely used measure than the variance because the units of the standard deviation are the same as the original units of the data, which makes the interpretation easier.

The standard error is often confused with the standard deviation measure. The standard deviation is used to measure the variability to understand how scattered some observations are in the dataset. The standard error of an estimate is the standard deviation of its sampling distribution. For example, the mean calculated from different samples will vary from one to another; and the variation can be described by a distribution called as the "sampling distribution" of the mean. The standard error is an estimate that is calculated to know how much sample means will vary from the standard deviation of this sampling distribution. The standard error of the sample mean is calculated by the following equation.

$$SE = \frac{s}{\sqrt{n}} \tag{5.7}$$

where s is the sample standard deviation, and n is the total number of observations in the sample. With the increase in the sample size, the standard error decreases but the standard deviation tends to remain unchanged.

5.2.2.4 Coefficient of variation

The standard deviation provides a measure of variability without considering the magnitude of variable values. The coefficient of variation (CV), also called relative standard deviation which is a unitless quantity, is a measure of relative variability that provides the dispersion of observations in a dataset around the mean. The CV for a sample is calculated using the following equation:

$$CV = \frac{s}{\overline{x}} \tag{5.8}$$

where $\overline{x}$ is the sample mean. The CV is often expressed in percentages and is a useful measure to compare the degree of variability from one sample to another as it does not have any unit. For example, if we want to compare the variation in traffic crashes between two facility types, then the CV shows which facility type has more variation than the other.

Exercise 5.1

The following are the crashes that occurred on 30 segments selected from Roadway Segment Dataset. Provide the summary statistics for crashes.

1,3,0,0,6,2,0,1,4,0,1,3,16,0,1,0,2,1,1,3,2,8,5,2,3,2,0,1,0,4.

Statistic	Value
Mean	2.4
Median	1.5
Mode(s)	0.0
Range	16.0
25th percentile	0.0
75th percentile	3.0
IQR	3.0
Variance	10.5
Standard deviation	3.2
CV	1.3

5.2.2.5 Symmetrical and asymmetrical data

The data follow a symmetrical distribution when its observations occur at regular frequencies and all the measures of central tendency (i.e., mean, median, and mode) occur at the same point. Fig. 5.1a shows a symmetrical distribution that appears in the shape of a bell curve. If a line was drawn dissecting the middle of the curve, the left side of the distribution mirrors the right side. Examples of data that are symmetrically distributed around their mean include free-flow speeds of vehicles (Berry and Belmont, 1951) and logarithm of crash rates (Ma et al., 2015). For a symmetrical distribution with a large sample size, it is recommended to express the results in terms of mean and standard deviation.

Asymmetrical distributions, also known as skewed distributions, can be either right-skewed or left-skewed and do not have the same value for the mean, median, and mode. A right-skewed or a positively skewed distribution has a longer tail on the right and the mean is on the right of the mode (see Fig. 5.1b). Examples of data that are usually asymmetrically distributed around their mean and follow a positively skewed distribution include the crash rate data (Ma et al., 2015), crash frequency data (Miaou, 1994), and travel time data (Berry and Belmont, 1951). A left-skewed or a negative distribution has a longer tail on the left, and the mean is on the left of the mode (see Fig. 5.1C). This distribution is rarely found for any variable in the traffic safety datasets. For asymmetrical distributions, or when the sample size is small, it is recommended to express the results in terms of median and IQR.

5.2.2.6 Skewness

Skewness is used to measure the degree of asymmetry of a distribution. In other words, it quantifies the degree of distortion from the normal distribution. It differentiates extreme values from one tail to the other. A symmetrical distribution (such as normal distribution) has a skewness of 0. The sample skewness can be calculated using the following equation.

$$g_1 = \frac{\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}\right)}{s^3} \tag{5.9}$$

where s is the sample standard deviation.

The data are assumed to be characterized as follows:

- Symmetrical, if $-0.5 \leq g_1 < +0.5$
- Moderately negative skewed, if $-1.0 \leq g_1 < -0.5$
- Moderately positive skewed, if $+0.5 \leq g_1 < +1.0$
- Highly negative skewed, if $g_1 \leq -1.0$
- Highly positive skewed, if $g_1 \geq +1.0$

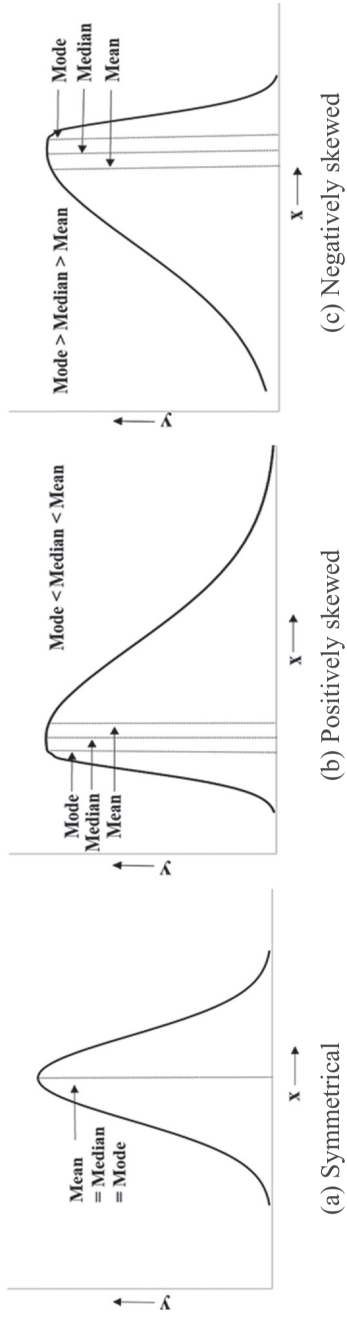

FIGURE 5.1 Symmetrical and skewed distributions.

5.2.2.7 Kurtosis

Kurtosis is the measure of the sharpness of the peak of a frequency distribution. The sample kurtosis can be calculated using the following equation.

$$g_2 = \frac{\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}\right)}{s^4} - 3 \tag{5.10}$$

If the kurtosis value is positive then it suggests heavy tails. Alternatively, a negative value means that there are light tails. The tail heaviness or lightness is in comparison with the normal distribution and it suggests whether the data distribution is flatter or less flat than the normal distribution. The kurtosis value is 3.0 for a standard normal distribution.

Kurtosis can be categorized into three measures, as shown in Fig. 5.2. If the kurtosis statistic of a distribution is similar to that of the normal distribution, or bell curve, then it is called a mesokurtic distribution. This type of distribution has similar extreme value characteristics as that of a normal distribution. If the kurtosis is greater than a mesokurtic distribution then it is called a leptokurtic distribution. This distribution has long tails (due to the presence of many outliers). The outliers stretch the horizontal axis and a lot of data appear in the narrow curve. The final type of distribution is a platykurtic distribution and it has kurtosis that is smaller than a mesokurtic distribution. This type of distribution has short tails (due to the presence of fewer outliers). When compared to the normal distribution, these distributions have fewer extreme values.

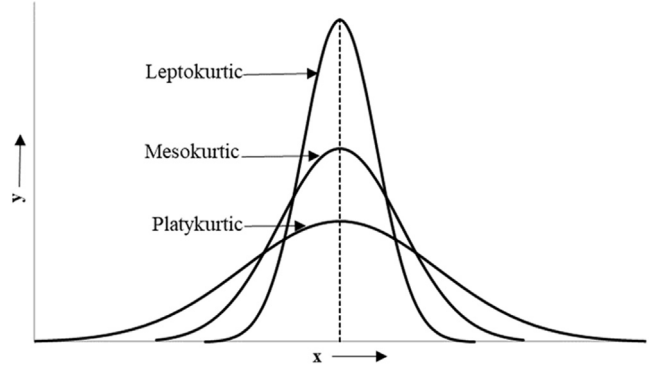

FIGURE 5.2 Kurtosis in the normal curve.

Exercise 5.2

Using the data presented in Exercise 5.1, calculate the skewness and kurtosis.

$$g_1 = \frac{\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}\right)}{s^3} = \frac{87.808}{3.2^3} = 2.6$$

$$g_2 = \frac{\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}\right)}{s^4} - 3 = \frac{1190.45}{3.2^4} - 3 = 7.89$$

As the skewness is 2.6, it is highly positively skewed, similar to the one shown in Fig. 5.1b. The mean, median, and mode statistics also confirm that the distribution is positively skewed.

The kurtosis value is 7.89, indicating a relatively "skinny" (lepto-kurtic) distribution. This distribution has longer tails, due to the presence of outliers. In this case, the site with 16 crashes is a potential outlier.

5.2.3 Measures of association

The measure of association is used to quantify the relationship between two variables. Correlation and regression analysis are among several methods that are used to quantify the measure of association. The correlation between two variables refers to a measure of the linear relationship, whereas association refers to any relationship between variables. The selection of the method to determine the strength of an association is dependent on the characteristics of data for a variable. Data can be observed on an interval/ratio (continuous) scale, an ordinal/rank (integer) scale, or a nominal/categorical (qualitative) scale.

5.2.3.1 Pearson's correlation coefficient

When the association between two variables that are measured on an interval/ratio (continuous) scale is sought, the appropriate measure of association is Pearson's correlation coefficient. Pearson's correlation coefficient r for a sample is defined by the following equation.

$$r = \frac{COV(X, Y)}{s_x s_y} \tag{5.11}$$

with,

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{5.12}$$

where $COV(X,Y)$ is the sample covariance between two random variables X and Y that are normally distributed with means $\overline{x}$ and $\overline{y}$ and standard deviations s_x and s_y respectively. To calculate the population correlation coefficient, the sample means $\overline{x}$ and $\overline{y}$ are replaced by population means μ_x and μ_y, and the sample standard deviations s_x and s_y are replaced by population standard deviations σ_x and σ_y respectively.

5.2.3.2 *Spearman rank-order correlation coefficient*

The Spearman rank-order correlation coefficient, a nonparametric method, is used to measure the strength of association between two variables when one or both are measured on an ordinal/ranked (integer) scale, or when both variables are not normally distributed. If one of the variables is on an interval scale, then it needs to be transformed to a rank scale to analyze with the Spearman rank-order correlation coefficient, although this may result in a loss of information. Once two variables are ranked and sorted in an ascending order, the spearman correlation coefficient r_s is calculated using the following equation.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{5.13}$$

where d_i (d_1, d_2, ….d_n) are the differences in ranks of two variables x_i (x_1, x_2, ….x_n) and y_i (y_1, y_2, ….y_n).

The correlation coefficient takes on the values from −1.0 to +1.0. A value of −1.0 indicates a perfect negative linear relationship between the two variables, which means as one variable increases, the other decreases. Similarly, a value of +1.0 indicates a perfect positive linear relationship between the two variables, which means as one variable increases, the other increases too. If the value is 0.0, then it indicates no linear relationship. Any coefficient value between −1.0 and 0.0 or 0.0 and +1.0 indicate a negative or positive linear relationship but not an exact straight line. Hinkle et al. (2003) provided a rule of thumb for interpreting the correlation coefficient, as shown in Table 5.1.

5.2.3.3 *Chi-square test for independence*

The chi-square test for independence is commonly used for testing relationships between two sets of data that are measured on the categorical scale. The chi-square test is used to measure the significance of the relationship but not the strength of the association. Before conducting the chi-square test, data must be arranged as a contingency table (a matrix

TABLE 5.1 Interpreting of correlation coefficient (Hinkle et al., 2003).

Correlation coefficient[a]	Interpretation
+0.9 to +1.0 (−0.9 to −1.0)	Very high correlation
+0.7 to +0.9 (−0.7 to −0.9)	High correlation
+0.5 to +0.7 (−0.5 to −0.7)	Moderate correlation
+0.3 to +0.5 (−0.3 to −0.5)	Low correlation
−0.3 to +0.3	Negligible correlation

[a] *"+" means positive correlation and "−" means negative correlation.*

that shows the frequency distribution of variables). The rows represent the bins or categories. The columns represent the frequencies for two variables of interest (one variable is represented as "observed" and other as "expected"). A two-way table is similar to a frequency distribution except that the two variable frequencies (observed and expected values) are shown simultaneously. The chi-square χ^2 statistic is calculated using the following equation.

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{5.14}$$

where O_i and E_i are the observed and expected frequencies in the i_{th} category (i=1,2 … k). The term df is the degrees of freedom, which is equal to $k - 1$. The calculated chi-square value will be compared to the critical value from a chi-square table for a chosen significance level α (e.g., $\alpha = 0.05$). If the calculated chi-square value is greater than the critical value, then it can be concluded that there is a significant difference between observed and expected frequencies. The chi-square statistic is extremely sensitive to the sample size within each category. If a particular category has expected frequency fewer than 5, then it should be combined with the adjacent category.

The chi-square test for independence should not be confused with the chi-square goodness-of-fit test, although the formula for the test is the same in both cases. The test for independence is used for testing the association between two sets of data, whereas the goodness-of-fit test is used to test if the data sample follows a certain distribution. It should be noted that the chi-square test can only be used for discrete distributions (e.g., Poisson, and binomial distributions). For continuous distributions (e.g., normal and uniform distributions), other tests such as Kolmogorov−Smirnov goodness of fit test should be used.

5.2.3.4 Relative risk and odds ratio

Relative risk and odds ratio are the two other measures used to test the association between categorical variables. The two tests are used to measure the strength of the association but do not provide the significance of the relationship. For calculating the relative risk or odds ratio, data must be arranged as a two-way contingency table (a two-by-two matrix that shows the frequency distribution of variables in two groups for two outcomes). The following example of two-way contingency table shows the frequency of two mutually exclusive outcomes (e.g., fatal vs. nonfatal crashes) for each of the two groups (e.g., cars installed with airbags vs. cars without airbags) to understand the role of airbags in saving lives when involved in a collision.

	Outcome	
Group	**Outcome 1**	**Outcome 2**
Treatment	A	B
Control	C	D

The relative risk (also known as risk ratio) is used to evaluate the risk (or probability) of an outcome in one group when compared to the risk of the same outcome in the other group. A relative risk of 1.0 indicates no difference in risk between the groups, whereas a relative risk other than 1.0 indicates that there is a difference between the groups. The relative risk RR is calculated using the following equation.

$$RR = \frac{A/(A+B)}{C/(C+D)} \tag{5.15}$$

The odds ratio is used to evaluate the odds of an outcome in one group when compared to the odds of the same outcome in the other group. Similar to the relative risk, odds ratio of 1.0 indicates no difference in risk between the groups, whereas an odds ratio other than 1.0 indicates that there is a difference between the groups. The odds ratio OR is calculated using the following equation.

$$OR = \frac{A/B}{C/D} = \frac{AD}{BC} \tag{5.16}$$

For rare events (such as traffic crashes), where the chance of occurrence is too low (<5%), both RR and OR provide similar results and can be used interchangeably. For all values of RR less than 1, OR is always lower than RR. However, when RR is greater than 1, OR is always greater than RR.

Exercise 5.3

Using the Naturalistic Driving Dataset, Owens et al. (2018) evaluated the crash risk of cell phone use while driving. The authors have identified 253 crash events, of which 83 involved cell phone usage while driving. Similarly, they have identified 849 no-crash events, of which 236 involved cell phone usage. Calculate the *RR* and *OR*.

First, summarize the data into a two-way contingency table.

	Outcome	
Group	**Crash events**	**No-crash events**
Cell phone use	83	236
No cell phone use	170	613

Second, calculate the *RR* and *OR*.

$$RR = \frac{83/(83+236)}{170/(170+613)} = \frac{0.26}{0.22} = 1.18$$

$$OR = \frac{83/236}{170/613} = \frac{0.35}{0.28} = 1.25$$

As the chance of crash outcome is much greater than 5% in this case, the results from *RR* and *OR* are different. As the *RR* and *OR* values are greater than 1, the results suggest that the cell phone use is associated with the increased risk of traffic crashes.

5.2.4 Confidence intervals

The estimate of mean is highly dependent on the sample considered and varies from one sample to another. Given the variability in sample means, it is valuable to consider the interval estimates rather than a single point estimate for mean. If the population mean is unknown, then a confidence interval provides an estimated range of values in which the unknown mean is likely included. The interval range is estimated from the sample data. The interval also provides an indication about the uncertainty exists in the estimate of true mean. If the interval is narrower, then it indicates that the estimate is more precise. A confidence level of 95% is generally considered; however, other levels such as 90% and 99% are also widely used. Higher confidence levels provide wider confidence intervals, which means it is highly likely to contain the population mean. It should be noted that an interval computed from the sample may or may

not contain the true mean. If many samples are used to compute multiple confidence intervals, then the proportion of samples where the true mean is expected to be contained is called the confidence coefficient.

5.2.4.1 Confidence intervals for unknown mean and known standard deviation

When a population with unknown mean μ and known standard deviation σ is considered, the confidence interval for the population mean, for a confidence level C, is calculated as

$$\overline{x} \pm Z\frac{\sigma}{\sqrt{n}} \tag{5.17}$$

where Z is the upper $(1-C)/2$ critical value for the standard normal distribution. The term $Z\frac{\sigma}{\sqrt{n}}$ is called margin of error.

For a relatively large sample from any population distribution, the sample mean $\overline{x}$ follows a normal distribution with mean μ (i.e., unknown population mean) and the standard deviation of mean (also called standard error) $\sigma/\sqrt{n}$, with n being the sample size.

5.2.4.2 Confidence intervals for unknown mean and unknown standard deviation

When the population standard deviation σ is unknown, it is replaced by standard deviation s that is estimated from the sample. In such a scenario, the sample mean $\overline{x}$ does not follow a normal distribution but a t distribution with mean μ and the standard deviation $s/\sqrt{n}$. When a population with unknown mean μ and unknown standard deviation is considered, the confidence interval for the population mean, for a confidence level C, is calculated as

$$\overline{x} \pm t\frac{s}{\sqrt{n}} \tag{5.18}$$

where t is the upper $(1-C)/2$ critical value for the t distribution with $n-1$ degrees of freedom.

5.2.4.3 Confidence intervals for proportions

It is sometimes of interest to estimate the confidence intervals for a population proportion (say, a proportion of Driving While Intoxicated or DWI between two adjacent cities). As the population proportion p is unknown, a sample proportion $\widehat{p}$ is used instead. For a large sample size, the sample proportion $\widehat{p}$ follows a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$. When a random sample is drawn from

the population then the confidence interval for the population proportion p, for a confidence level C, is calculated as

$$\widehat{p} \pm Z\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}} \tag{5.19}$$

where Z is the upper (1−C)/2 critical value for the standard normal distribution.

Exercise 5.4

Majority of the crashes on horizontal curves are speed-related. Advisory speeds are set to inform motorists about the safe speeds when traversing along horizontal curves. The advisory speed setting is based on the average truck speed, so an agency is interested in knowing the truck proportion in the traffic. A survey was conducted at two horizontal curves for a short period of time in Texas. At the first horizontal curve, 296 passenger cars and 43 trucks, and at the second curve, 324 passenger cars and 72 trucks were observed. What is the confidence interval for the proportion of trucks at 95% level?

The sample truck proportion in the traffic is $(43+72)/(296+43+324+72) = 0.156$. The z-value for the 95% level $\left(\text{significance level} = \frac{1-C}{2} = \frac{1-0.95}{2} = 0.025\right)$ is 1.96. The confidence interval for the truck proportion is obtained as

$$\left[\widehat{p}+Z\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}}, \widehat{p}-Z\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}}\right] = \left[0.156+1.96\sqrt{\frac{0.156(1-0.156)}{735}}, 0.156-1.96\sqrt{\frac{0.156(1-0.156)}{735}}\right] = [0.13, 0.182]$$

5.2.4.4 *Confidence intervals for the population variance and standard deviation*

The sample variance shown in Eq. (5.3) provides an unbiased estimate of the population variance, although the sample standard deviation shown in Eq. (5.5) may provide a biased estimate for the population

standard deviation. As both variance and standard deviation are nonnegative numbers, they do not follow a normal distribution. If x_1, x_2, ….x_n are normally distributed, then the term $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution (a nonsymmetric distribution) with $n-1$ degrees of freedom. The confidence interval for the population variance and standard deviation, for a confidence level C, is calculated using the following equations, respectively.

$$\frac{(n-1)s^2}{\chi^2_{\frac{(1-C)}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{(1-C)}{2},n-1}} \tag{5.20}$$

$$\sqrt{\frac{(n-1)}{\chi^2_{\frac{(1-C)}{2},n-1}}}s \leq \sigma \leq \sqrt{\frac{(n-1)}{\chi^2_{1-\frac{(1-C)}{2},n-1}}}s \tag{5.21}$$

Exercise 5.5

The operating speeds of passenger cars were measured for a few hours on a random day at the midpoint of a horizontal curve in Texas. In total, 952 car speeds were observed and it was found that the mean operating speed is 64.1 mph and the sample standard deviation is 5.3 mph. What is the confidence interval at 95% level for the mean operating speed and the standard deviation of speeds?

As the population mean and standard deviations are unknown, the confidence interval for the population mean shown in Eq. (5.18) should be used. The t-value from the t-distribution for the 95% level is 1.96. The confidence interval for the mean operating speed is obtained as

$$\bar{x} \pm t\frac{s}{\sqrt{n}} = 64.1 \pm 1.96\frac{5.3}{\sqrt{952}} = [63.76, 64.44]$$

For the standard deviation of speeds, the confidence interval shown in Eq. (5.21) should be used. From the χ^2 table, the approximate value of $\chi^2_{0.025,951}$ is 1037.3 and $\chi^2_{0.975,951}$ is 866.5. The confidence interval for the standard deviation of speeds is obtained as

$$\left[\sqrt{\frac{(n-1)}{\chi^2_{\frac{(1-C)}{2},n-1}}}s, \sqrt{\frac{(n-1)}{\chi^2_{1-\frac{(1-C)}{2},n-1}}}s\right] = \left[\sqrt{\frac{951}{1037.3}} \times 5.3, \sqrt{\frac{951}{866.5}} \times 5.3\right]$$
$$= [5.08, 5.55]$$

5.2.5 Hypothesis testing

The test to determine whether a hypothesis is true or not with the use of statistics based on a sample data is called hypothesis testing. There are four steps in the hypothesis testing.

Step 1—State the hypotheses. The first step involves stating the null and alternative hypotheses. Both hypotheses are stated in a mutually exclusive manner. That is, if null is true, then the alternative is false or vice-versa.

H_0 (null hypothesis): no variation exists between variables or that a single variable is no different than its mean.

H_1(alternate hypothesis): variation exists between variables or that a single variable is different than its mean.

Step 2—Select confidence level. This step involves selecting an appropriate confidence level (C). Many studies use the significance level α, which is 1-C. A significance level of 0.05 (i.e., confidence level of 95%) is widely used but other significance levels equal to 0.01, or 0.10 are not uncommon either. Note that smaller significance levels require more data to properly detect the difference.

Step 3—Choose the test method and compute the probability. The test method is highly dependent on the data sampling distribution. The test method typically involves a test statistic that might be a mean score, proportion, difference between means, difference between proportions, etc. Compute the probability (*P*-value) that provides an evidence whether to accept or reject the null hypothesis.

Step 4 - Interpret results. The *P*-value is compared against the significance level (1-C) selected in Step 2. If the *P*-value is less than 1-C, then there is an evidence to reject the null hypothesis which states that the observed effect is statistically significant, and the alternative hypothesis is considered valid. Alternatively, if *P*-value is greater than the significance level, the null hypothesis cannot be rejected, which states that the observed effect is not statistically significant. As the *P*-value becomes smaller, the evidence against the null hypothesis becomes stronger.

5.2.5.1 Decision errors

When a decision is made based on the hypothesis testing, there is a chance that there is an error in the decision made. Two types of errors can result from a hypothesis test.

- Type I error. A Type I error occurs when a null hypothesis is rejected even though it is true. The probability of committing a Type I error is nothing but the significance level α selected in Step 2 of a hypothesis test.

- Type II error. A Type II error occurs when a null hypothesis is not rejected even though it is false. The probability of committing a Type II error is denoted by β. The probability of not committing a Type II error is called the Power of the test, and is denoted by 1- β.

5.2.5.2 *Two-tailed hypothesis test*

The two-tailed test is a method in which the rejection region is on two sides of the sampling distribution. It tests whether a sample mean is equal to a certain preselected value (also called null value) or the mean for the corresponding population. It does not matter if the sample mean is greater than or less than the population mean. The two-tailed test is also known as nondirectional test. The critical value which is a threshold that defines the boundaries is used to define the critical areas. If the calculated test statistic falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis. Fig. 5.3 shows the critical values for a two-tailed test with a 0.05 significance level. As the testing area is under two tails of a normal distribution, it is called the two-tailed test. In Fig. 5.3, the regions beyond the critical values are called the rejection regions. When the test statistic value is in the rejection region, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

5.2.5.3 *One-tailed hypothesis test*

A one-tailed test is a statistical test in which the rejection region is one-sided of the sampling distribution. It tests whether a sample mean is either greater than or less than a preselected value (also called null value) or the mean for the corresponding population. It is also known as a directional test and can be classified as upper-tailed or lower-tailed test. The upper-tailed test corresponds to testing if the sample mean is greater than the null value, whereas the lower-tailed test corresponds to the sample mean is less than the null value. If the test statistic falls into the

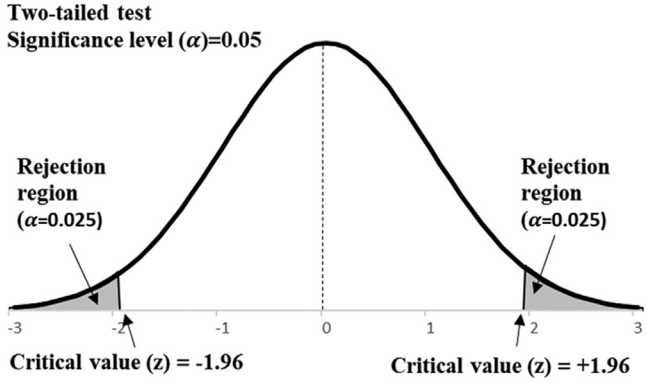

FIGURE 5.3 Critical values for a two-tailed (nondirectional) test.

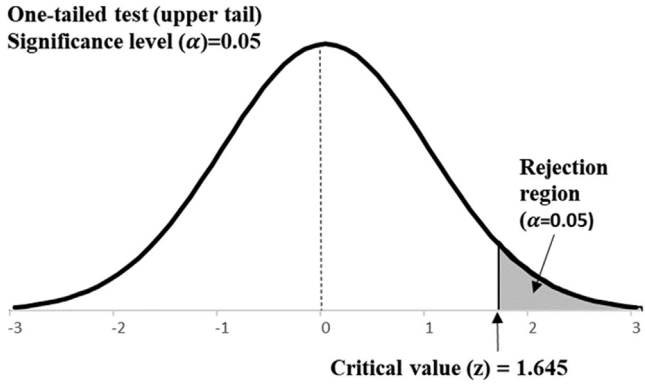

FIGURE 5.4 Critical values for a one-tailed (directional) test.

one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis. Fig. 5.4 shows the critical values for a one-tailed (directional) test with a 0.05 significance level. As the testing area is under one tail of a normal distribution, it is called the one-tailed test. A one-tailed test is used for testing in one direction of interest and disregarding the possibility of a relationship in the other direction. Similar to the two-tailed test, when the test statistic value is in the rejection region, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

5.2.5.4 *Hypothesis testing for one sample*

A standardized normal test statistic is used to conduct the hypothesis testing. The test statistic is called the z statistic and it is used to convert any sampling distribution into a standard normal distribution. The z statistic is used to determine the number of standard deviations in a standard normal distribution that a sample mean deviates from the population mean. When the population mean and standard deviation are known, the z statistic is calculated as

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{5.22}$$

The calculated value is the value of a test statistic and it is compared to the critical value(s) of a hypothesis test to make a decision (see Table 5.2). When the obtained value exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

When the population mean is known and the standard deviation is unknown, the test statistic is calculated as:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{5.23}$$

TABLE 5.2 Critical values for different levels of significance.

Significance level α	One-tailed test	Two-tailed test
0.10	+1.282 or −1.282	±1.645
0.05	+1.645 or −1.645	±1.96
0.01	+2.33 or −2.33	±2.58
0.001	+3.09 or −3.09	±3.30

The calculated test statistic is compared to the critical value of a hypothesis test to accept or reject the null hypothesis. The critical value is obtained from a t-distribution with $n-1$ degrees of freedom. When the calculated test statistic exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

For testing the population proportion, the test statistic is calculated as follows:

$$z = \frac{\left(\widehat{p} - p\right)}{\sqrt{p(1-p)/n}} \tag{5.24}$$

When the obtained value exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

If the population variance is of interest, then the test statistic is calculated as follows:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{5.25}$$

Here s^2 is the variance estimated from a sample. The critical value is obtained from a χ^2-distribution with $n-1$ degrees of freedom. When the obtained value exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

5.2.5.5 *Hypothesis testing for two samples*

Comparing two samples is of great interest to understand the difference between the two groups. The groups can be either dependent or independent with each other. In dependent groups, the observations from one group are paired with observations in the other group, so it is called matched pairs. When independent groups are considered, observations selected from one group are completely independent from the observations selected in the second group.

Dependent Samples. The dependent samples, also known as matched or paired samples, are those where observations in one group are paired with the observations in the other group.

Paired t-test. The paired sample t-test is the most common statistical procedure used for dependent samples. This test is useful for evaluating the differences in two time periods for the same observation or for comparing the two different treatments applied at the same site in different times. The first step to apply the test is to pair the observations in two different samples. Let us call the two samples as x_i (x_1, x_2, ….x_n) and y_i (y_1, y_2, ….y_n). The second step is to calculate the difference between the two paired observations d_i (d_1, d_2, ….d_n) where $d_i = x_i - y_i$. In the next step, we calculate the mean difference $\overline{d}$ and the standard deviation s. Using Eq. (5.23), we calculate the test statistic. Either a one-tailed or two-tailed hypothesis test can be used depending on the objective of the study. The calculated test statistic is compared to the critical value of a hypothesis test to accept or reject the null hypothesis. As described earlier, the critical value is obtained from a t-distribution with $n - 1$ degrees of freedom. When the calculated test statistic exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

The paired t-test is based on the assumption that the observations within each group are independent of each other. The other assumption is that the differences between the paired observations are approximately normally distributed and should not contain any outliers. If the normal distribution assumption is violated, then other tests such as the Wilcoxon signed rank test can be used. This is a nonparametric test that has been used in many transportation studies (Pratt et al., 2019; Hallmark et al., 2015; Jiang et al., 2011), and these studies showed Wilcoxon signed rank test is superior over traditional test approaches.

Exercise 5.6

A study was conducted to check the influence of consuming two beers on the driver impairment (UF Biostatistics). The study chose a sample of 20 drivers and it used an obstacle course to measure the reaction times before and after drinking two beers. Conduct an analysis to check whether consuming two beers will make the driver impaired or not.

First, summarize the data, and calculate the differences and its mean and standard deviation.

Before reaction times (x_i)	**After reaction times** (y_i)	**Difference** (d_i)
6.25	6.85	−0.60
2.96	4.78	−1.82
4.95	5.57	−0.62
3.94	4.01	−0.07

Exercise 5.6 (*cont'd*)

—cont'd

Before reaction times (x_i)	After reaction times (y_i)	Difference (d_i)
4.85	5.91	−1.06
4.81	5.34	−0.53
6.60	6.09	0.51
5.33	5.84	−0.51
5.19	4.19	1.00
4.88	5.75	−0.87
5.75	6.25	−0.5
5.26	7.23	−1.97
3.16	4.55	−1.39
6.65	6.42	0.23
5.49	5.25	0.24
4.05	5.59	−1.54
4.42	3.96	0.46
4.99	5.93	−0.94
5.01	6.03	−1.02
4.69	3.72	0.97
Mean difference ($\bar{d}$)		−0.5015
Standard deviation (s)		0.8686

Second, state the hypotheses.

H_0 (null hypothesis): consuming two beers will not change driver reaction times.

H_1(alternate hypothesis): consuming two beers will change driver reaction times.

Third, a significance level of 0.05 (i.e., the confidence level of 95%) is selected.

Fourth, as the test is to check whether there is a difference or not (i.e., it is irrelevant whether there is an increase or decrease in reaction times), we conduct a two-sided hypothesis testing.

$$t = \frac{-0.5015 - 0}{0.8686/\sqrt{19}} = -2.58$$

continued

Exercise 5.6 (*cont'd*)

The critical value for a two-tailed test from a *t*-distribution with 19 degrees of freedom for 0.05 significant level is 2.086, which is less than the absolute computed value. Thus, the null hypothesis can be rejected.

Alternatively, from a *t*-distribution, the critical values 2.539 and 2.861 correspond to a *P*-value of 0.02 and 0.01, respectively. With interpolation, the approximate *P*-value for a critical value of 2.58 is 0.018. The *P*-value is less than the selected significant level of 0.05. Thus, we reject the null hypothesis. The conclusion is that we have enough evidence to show that drinking two beers has an impact in driver's reaction times.

Independent Samples. The parameters tested using independent samples are either population means or population proportions.

Two Population Means with Unknown Standard Deviations. Let us denote the first random sample as X_1 and the second random sample as X_2. The population mean of the first sample is μ_1 and the second sample is μ_2. For testing the difference in population means, the random variable considered is the difference in sample means, $\overline{X}_1 - \overline{X}_2$. The difference between two samples depends on both means and standard deviations. As the population standard deviations σ_1 and σ_2 are unknown and are assumed to be unequal, they are estimated using the two sample standard deviations from the independent samples. The standard deviation, or standard error of the difference in sample means, $\overline{X}_1 - \overline{X}_2$, is estimated using the following equation.

$$s = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{5.26}$$

where the sample standard deviations s_1 and s_2 are estimates of unknown population standard deviations σ_1 and σ_2, respectively. The sample size of the first and second samples are n_1 and n_2, respectively.

When the sample size is large, the populations approximately follow normal distribution. The test statistic for testing the difference in population means is calculated as follows:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s} \tag{5.27}$$

where $\overline{x}_1$ and $\overline{x}_2$ are the sample means, and μ_1 and μ_2 are the population means.

When sample sizes are small ($5 \leq n_1 \leq 25$ and $5 \leq n_2 \leq 25$), the test statistic is approximated by the Student's t-distribution with degrees of freedom (df) as follows.

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_1}\right)^2} \tag{5.28}$$

When the population standard deviations σ_1 and σ_2 are equal, a pooled standard deviation is estimated from the sample standard deviations. The pooled standard deviation (s_p) is given as

$$s_p = \frac{(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2}{n_1 + n_2 - 2} \tag{5.29}$$

The test statistic for testing the difference in population means with equal population standard deviations is calculated as follows:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{5.30}$$

The test statistic is approximated by the Student's t-distribution with degrees of freedom (df) equal to $n_1 + n_2 - 2$.

Two Population Proportions. Similar to comparing two means, it is a common practice to compare two proportions. The difference in estimated proportions may be due to a difference in the populations or it could be just by chance. A hypothesis test is useful in determining if the difference in estimated proportions and the difference in population proportions are the same. When conducting a hypothesis test that compares two independent population proportions, it is important that the two independent samples are random and have no dependency. The rule of thumb is that the number of first outcomes (i.e., successes) is at least five, and the number of second outcomes (i.e., failures) is at least five, for each of the samples. If the difference of two proportions follows an approximately normal distribution, then the pooled proportion (p_p) is calculated as follows:

$$p_p = \frac{x_1 + x_2}{n_1 + n_2} \tag{5.31}$$

where the sample proportion for sample X_1 is $\widehat{p}_1 = x_1/n_1$ and for sample X_2 is $\widehat{p}_2 = x_2/n_2$.

If the proportions $\widehat{p}_1$ and $\widehat{p}_2$, and their difference $\widehat{p}_1 - \widehat{p}_2$ follow a normal distribution, then the test statistic for testing the difference in population proportions is calculated as follows:

$$z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{p_p(1 - p_p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{5.32}$$

5.2.5.6 Hypothesis testing for multiple samples

When we have three or more independent groups, the analysis of variance (ANOVA) is used to determine whether the population means are statistically different from each other. The ANOVA test is based on the assumption that the observations within each group are independent of each other. The other assumption is that the data are approximately normally distributed and should not contain any outliers. Additionally, it relies on the homogeneity of variance that means that the variance among the groups is approximately equal. ANOVA is dependent on two variances—variation within group observations and variation among groups.

The hypotheses of ANOVA can be stated as follows:

H_0 (null hypothesis): all population means are equal.

H_1(alternate hypothesis): at least one of the population means is different from others.

The test statistic for ANOVA is calculated as follows.

$$F = \frac{\sum_{j=1}^{k} n_j \left(\overline{x}_j - \overline{x}\right)^2 / (k-1)}{\sum_{j=1}^{k} (n_j - 1) s_j^2 / (N-k)} \tag{5.33}$$

where $\overline{x}_j$, s_j^2, n_j are the mean, variance, and sample size of group $j(j = 1, 2 \ldots k)$ respectively, k is the total number of groups, and N is the total number of observations in all groups together. The critical value is obtained from the F-distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = N - k$. When the obtained value exceeds a critical value, the null hypothesis is rejected; otherwise, we retain the null hypothesis.

Exercise 5.7

The operating speeds for passenger cars are measured on the straight section just before the start of horizontal curves on different highway types. Using ANOVA, conduct a test to verify if there is a difference in mean operating speeds between highway types.

Site no.	4-lane undivided	4-lane divided	4-lane freeways
1	73	74	83
2	77	72	88
3	71	73	82
4	72	77	82
5	70	81	82
6	70	72	82
7	73	76	80

Exercise 5.7 (*cont'd*)

We perform an ANOVA to test the null hypothesis that states all the mean operating speeds are equal. The test statistic for ANOVA is calculated as follows.

$$F = \frac{\sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2 / (k-1)}{\sum_{j=1}^{k} (n_j - 1) s_j^2 / (N-k)} = \frac{204.262}{7.442} = 27.449$$

The critical value for a F-distribution with degrees of freedom $df_1 = 2$ and $df_2 = 18$ for 0.05 significant level is 3.5546. As the calculated F-value is larger than the critical value, the null hypothesis is rejected. The conclusion is that there is enough evidence at least one highway has significantly different operating speeds than others.

5.3 Graphical techniques

This section describes the various graphical techniques that can be used to explore safety data. Eleven different techniques are covered.

5.3.1 Box-and-whisker plot

A box-and-whisker plot (also called a box plot) shows the data graphically through their quartiles (lower, middle, and upper quartiles described earlier) in a box-shaped plot. Whiskers are the lines extended both sides of the box indicating variability outside the upper and lower quartiles, hence the term box-and-whisker plot. Box plots received their name from the box in the middle. These plots are nonparametric and show variations in the data observations without making any assumptions of the underlying statistical distribution. The spacing between the different parts of the box indicates the degree of dispersion and skewness in the data. The plots are useful in estimating measures such as the interquartile range, median, and range. Box plots can be drawn either horizontally or vertically. Boxplot displays the five-number summary of a set of data, as shown in Fig. 5.5. The five-number summary is the minimum, first or lower quartile (Q1), median or middle quartile (Q2), third or upper quartile (Q3), and maximum. Q1 is defined as the middle number between the lowest observation and the median of the dataset. Q3 is the middle value between the median and the highest observation of the

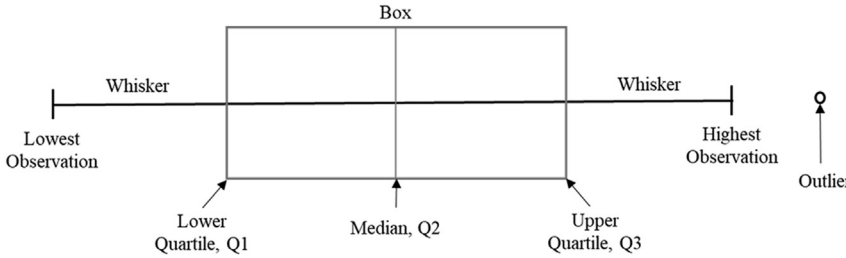

FIGURE 5.5 Box plot showing different measures.

dataset. Outlier data will not be included between the whiskers but are plotted as individual points outside the highest or lowest observations with a dot, or small circle. Outliers are not always plotted in a box plot.

Fig. 5.6 shows an example of a box plot that illustrates the traffic death rate in Africa from 1980 to 2015 (Adeloye et al., 2016). The middle quartile is closer to the lower quartile in almost all cases, which informs that the distribution of road traffic deaths is positively skewed. Due to the presence of outliers and skewed distributions, the IQR and median are the preferred measures of variability and central tendency, respectively, in this case.

5.3.2 Histogram

A histogram is a plot that shows the distribution of a continuous variable. The first step in constructing the histogram is to divide the entire range of variable values into a series of intervals called "bins". The second

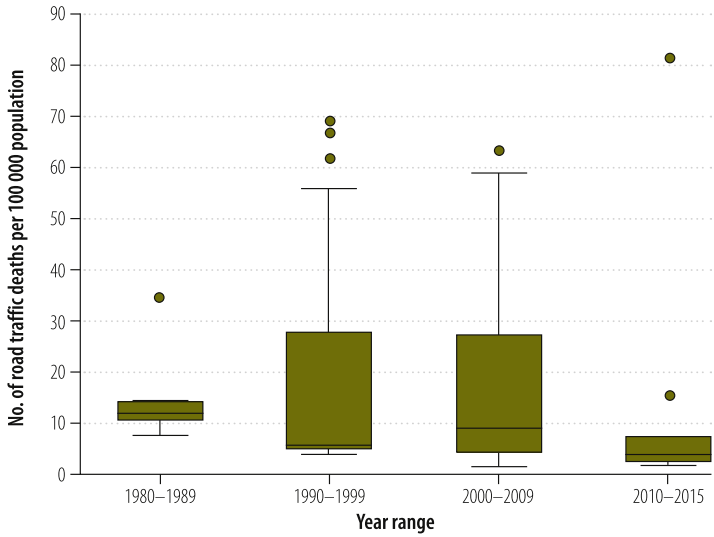

FIGURE 5.6 Box plot showing the traffic death rate in Africa. *From Adeloye D, Thompson JY, Akanbi MA, Azuh D, Samuel V, Omoregbe N, et al. The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and metaanalysis. Bull World Health Organ. 2016;94(7):510–21A.*

step is to count the frequency that will fall into each bin and show the frequency as a rectangle erected over the bin with height proportional to the frequency. The bins (intervals) are usually specified as consecutive and nonoverlapping intervals of a variable. The bins must be adjacent without any gap between them. Due to this, the rectangles of a histogram touch each other to indicate that the variable is continuous. The bins are usually of equal size but it is not a requirement. A histogram may also be normalized to display relative frequencies. In such a case, the histogram shows the proportion of observations that fall into several categories, with the sum of the heights equal to 1. If the unequal width bins are used, then the erected rectangle is defined to have its area proportional to the frequency of observations in the bin. The vertical axis then represents frequency density and not the frequency.

A histogram of blood alcohol concentration (BAC) values for passenger vehicle drivers involved in fatal crashes from the years 2000−04 is shown in Fig. 5.7 (NHTSA, 2007) The Fatality Analysis Reporting System (FARS) contains BAC levels for drivers involved in fatal crashes if they are tested positive for BAC. The ranges of BAC values are specified on the horizontal axis and the count of fatal crashes is specified on the vertical axis. Thus, the height of the rectangle in Fig. 5.7 is directly proportional to the number of observations corresponding to that range. The histogram shows that the highest count of fatal crashes occurs in the range with midpoint BAC = 0.10 g/dL. The histogram shows the mean value (0.194 g/dL) and the standard deviation (0.119 g/dL).

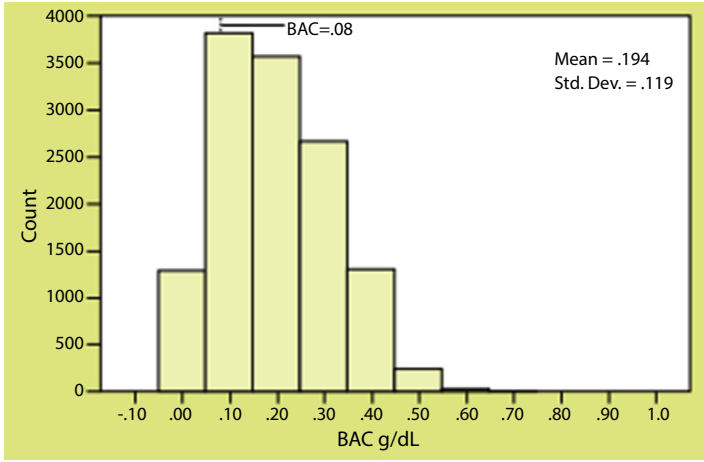

FIGURE 5.7 Histogram of passenger car driver BAC values. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.*

Histograms provide an approximate density of the underlying distribution of the variable of interest and can be used for estimating the probability density function. The probability density uses the total area of a histogram and is always normalized to 1. A kernel density is an extension of the histogram that uses a kernel to smooth frequencies over the bins. By doing this, we will get a smoother probability density function, which more accurately reflects the distribution of the variable of interest. The kernel density is a plot of a mathematical function where each value on the horizontal axis corresponds to a kernel density estimate on the vertical axis. Instead of a histogram, the density estimate could be plotted and is usually shown as a curve rather than a set of rectangular boxes. Although the kernel density function provides useful information about the data variable, histograms are generally preferred because they are simple and easy to construct. Fig. 5.8 shows a kernel density superimposed on the histogram of BAC values for passenger vehicle drivers involved in fatal crashes (NHTSA, 2007). The kernel density conveys similar information as the histograms about the density of the data.

The main advantage of kernel density estimation is that two or more distributions can be plotted on the same axis and compared against each other. It is not possible to produce simultaneous histograms on the same plot. The kernel density function for passenger car and motorcycle operator BAC values is shown in Fig. 5.9 on the same axis (NHTSA, 2007). This figure is useful in making the direct comparison of BAC distributional behavior of passenger cars and motorcycles. The figure also allows each distribution to be displayed in comparison to the legal limit of BAC = 0.08.

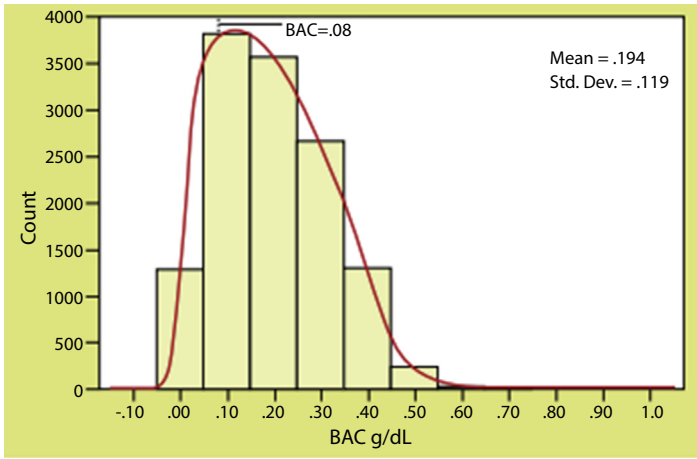

FIGURE 5.8 Histogram of passenger car driver BAC values with kernel density superimposed. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.*

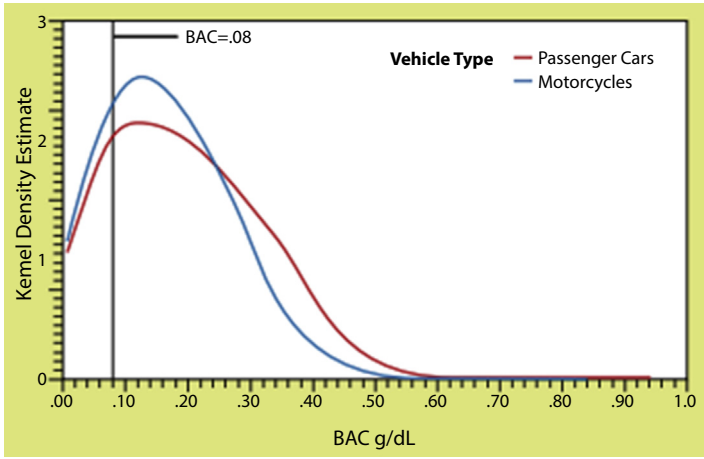

FIGURE 5.9 Kernel density plots: passenger car driver and motorcycle operator BAC values. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.

5.3.3 Bar graphs

The bar graph relates two variables as opposed to the histograms that relate only one. A histogram is a plot for continuous data, while a bar chart is used for categorical variables. The other difference is that the bar charts have gaps between the rectangles, whereas histograms do not.

Fig. 5.10 shows a comparison of male and female drivers involved in fatal crashes. The bar graph shows that female drivers are involved in a fewer fatal crashes than male drivers.

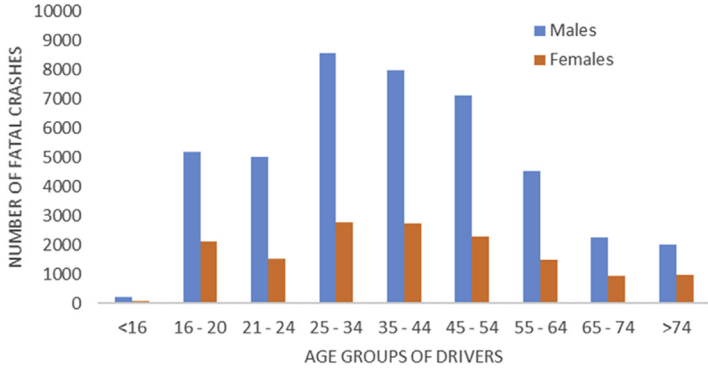

FIGURE 5.10 Male versus female drivers fatal crashes. *Based on data available at StatCrunch:* https://www.statcrunch.com/5.0/viewreport.php?reportid=35500.

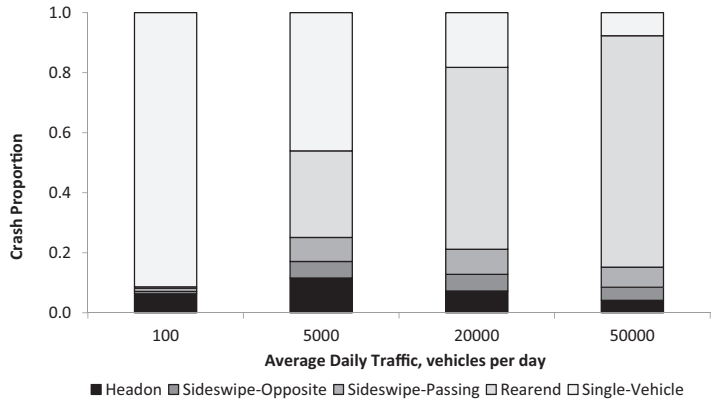

FIGURE 5.11 Crash proportion by collision type.

Instead of showing the rectangular bars side-by-side, they can be stacked, and those are called as stacked bar graphs. These bars are used to show how the total amount is split into smaller categories and what the relationship of each category has on the total amount. Stacked bar graphs can be constructed by placing categories one after the other. The total value of the bar is the sum of all categories. These graphs are ideal for comparing the total amounts across different groups. Alternatively, stack bar graphs can be constructed to show the percentage of each category when compared to the total amount in each group. Consequently, relative differences between quantities in each group are easily visible. The limitation with the stacked bar graphs is that they become harder to read with the large increase in categories. Fig. 5.11 shows that the proportion of collision types by different volume ranges for the crashes occurred in Texas on two-lane highway segments. The figure shows that single-vehicle crashes are more prevalent on low-volume roads and rear-end crashes are more prevalent on high-volume highways.

Another type of bar graph is known as the mosaic plot. This plot is similar to stacked bar graphs but adds another dimension, which is captured by the width of the column. The width can, for example, show the sample size for each category. Fig. 5.12 shows the proportion of crashes attributed to tire debris on Texas highways by speed limit (based on a sample of highways) when compared to remaining crash types (denoted as "Other") (Avelar et al., 2017). The width of the columns in this figure describes the proportion of the data sampled for each speed limit. The largest sample was for the speed limit of 65 mph.

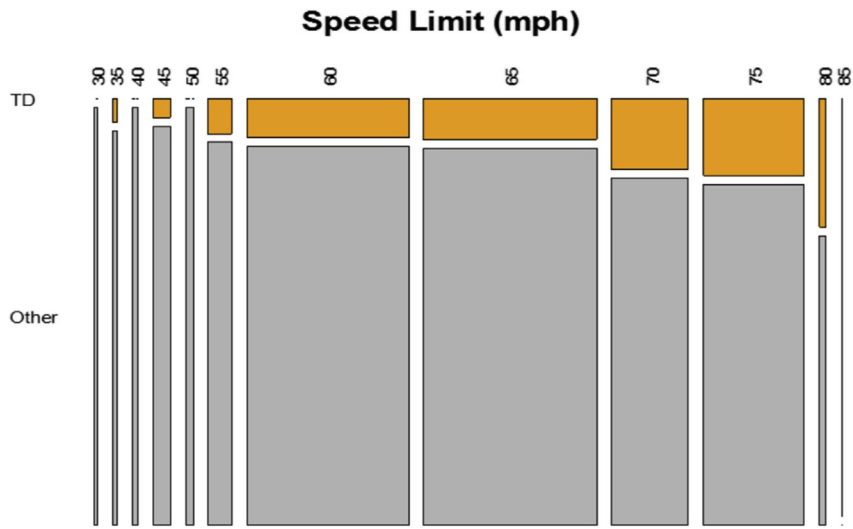

FIGURE 5.12 Crashes caused by tire debris by speed limit in Texas. *From Avelar, R.E., M.P. Pratt, J.D. Miles, T. Lindheimer, N. Trout, and J. Crawford (2017) report Develop Metrics of Tire Debris on Texas Highways: Technical Report. FHWA/TX-16/0-6860-1. Texas A&M Transportation Institute, College Station, TX.*

5.3.4 Error bars

An error bar indicates the uncertainty in the estimated measurement. Error bars are used to display the standard error to give a sense of range or spread of estimate. Although not common, they can be used to show a confidence interval or the minimum and maximum values in a dataset. Depending on whether the bar graphs are shown horizontally or vertically, the error bar lines must always run parallel to the bars. The length of the error bar reveals how much uncertainty exists in the estimate. For instance, a short error bar indicates that the estimate is accurate and a long error bar shows that the estimate is less reliable. The error bars can also be used to compare the difference in the estimate in different categories. Fig. 5.13 shows the error bars for run-off-the-road (ROR) events (crash and near-crash events) by age group for the drivers in the 100-Car Study (McLaughlin et al., 2009). The error bars reveal that the frequency of ROR events was not significantly different between participants in the 18−to 20-year-old and 21− to 24-year-old age groups. Similarly, there is no significant difference for the age groups above 25-year old. Although the estimate for the 35−to 44-year-old and 45− to 54-year-old age groups are similar (as shown by the tip of the bar or bar size), the length of error lines reveal that the estimate for 35−to 44-year-old is less reliable.

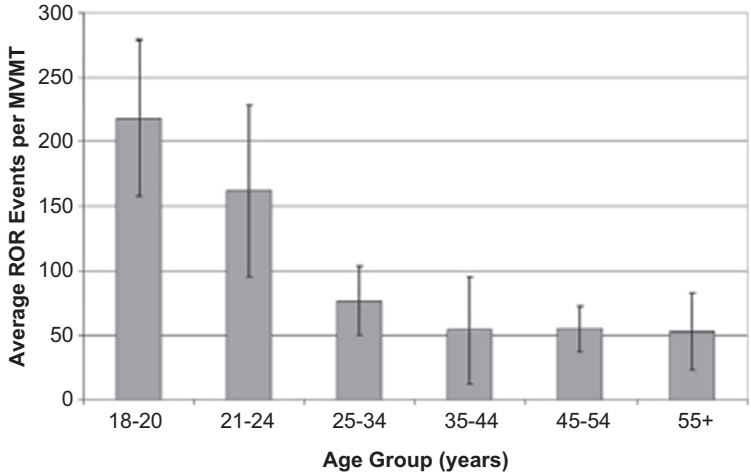

FIGURE 5.13 Average ROR events per MVMT by age group. *From McLaughlin, S.B., Hankey, J.M., Klauer, S.G., Dingus, T.A., 2009. report Contributing Factors to Run-Off-Road Crashes and Near-Crashes, National Highway Traffic Safety Administration, Report DOT HS 811 079.*

5.3.5 Pie charts

A pie chart is circular in shape and is used to show proportions of various categories, by dividing a circle into segments. In a pie chart, each arc length represents a proportion of each category, with all categories summing up to 100%.

Although pie charts provide a simple way of showing proportional distribution of the data, they are often criticized. They cannot show more than a few categories, because as the categories increase, the size of each section becomes smaller making them unsuitable for large categories of data. Fig. 5.14 shows the fatality composition of different vehicle types obtained from the FARS database for the years 2006 and 2015 (NHTSA, 2015). The pie charts show the percentage of passenger car occupant fatalities decreased from 43% in 2006 to 36% in 2015. The percentage of light-truck occupant fatalities decreased from 30% to 28% in the 10-year period. However, the charts show that the proportion of motorcyclist fatalities increased by 3%.

5.3.6 Scatterplots

A scatter plot uses dots to show the relationship between two variables. The two variables are displayed on horizontal and vertical axes to show if a relationship or an association between the two variables exists. The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

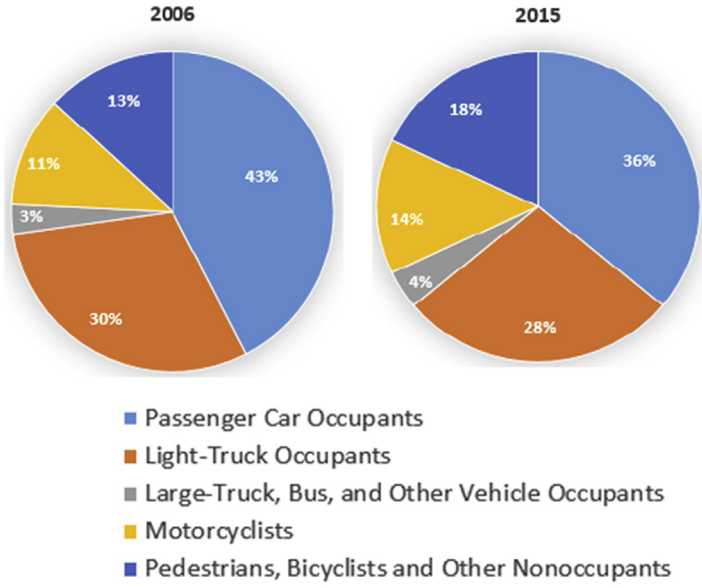

FIGURE 5.14 Fatality composition by vehicle type, 2006 and 2015. *Based on data available at: National Highway Traffic Safety Administration, report National Center for Statistics and Analysis. (2016, August). 2015 motor vehicle crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS 812 318). Washington, DC: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318.*

The patterns in scatterplots can be used to identify various types of correlation. Fig. 5.15 shows the correlations that are positive (value of one variable increases with other), negative (value of one variable decreases as the other increases), negligible or null (no correlation). The strength of the correlation can be determined by how closely one point exists when compared to the other point on the graph. Outliers are those points that are far outside the point clusters.

Sometimes, an equation is developed to fit a line or curve that can be used for predicting the future values via extrapolation. This fitted line is called the Line of Best Fit or a Trend Line. Scatterplots are ideal for paired numerical data to check the influence of one variable on the other. However, it is important to note that the correlation is not causation and another unobserved variable may be influencing results. Fig. 5.16 illustrates a scatterplot of traffic fatalities and vehicle-miles traveled (VMT) from 2005 to 2018 in Texas. A linear trend line is fitted and it can be used to estimate the fatalities for a given VMT in the future years.

FIGURE 5.15 Scatterplots showing types of correlation.

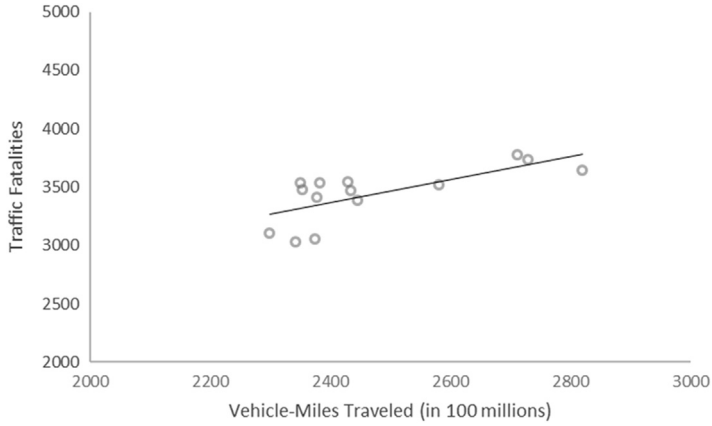

FIGURE 5.16 Miles driven and fatality rate scatterplot.

5.3.7 Bubble chart

Bubble chart is an extension of the scatterplot that is used to investigate the relationship between three variables. Each observation is represented by a circle (also called as bubble) and its location is indicated by the horizontal axis and vertical axis. The location of the circle provides the relationship between two variables (one variable shown on the horizontal axis and the second on vertical axis). The size of the circle illustrates the value of the third variable. Each circle can be displayed by a different color but it is not mandatory. Fig. 5.17 shows the number of vehicles registered and total lane-miles of the top five counties in Texas along with

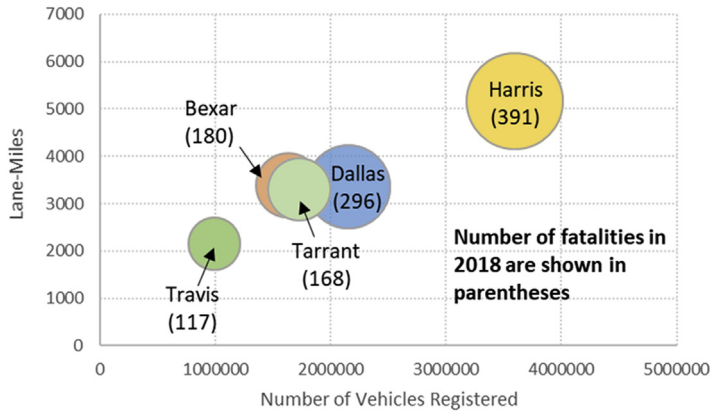

FIGURE 5.17 Bubble chart showing the relationship between fatalities, vehicles registered and lane-miles.

the number of fatalities in 2018. The size of the circle shows the number of fatalities and its location illustrates the relationship between vehicles registered and lane-miles. The figure reveals a positive correlation between fatalities and exposure variables.

5.3.8 Radar/web plot

A radar or web plot is a two-dimensional figure used for examining several variables at the same time or on the same plane for a single unit. The figure resembles more or less a bullseye where each variable is represented by an axis that originates in the middle of the radar plot, usually represented, but not always, by the value 0, and extends to its margin. The analyst puts the value for each variable along the axis. Then, a line is used to connect each point on all the axes to form a polygon. Often, the area inside the polygon can be colored to enhance the visual presentation. Other units can be added to the plot using different colors. Although the radar plot can be a powerful visual tool, the figure can be difficult to read as more units are added, is dependent on the scale used for each axis (i.e., changing the scale can influence the perception of the relationship between the variables), and is not able to able to rank the units for the variables investigated. Fig. 5.18 shows a radar plot for the relationship between the average speed with low-speed variation at upstream and average speed with high-speed variation at downstream of a location on an urban freeway section (see Exercise 10.1 in Chapter 10—*Capacity, Mobility and Safety*).

5.3.9 Heatmap

It is not always easy to understand the data if presented numerically. A heat map is a representation of data in the form of a map in which data values are represented by colors. Heat maps provide a convenient tool to communicate relationships between data values and to explore large datasets. For large-scale datasets, it is crucial for data to be properly classified and visualized for proper interpretation. Heat maps make use of clustering analysis and kernel density estimation methods to show diverse datasets in an effective and efficient manner. These methods are described in detail in Chapter 9—*Models for Spatial Data*. Unlike other data visualizations, heat maps are self-explanatory. The darker the shade, the greater the frequency/severity. When heat maps are used in combination with other existing data visualizations, it becomes much easier to understand key data elements. Fig. 5.19 provides a heat map of high-risk locations for crashes and crimes in College Station, Texas from 2005 to 2010 (Kuo et al., 2013). The hotspots for crashes are represented by warm

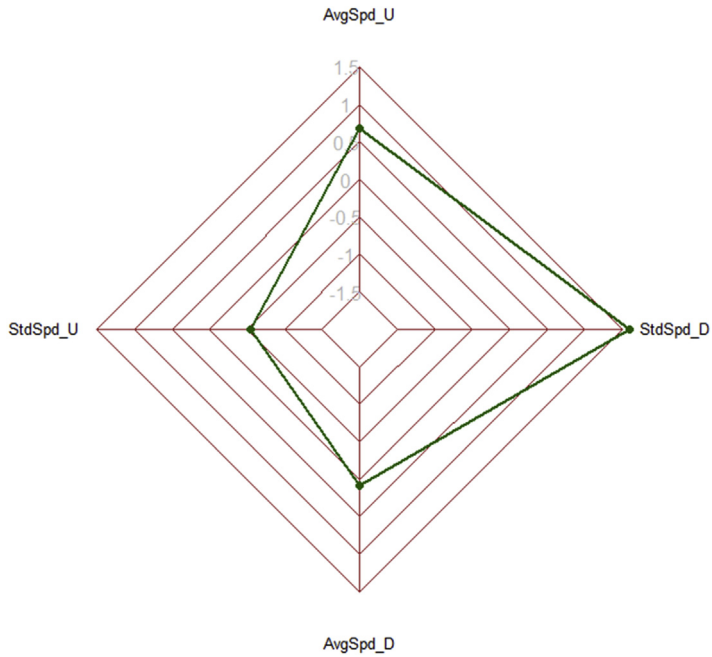

FIGURE 5.18 Relationship between the average speed with low-speed variation at upstream and average speed with high-speed variation at downstream of an urban freeway segment (see Exercise 10.1).

colors (i.e., red [mild gray in printed version] and yellow [light gray in printed version]) and cold colors (i.e., purple [gray in printed version] and blue [dark gray in printed version]) are used for crime hotspots. The heat maps show a spatial relationship between crashes and crimes because the hotspots are near each other.

5.3.10 Contour plot

A contour plot is a graphical technique that uses constant z-slices, called contours, on a two-dimensional plane to show a three-dimensional surface. One predictor variable is represented on the horizontal axis and a second predictor variable is represented on the vertical axis. The third variable is the response variable shown as contours represented by a color gradient and isolines (lines of constant value). Fig. 5.20 shows a contour plot of the predicted slight injury crashes per kilometers traveled as a function of speed and Average Annual Daily Traffic (Imprialou et al., 2016). The plot was developed using the crash data obtained from the National Road Accident Database of the United Kingdom for the year 2012.

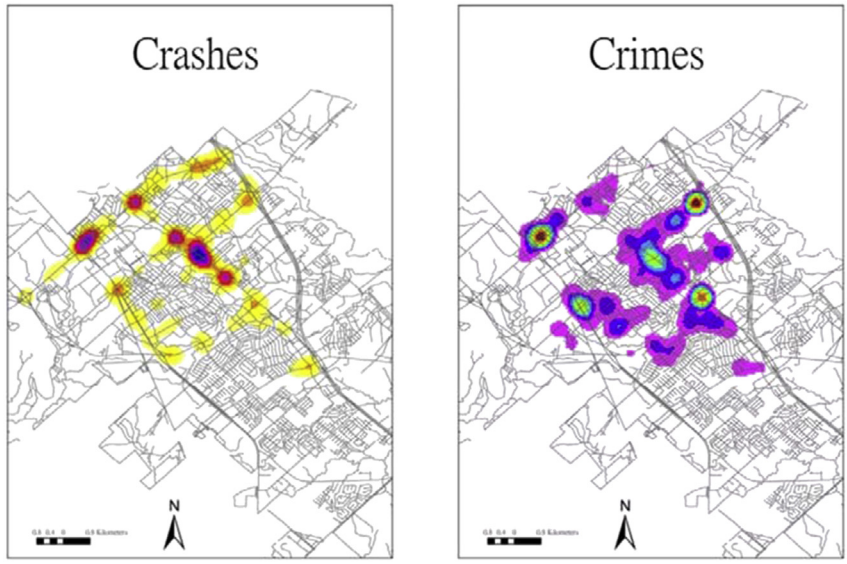

FIGURE 5.19 Heat map of high-risk locations for crashes and crimes. *From Kuo, P.-F., D. Lord, and T.D. Walden (2013) Using geographical information systems to organize police patrol routes effectively by grouping hot spots of crash and crime data. J. Transport Geogr., Vol. 30 (June), pp. 138–148.*

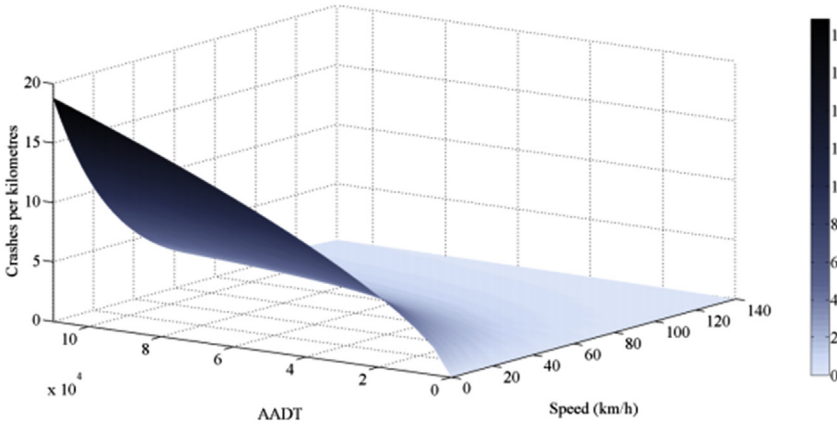

FIGURE 5.20 Contour plot of slight injury crashes. *From Imprialou, M-I, M. Quddus, D. Pitfield, D. Lord (2016) Re-visiting crash-speed relationships: a new perspective in crash modelling. Accid. Anal. Prev., Vol. 86, pp. 173–185.*

5.3.11 Population pyramid

A population pyramid illustrates graphically the distribution of various age groups in a population by gender for a particular variable,

which forms the shape of a pyramid when the population is growing. The horizontal axis is used to plot the population frequency or relative frequency and the vertical axis lists all age groups. Population Pyramids are ideal for identifying changes or differences in population patterns. For comparing patterns across various population groups, multiple population pyramids can be used. The shape of a population pyramid can be used to interpret a population. For example, a pyramid with a wide base and a narrow top section suggests that the younger population has more risk than the older population.

Fig. 5.21 shows the population pyramid for fatal and serious injury speeding crashes in Texas. It shows the overrepresentation for specific age

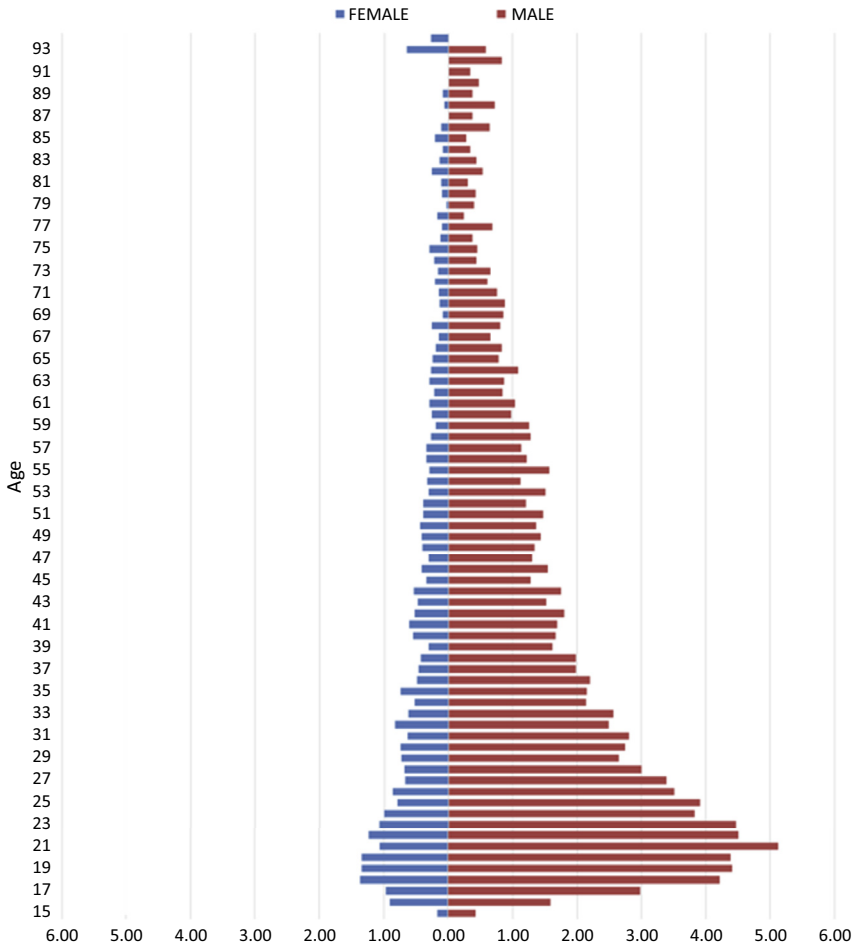

FIGURE 5.21 Population pyramid for fatal and serious injury speeding crashes. *This figure is taken from the link: https://www.texasshsp.com/wp-content/uploads/2019/02/SHSP-2019-v3.pdf.*

and gender groups after taking into account population size. The population pyramid helps to identify age and gender groups that experienced a greater number of fatalities and serious injuries than would be expected given their population size. Instead of showing the crash frequency, an index is computed to identify overrepresentation. The index is defined as the proportion of crashes relative to (or divided by) the proportion of the population occupied by each age/gender group. If the index values are over 1.00, then it indicates an excess for that age and gender group. The population pyramid shows that the males in Texas are greatly overrepresented across nearly all age groups. Among females, the problem is greatest among those under 25 years.

References

Adeloye, D., Thompson, J.Y., Akanbi, M.A., Azuh, D., Samuel, V., Omoregbe, N., et al., 2016. The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and metaanalysis. Bull. World Health Organ. 94 (7), 510–21A.

Avelar, R.E., Pratt, M.P., Miles, J.D., Lindheimer, T., Trout, N., Crawford, J., 2017. Develop Metrics of Tire Debris on Texas Highways: Technical Report. FHWA/TX-16/0-6860-1. Texas A&M Transportation Institute, College Station, TX.

Berry, D.S., Belmont, D.M., 1951. Distribution of vehicle speeds and travel times. In: Proceedings of 2nd Berkeley Symposium on Mathematical and Statistical Probabability, pp. 589–602.

Hallmark, S.L., Tyner, S., Oneyear, N., Carney, C., McGehee, D., 2015. Evaluation of driving behavior on rural 2-lane curves using the shrp 2 naturalistic driving study data. J. Saf. Res. 54, 17–27.

Hinkle, D.E., Wiersma, W., Jurs, S.G., 2003. Applied Statistics for the Behavioral Sciences, fifth ed. Houghton Mifflin, Boston.

Imprialou, M.-I., Quddus, M., Pitfield, D., Lord, D., 2016. Re-visiting crash-speed relationships: a new perspective in crash modelling. Accid. Anal. Prev. 86, 173–185.

Jiang, X.M., Yan, X.D., Huang, B.S., Richards, S.H., 2011. Influence of curbs on traffic crash frequency on high-speed roadways. Traffic Injury Prevent. 12 (No. 4), 412–421.

Kuo, P.-F., Lord, D., Walden, T.D., 2013. Using geographical information systems to organize police patrol routes effectively by grouping hot spots of crash and crime data. J. Transport Geogr. 30 (June), 138–148.

Ma, L., Yan, X., Weng, J., 2015. Modeling traffic crash rates of road segmentsthrough a lognormal hurdle framework with flexible scale parameter. J. Adv.Transp. 49 (8), 928–940.

McLaughlin, S.B., Hankey, J.M., Klauer, S.G., Dingus, T.A., 2009. Contributing Factors to Run-Off-Road Crashes and Near-Crashes. National Highway Traffic Safety Administration. Report DOT HS 811 079.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26 (4), 471–482.

National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.

National Highway Traffic Safety Administration, 2015. National Center for Statistics and Analysis. (2016, August) motor vehicle crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS 812 318). Washington, DC. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318.

Owens, J.M., Tefft, B., Guo, F., Fang, Y., Perez, M., McClafferty, J., Dingus, T., 2018. Crash Risk of Cell Phone Use while Driving: Case-Crossover Study of SHRP 2 Naturalistic Driving Data. AAA Foundation for Traffic Safety Report, Washington, D.C.

Pratt, M.P., Geedipally, S., Dadashova, B., Wu, L., Shirazi, M., 2019. Familiar versus unfamiliar drivers on curves: a naturalistic data study. Transport. Res. Rec.: J. Transport. Res. Board 2673 (6), 225–235.

UF Biostatistics. Open Learning Textbook. Paired Samples. Data available at: https://bolt.mph.ufl.edu/6050-6052/unit-4b/module-13/paired-t-test/#drinking.