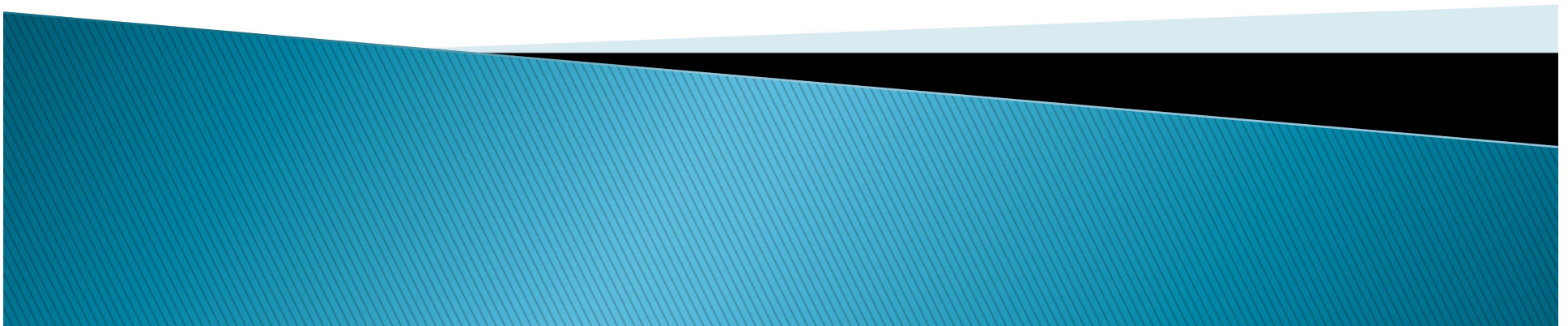


Exploratory Analyses of Crash Data

Fall 2021



How to “map” crashes on network

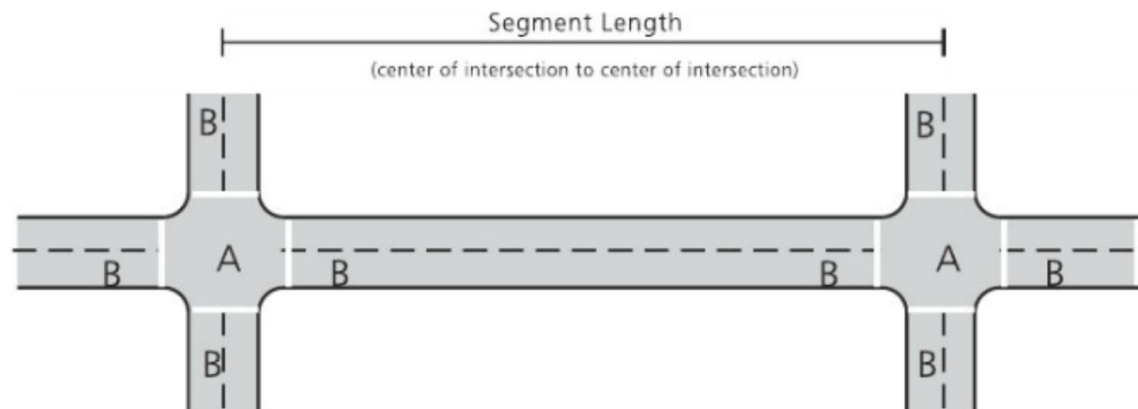
- ▶ In the previous lecture, I showed how to link different databases together.
- ▶ Usually, we link them via the location (control-section or GIS) and then match the crashes to the characteristics of the location.
- ▶ Not only do we need to match the crashes, but we also need to use the total sampling frame, even those that do not include crashes.
- ▶ Even the sites or observations that have 0 crash provides information.



Sampling Frame

Sampling frame: the sampling frame is the list of the population (this is a general term) from which the sample is drawn. It is important to understand how the sampling frame defines the population represented.

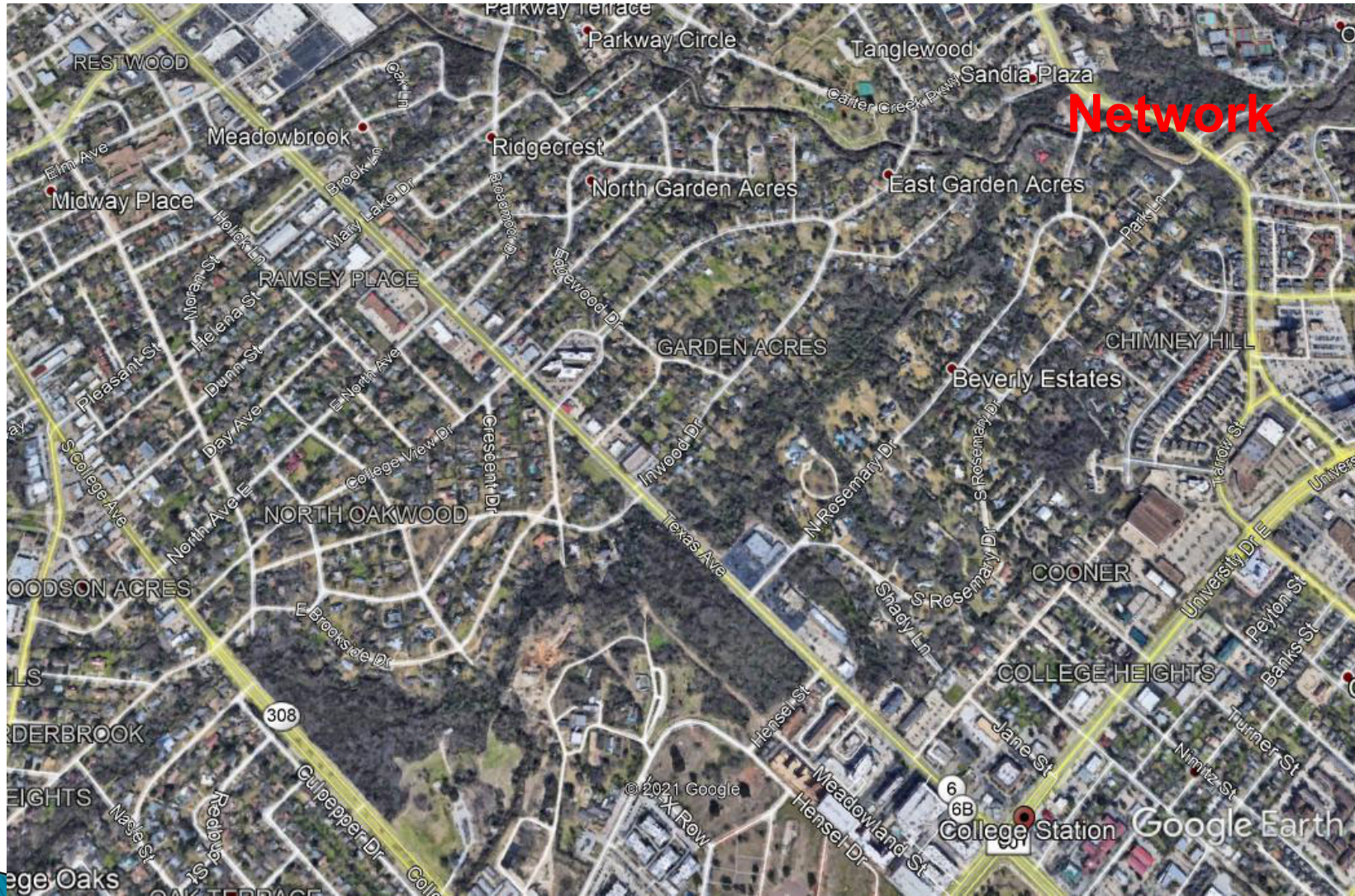
Example: Texas Ave is separated into 8 segments (next few slides). Crashes have been reported for two years. For this exercise, we are grouping section and intersection crashes together. Usually, we separate intersections from those that occurred on segments/sections. See Highway Safety Manual.



- A All crashes that occur within this region are classified as intersection crashes.
- B Crashes in this region may be segment or intersection related, depending on the characteristics of the crash.

Figure 12-2. Definition of Roadway Segments and Intersections

Sampling Frame



Sampling Frame



Mapping Crashes



Mapping Crashes




Assembling Database

Database separated by year

Site #	Year	Crashes	Variable 1	Variable 2
1	2020	2		
2	2020	1		
3	2020	0		
4	2020	0		
5	2020	2		
6	2020	1		
7	2020	0		
8	2020	3		
1	2021	2		
2	2021	0		
3	2021	1		
4	2021	0		
5	2021	1		
6	2021	2		
7	2021	0		
8	2021	4		

Variables: AADT, segment length,
lane width, shoulder width, etc.



Assembling Database

Aggregated Data

Site #	Number of Years	Crashes	Variable 1	Variable 2
1	2	4		
2	2	1		
3	2	1		
4	2	0		
5	2	3		
6	2	3		
7	2	0		
8	2	7		

Variables: AADT, lane width, shoulder width, etc. But they need to stay the same for the time period. Not all variables stay the same, such as AADT. For that variable, you need to take the average for the aggregated time period. For some, if they change during the period, the site will need to be removed.



Objectives

Exploratory data analyses 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.

Note. check Chapter 5 for all the important equations.

Quantitative Techniques

- ▶ Measures of central tendency
 - Mean, Median, Mode
- ▶ Measures of variability
 - Range, Quartiles and interquartile range, Variance, standard deviation and standard error, Coefficient of variation
 - Symmetrical and asymmetrical data, Skewness, Kurtosis (measure of the sharpness of the peak of a frequency distribution; see next slide).



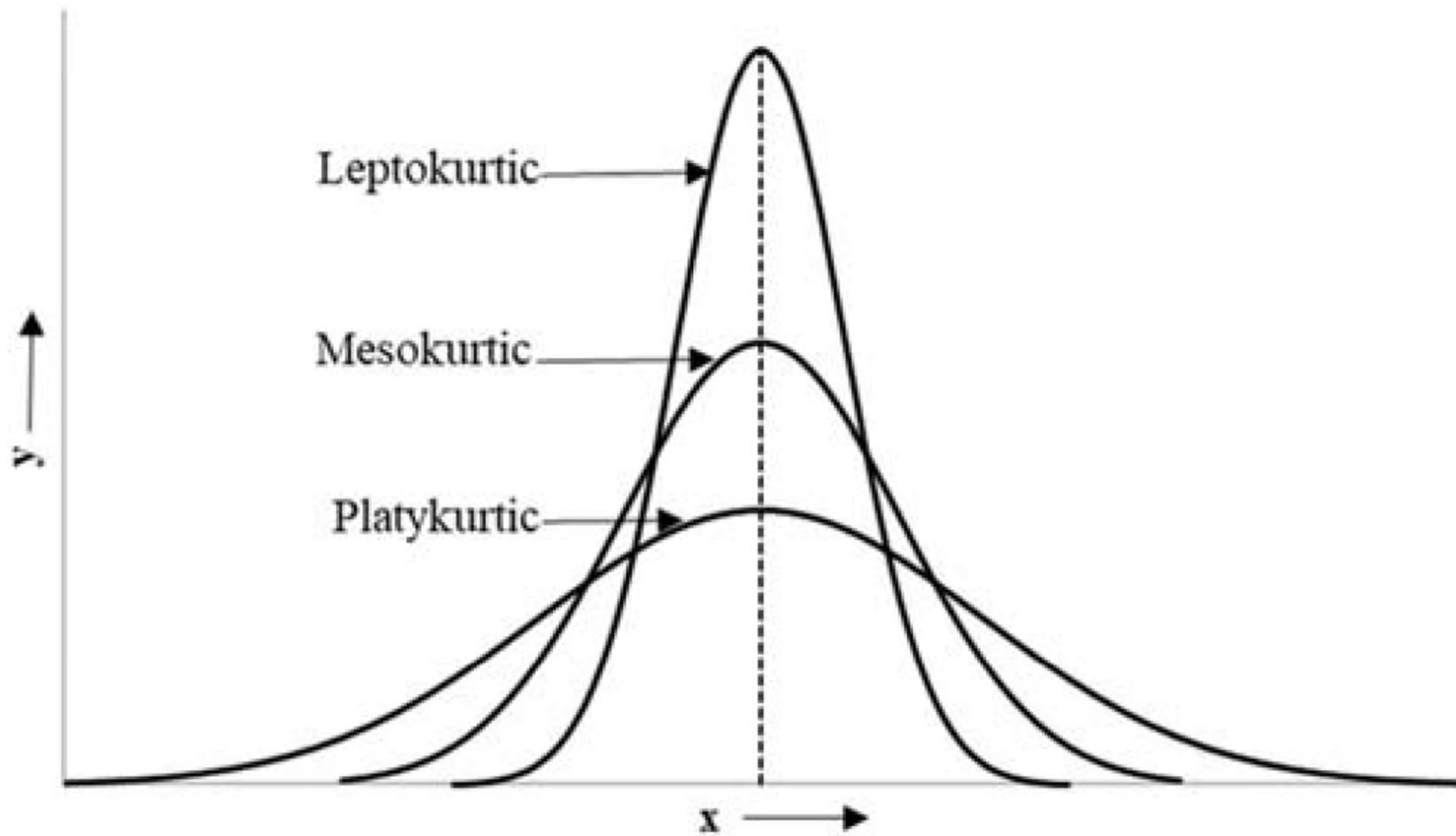


FIGURE 5.2 Kurtosis in the normal curve.



Quantitative Techniques

- ▶ Measures of association
 - Pearson's correlation coefficient, Spearman rank-order correlation coefficient, Chi-square test for independence (next slide), Relative risk and odds ratio (below)

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

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

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



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.



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.

Group	Outcome	
	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.

Confidence Intervals

Statistics are usually calculated from samples, such as the sample average \bar{X} , variance s^2 , the standard deviation s , are used to estimate the population parameters. For instance:

\bar{X} is used as an estimate of the population μ_x

s^2 is used as an estimate of the population variance σ^2

Interval estimates, defined as Confidence Intervals, allow inferences to be drawn about the population by providing an interval, a lower and upper value, within which the unknown parameter will lie with a prescribed level of confidence. In other words, the true value of the population is assumed to be located within the estimated interval.



Confidence Intervals

Confidence intervals for unknown mean and known standard deviation

$$\bar{x} \pm Z \frac{\sigma}{\sqrt{n}}$$

Confidence intervals for unknown mean and unknown standard deviation

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

Confidence intervals for proportions

$$\hat{p} \pm Z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$



Confidence Intervals

Confidence intervals for the population variance and standard deviation

$$\frac{(n-1)s^2}{\chi_{\frac{(1-C)}{2}, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\frac{(1-C)}{2}, n-1}^2}$$

$$\sqrt{\frac{(n-1)}{\chi_{\frac{(1-C)}{2}, n-1}^2}} s \leq \sigma \leq \sqrt{\frac{(n-1)}{\chi_{1-\frac{(1-C)}{2}, n-1}^2}} s$$



Confidence Intervals

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 (significance level = $\frac{1-C}{2} = \frac{1-0.95}{2} = 0.025$) is 1.96. The confidence interval for the truck proportion is obtained as

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

Hypothesis Testing

Test to determine whether a hypothesis is true or not based on sample data.

Step 1 – State the hypothesis

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

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

Both hypotheses are mutually exclusive.

Step 2 – Select confidence interval

This step involves selecting the appropriate confidence interval (C). The significance level could be equal to 0.01, 0.05 or 0.10.



Hypothesis Testing

Step 3 – Choose the test method and compute 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.



Hypothesis Testing

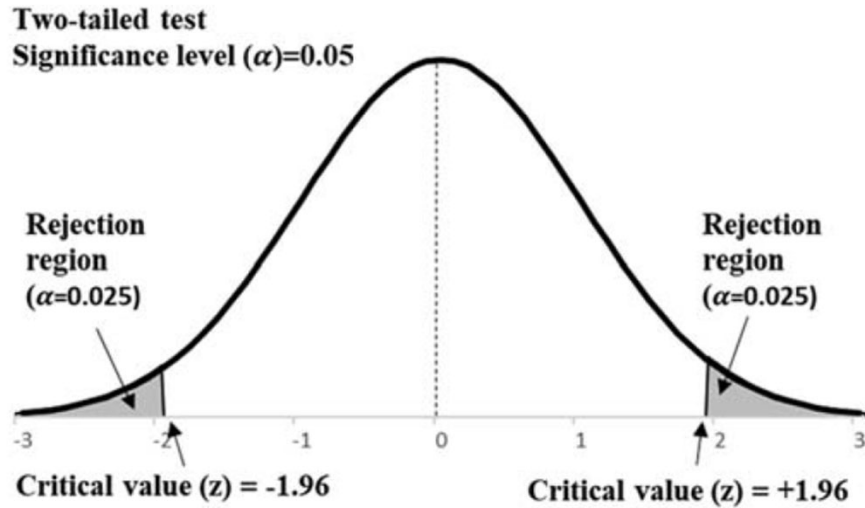


FIGURE 5.3 Critical values for a two-tailed (nondirectional) test.

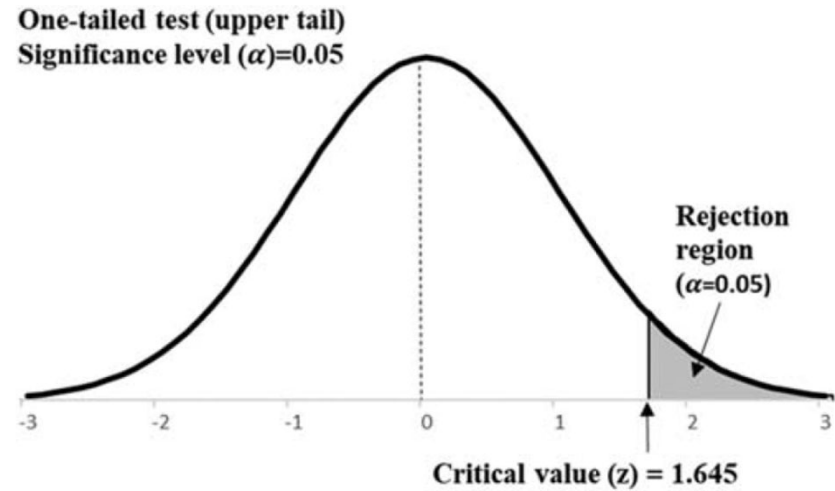


FIGURE 5.4 Critical values for a one-tailed (directional) test.



Hypothesis Testing

Decision Errors

- **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 $1 - \beta$.

Two-tailed hypothesis test

The two-tailed test is a method in which the rejection region is on two sides of the sampling distribution.

One-tailed hypothesis test

A one-tailed test is a statistical test in which the rejection region is on one side of the sampling distribution.

Hypothesis Testing

Hypothesis testing for one sample

When the population mean and standard deviation are known, the z statistic is calculated as

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

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

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

You compare the value with the one documented in the table on the next slide.



Hypothesis Testing

Hypothesis testing for one sample

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

See Chapter 5 in textbook for the proportion and variance calculations.



Hypothesis Testing

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.




Hypothesis Testing

Hypothesis testing for two samples

Dependent Samples

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 difference is then tested using the same equation described above with $\mu=0$:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$


Hypothesis Testing

Hypothesis testing for two samples

Independent Samples

The parameters tested using independent samples are either population means or population proportions. For this kind of analysis, the sample size (n) and the standard deviation (s) will be different for each population. The testing will be dependent on the sample size for both populations. When it is large, the normal distribution can be used and when they are small (~ 5 to 25), the student-t distribution needs to be used.

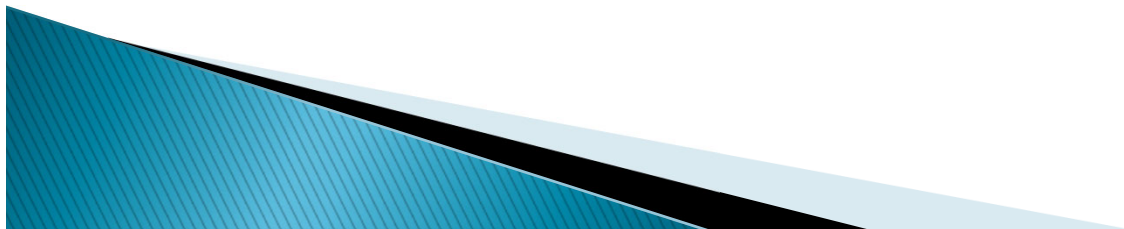
See section 5.2.5.5 in Chapter 5 for more details.

Multiple Independent Samples: Use ANOVA (also described in Chapter 5)

Summary Statistics

Table 1
Summary Statistics of datasets

Dataset	Variables	Min	Max	Average	Standard Deviation
Texas	Number of crashes	0	15	0.86	1.65
	Average 5-years AADT (vpd)	43	1166	313.8	253
	Segment length (miles)	0.10	4.41	0.96	0.93
Virginia	Number of crashes	0	8	2.01	2.09
	Average 5-years AADT (vpd)	163	5180	694	625
	Segment length (miles)	0.13	5.67	1.35	1.08



Graphical Methods

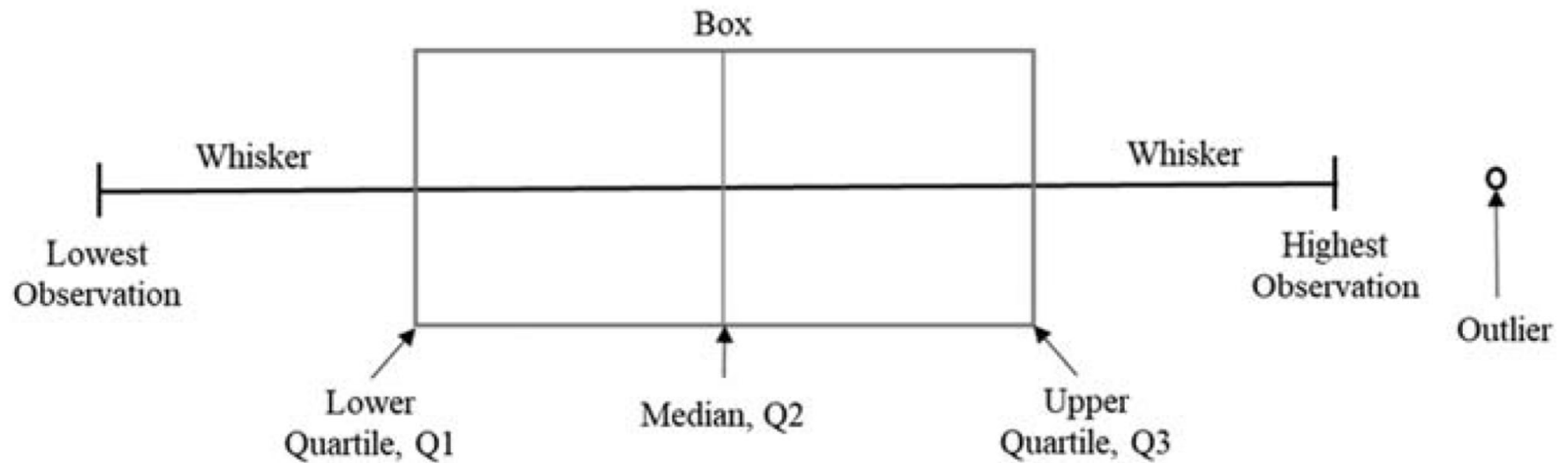


FIGURE 5.5 Box plot showing different measures.

Box Plot and Whiskers

Graphical Methods

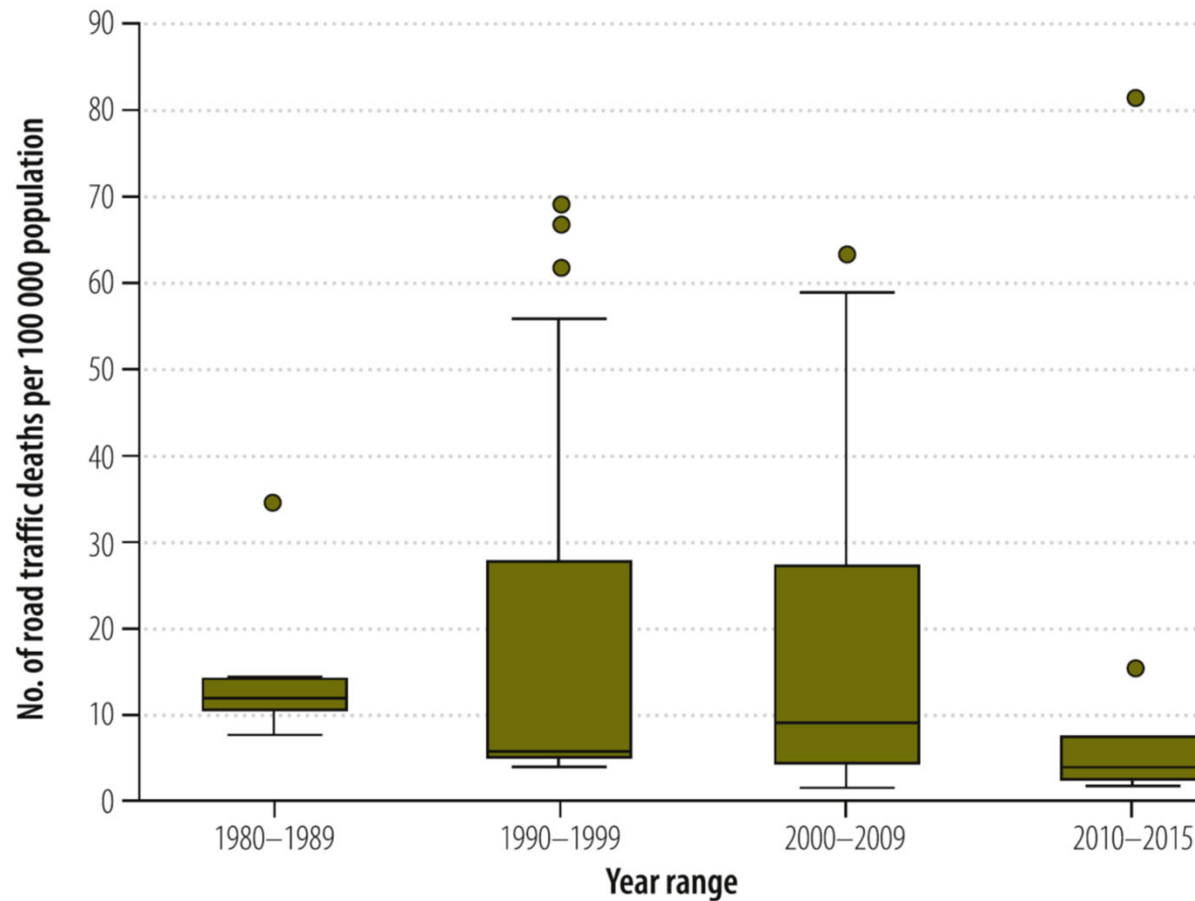
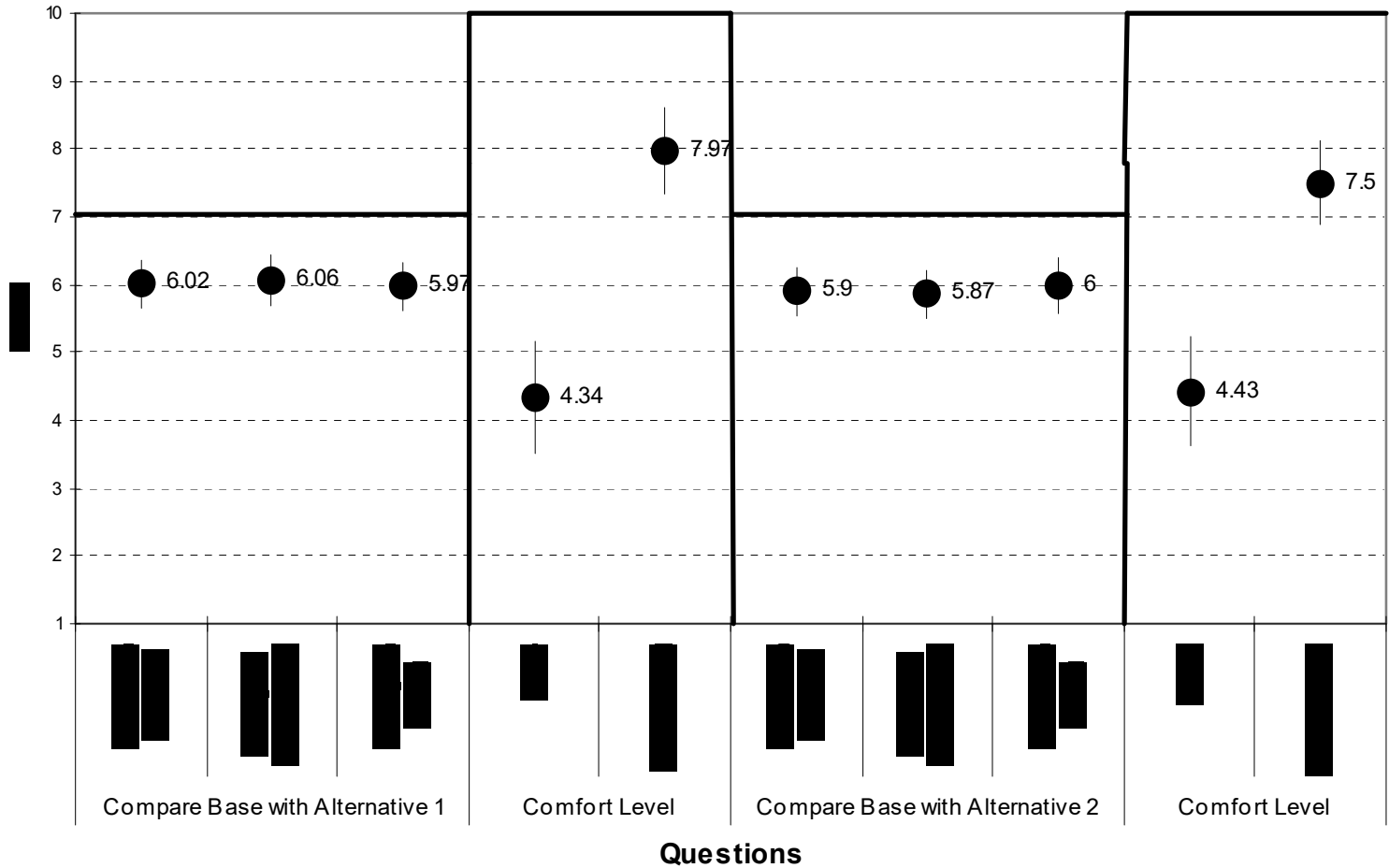


FIGURE 5.6 Box plot showing the traffic death rate in Africa. From Adeloje D, Thompson JY, Akanbi MA, Azuh D, Samuel V, Omoregbe N, et al. *The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and metaanalysis.* Bull World Health Organ. 2016;94(7):510–21A.

Bar Graphs

Graphical Methods



Box Plot



Graphical Methods

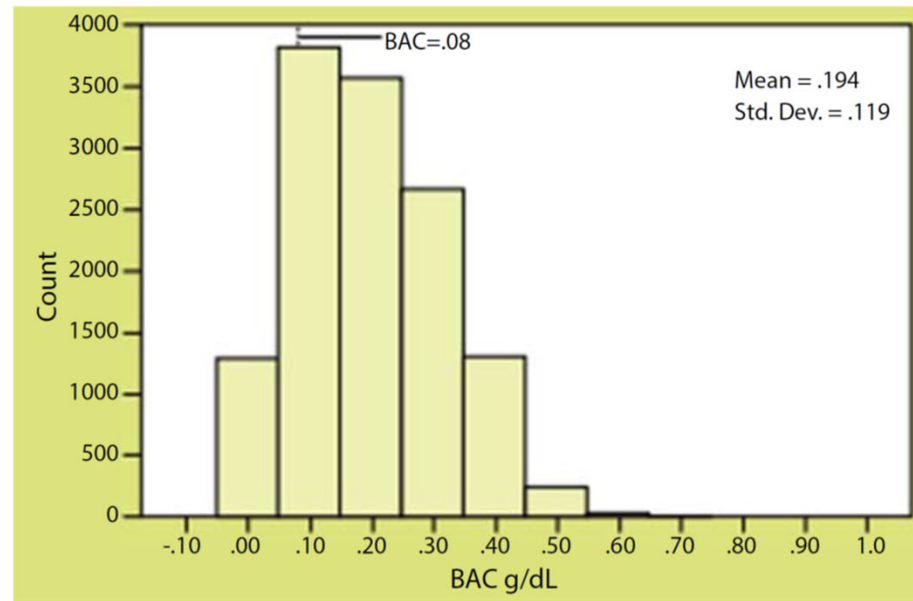


FIGURE 5.7 Histogram of passenger car driver BAC values. From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754>.

Histograms

Graphical Methods

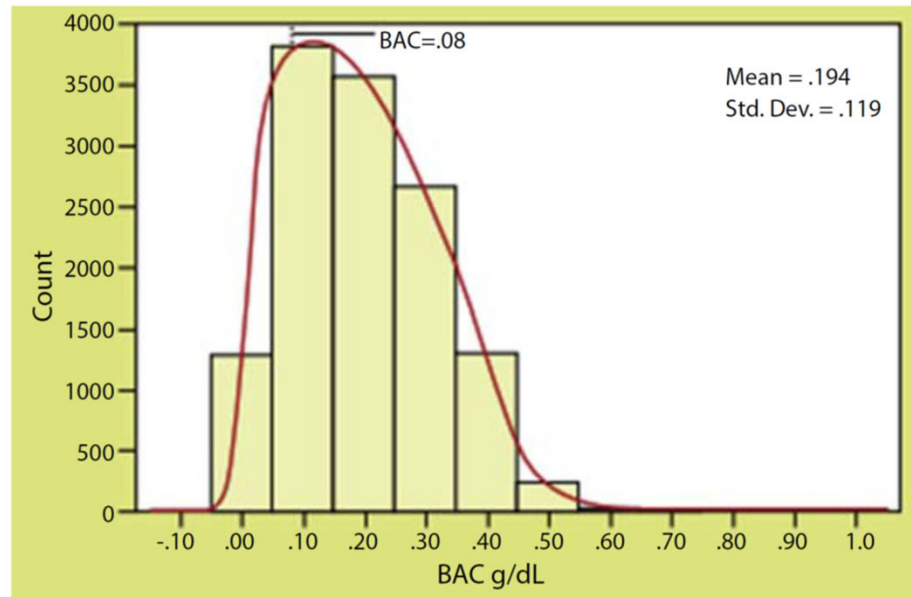


FIGURE 5.8 Histogram of passenger car driver BAC values with kernel density superimposed. From National Highway Traffic Safety Administration, 2007. *Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type.* <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754>.

Histograms

Graphical Methods

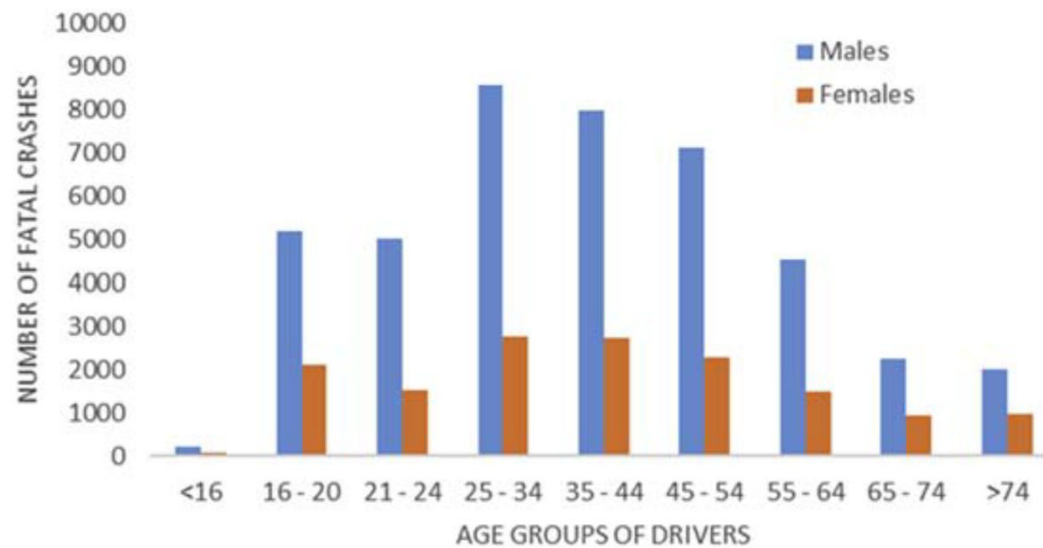
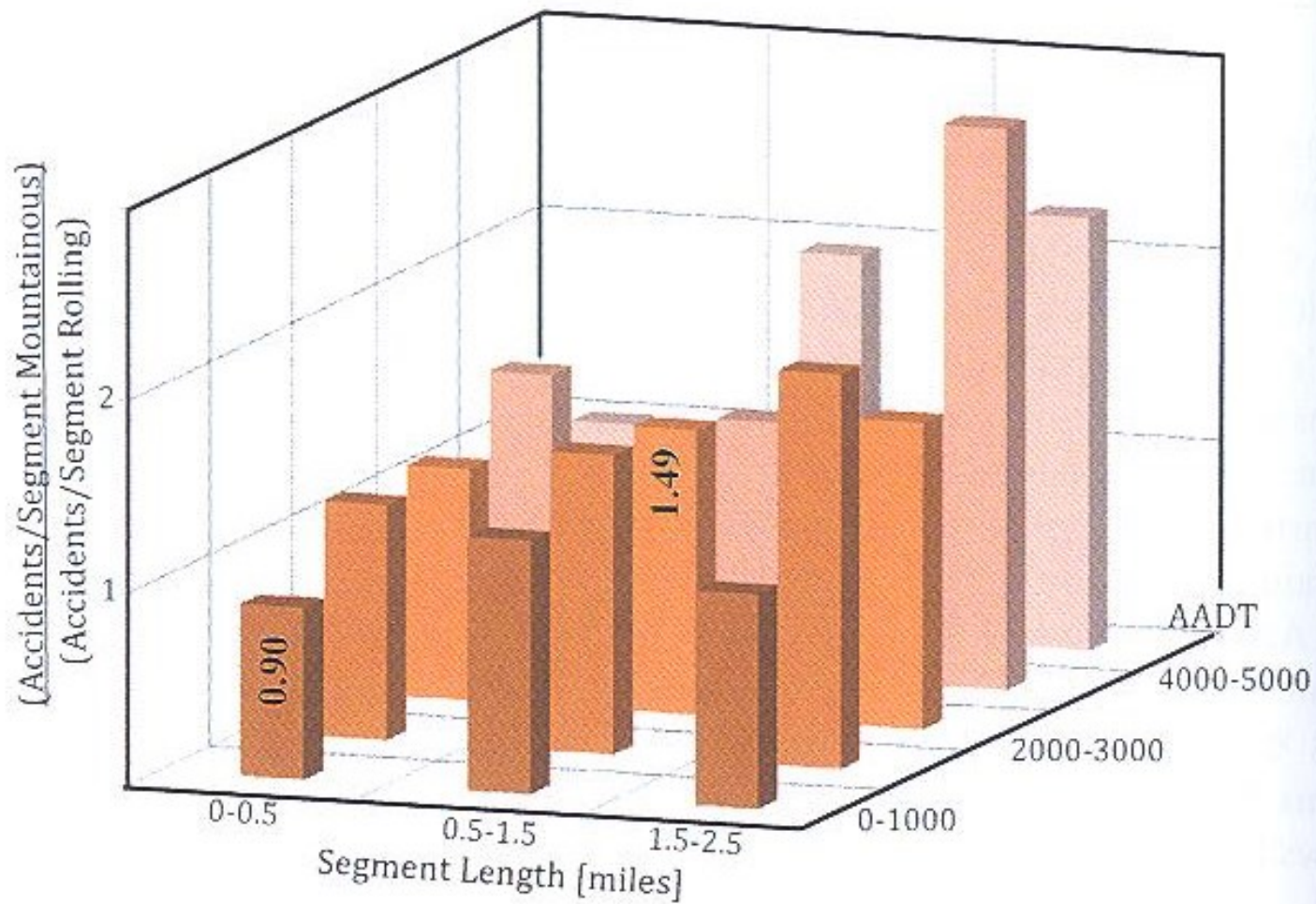


FIGURE 5.10 Male versus female drivers fatal crashes. Based on data available at StatCrunch: <https://www.statcrunch.com/5.0/viewreport.php?reportid=35500>.

Bar Graphs

Graphical Methods

3 Exploratory Data Analy



3D Histograms

Graphical Methods

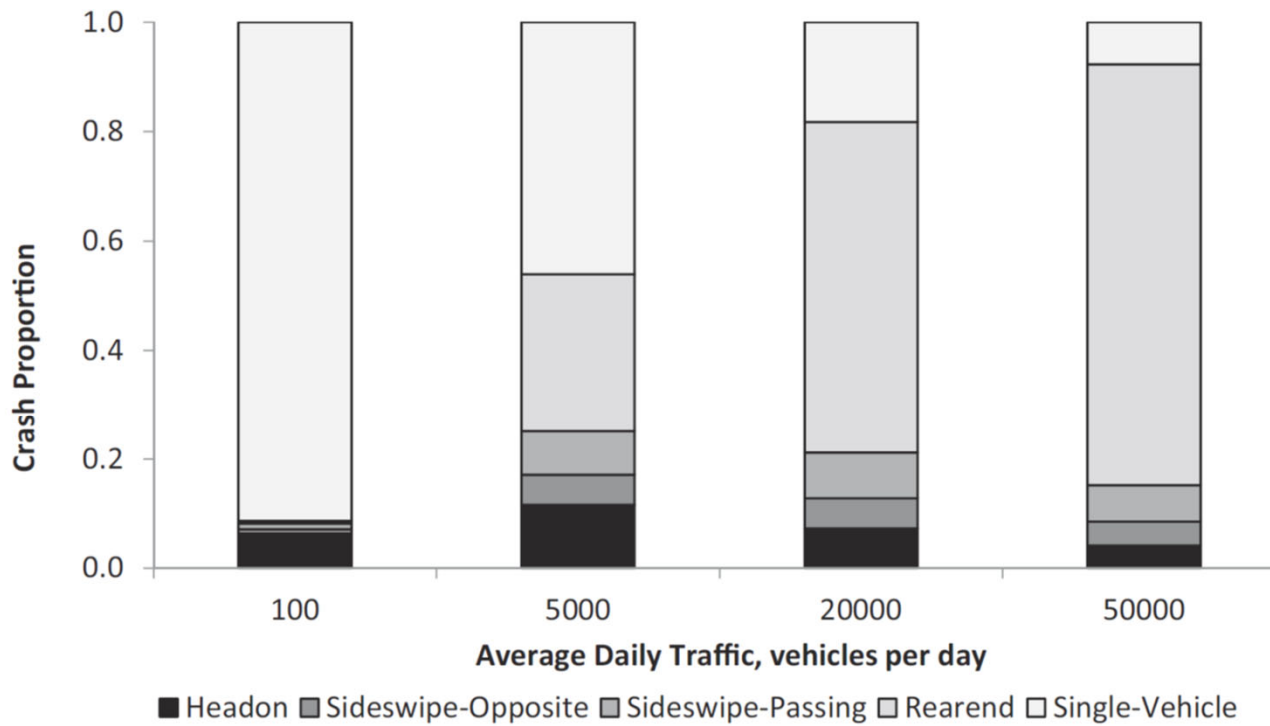


FIGURE 5.11 Crash proportion by collision type.

Histogram with Proportions



Graphical Methods

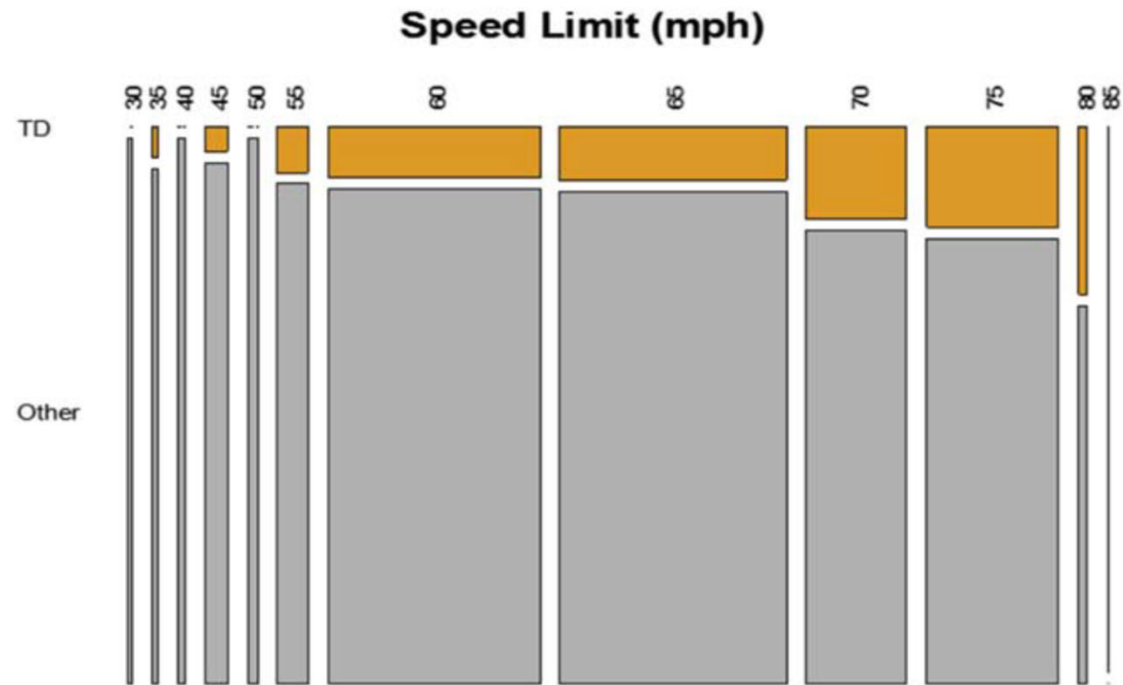


FIGURE 5.12 Crashes caused by tire debris by speed limit in Texas. From Avelar, R.E., M.P. Pratt, J.D. Miles, T. Lindheimer, N. Trout, and J. Crawford (2017) report *Develop Metrics of Tire Debris on Texas Highways: Technical Report. FHWA/TX-16/0-6860-1*. Texas A&M Transportation Institute, College Station,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

Mosaic Plot



Graphical Methods

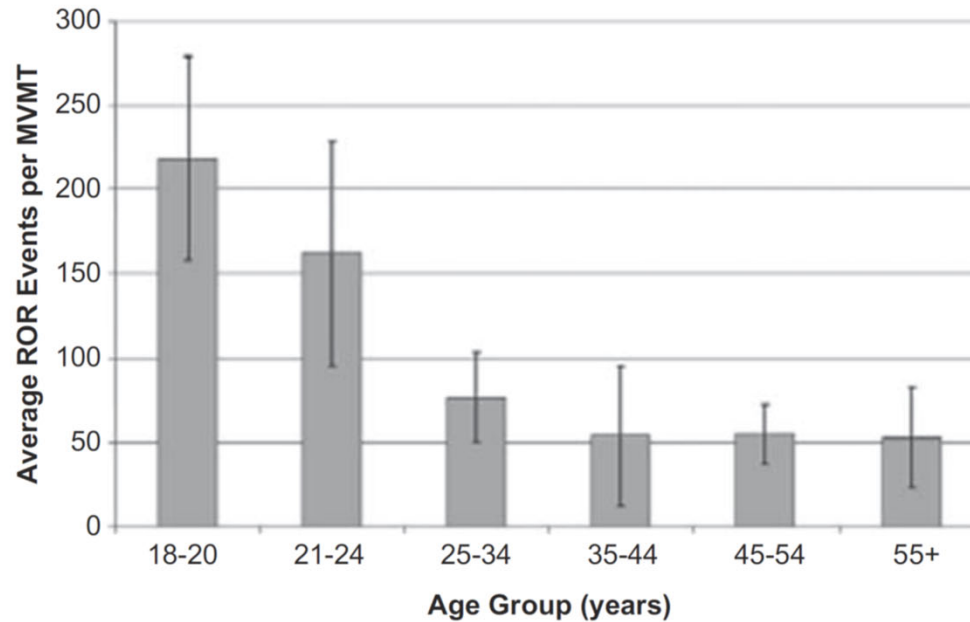


FIGURE 5.13 Average ROR events per MVMT by age group. From McLaughlin, S.B., Hankey, J.M., Klauer, S.G., Dingus, T.A., 2009. report *Contributing Factors to Run-Off-Road Crashes and Near-Crashes*, National Highway Traffic Safety Administration, Report DOT HS 811 079.

Error plots

Graphical Methods

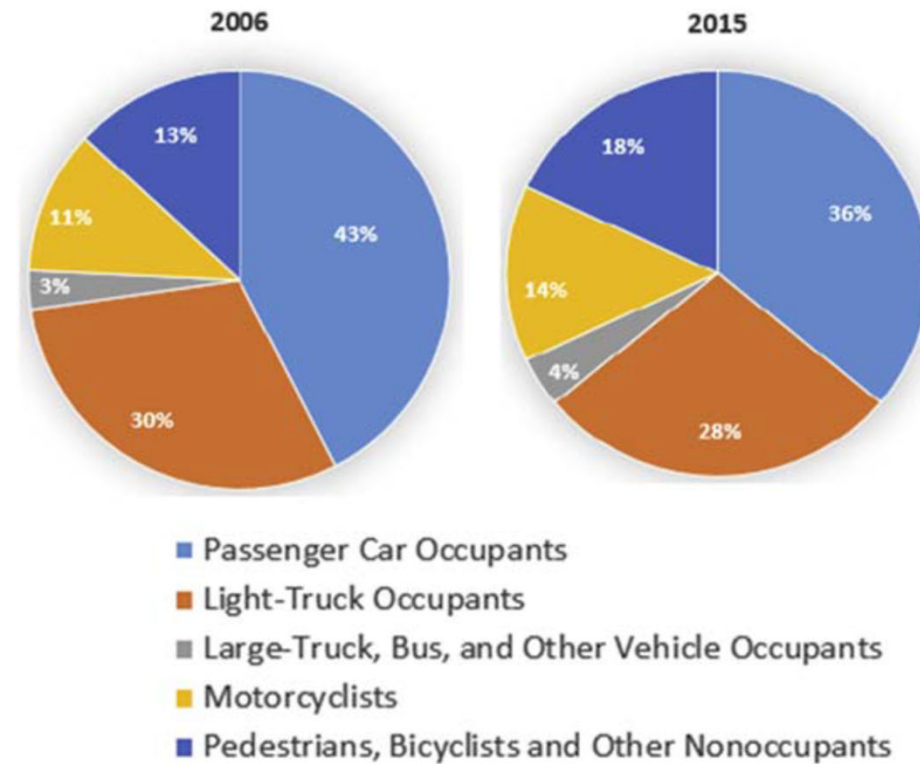


FIGURE 5.14 Fatality composition by vehicle type, 2006 and 2015. Based on data available at: National Highway Traffic Safety Administration, report National Center for Statistics and Analysis. (2016, August). 2015 motor vehicle crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS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

Pie charts

Graphical Methods

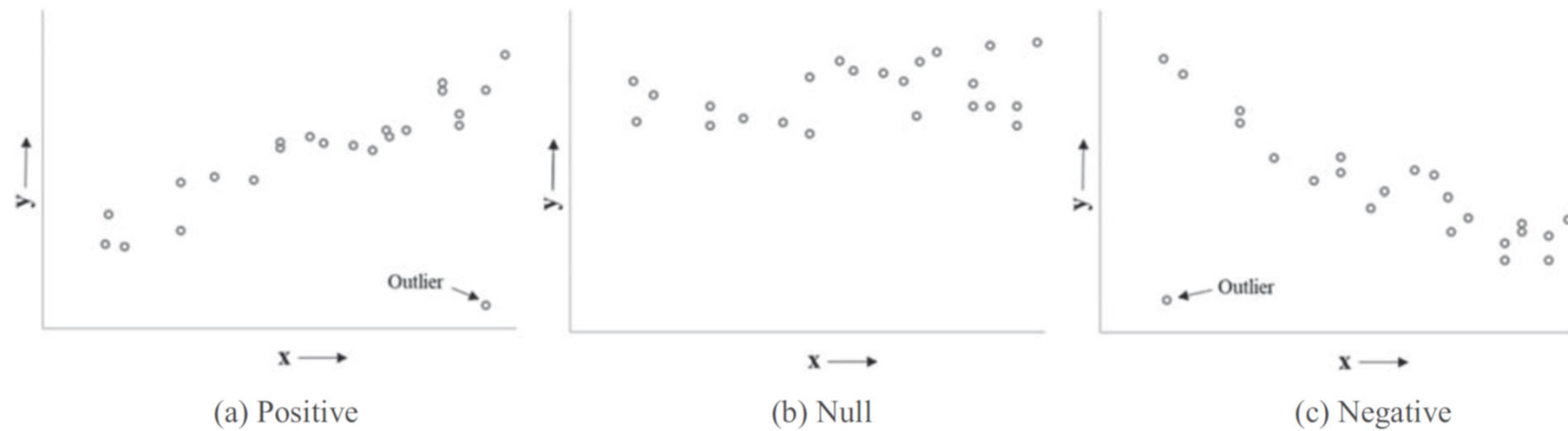
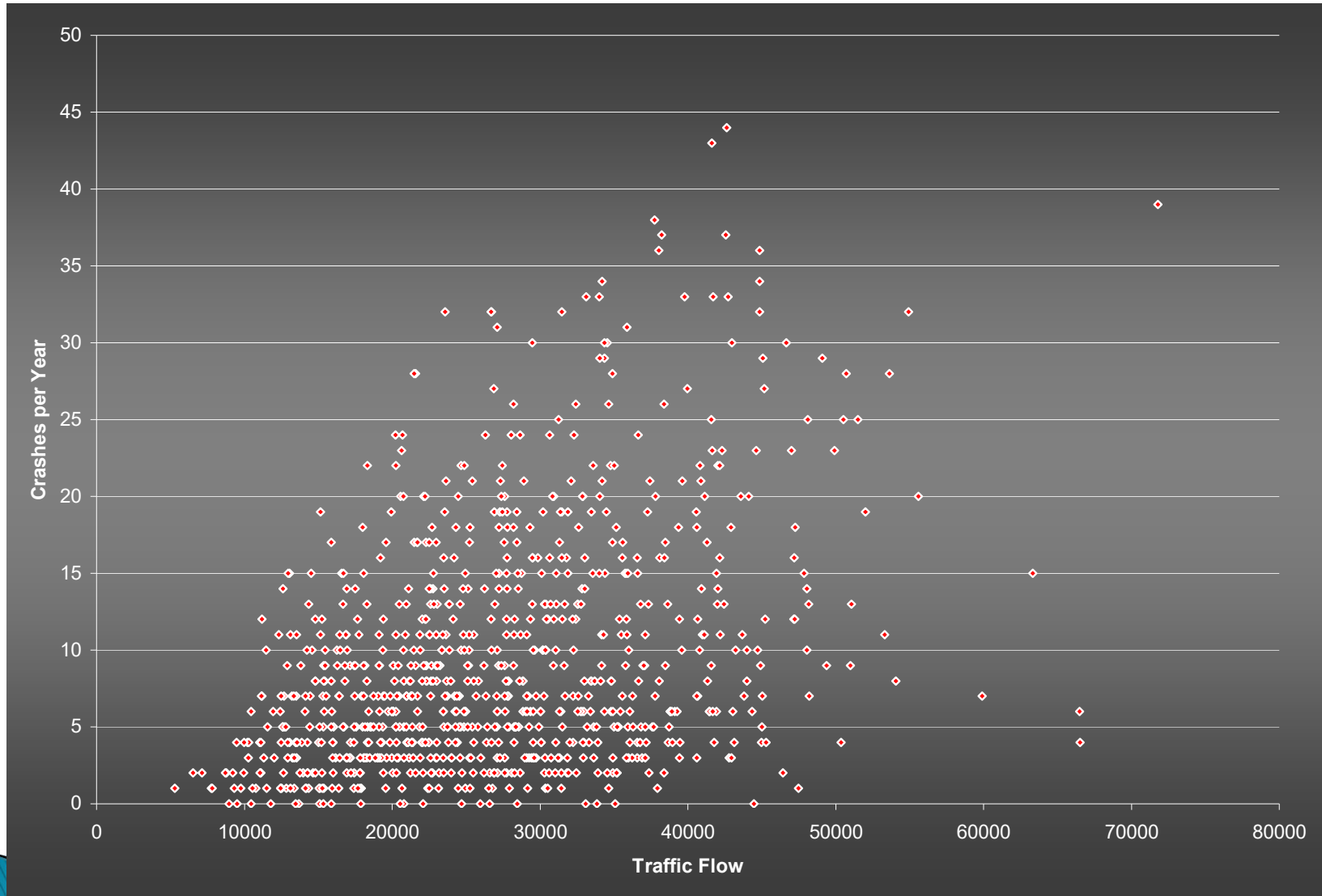


FIGURE 5.15 Scatterplots showing types of correlation.

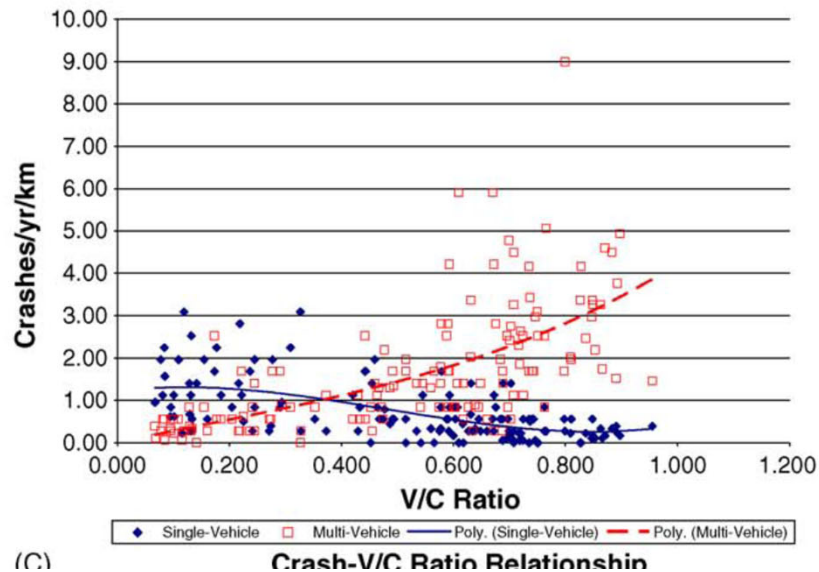
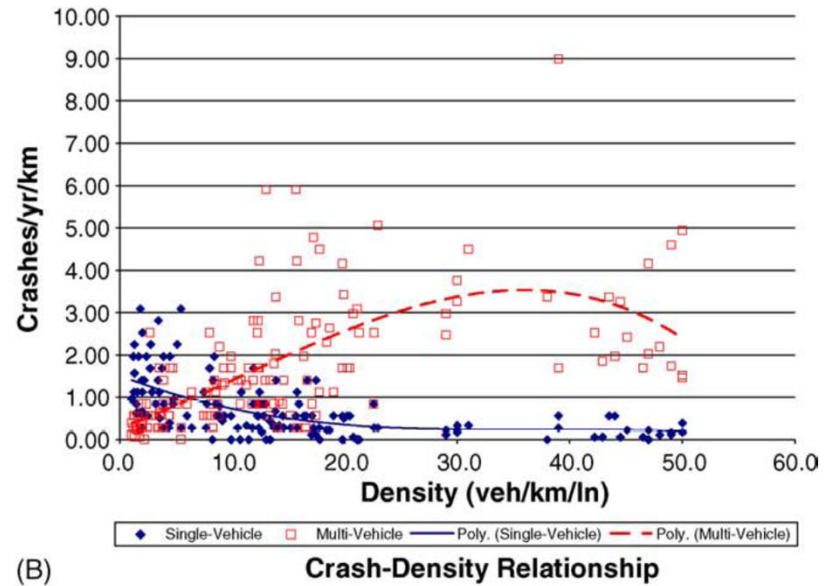
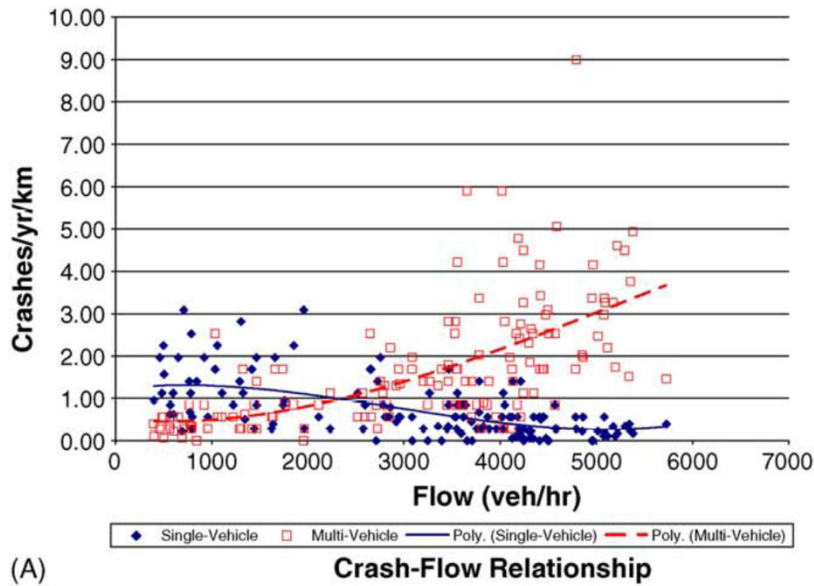
Scatter Plots

Graphical Methods

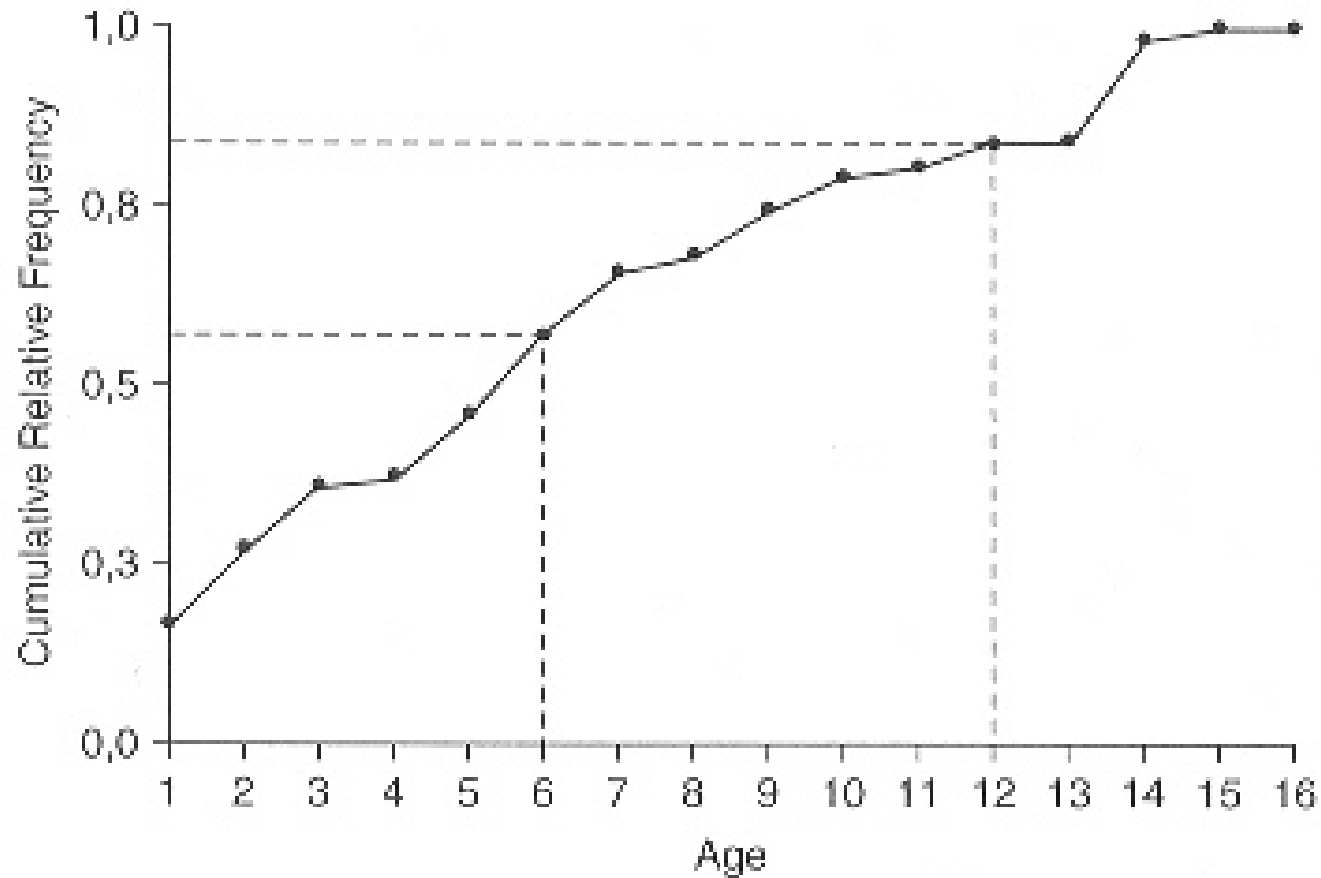


Scatter Plots

Graphical Methods



Graphical Methods



Ogives. Source: Washington et al. (2003)

Graphical Methods

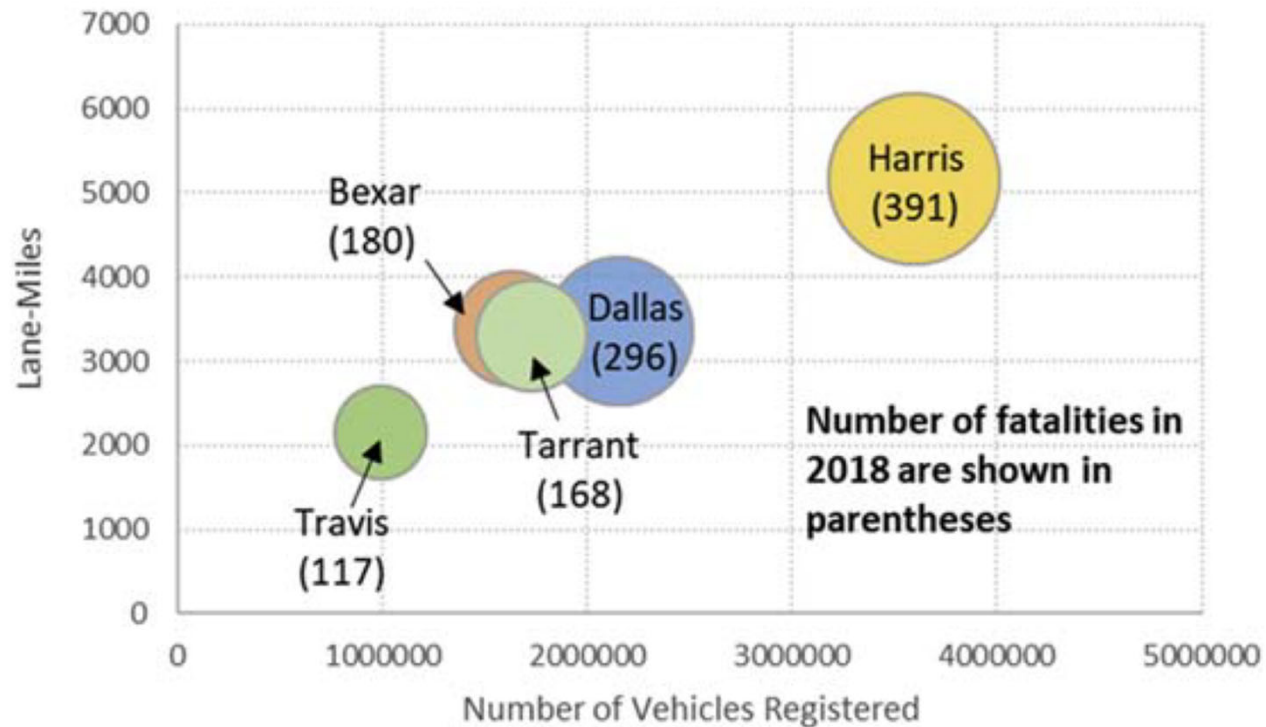


FIGURE 5.17 Bubble chart showing the relationship between fatalities, vehicles registered and lane-miles.

Bubble Chart

Graphical Methods

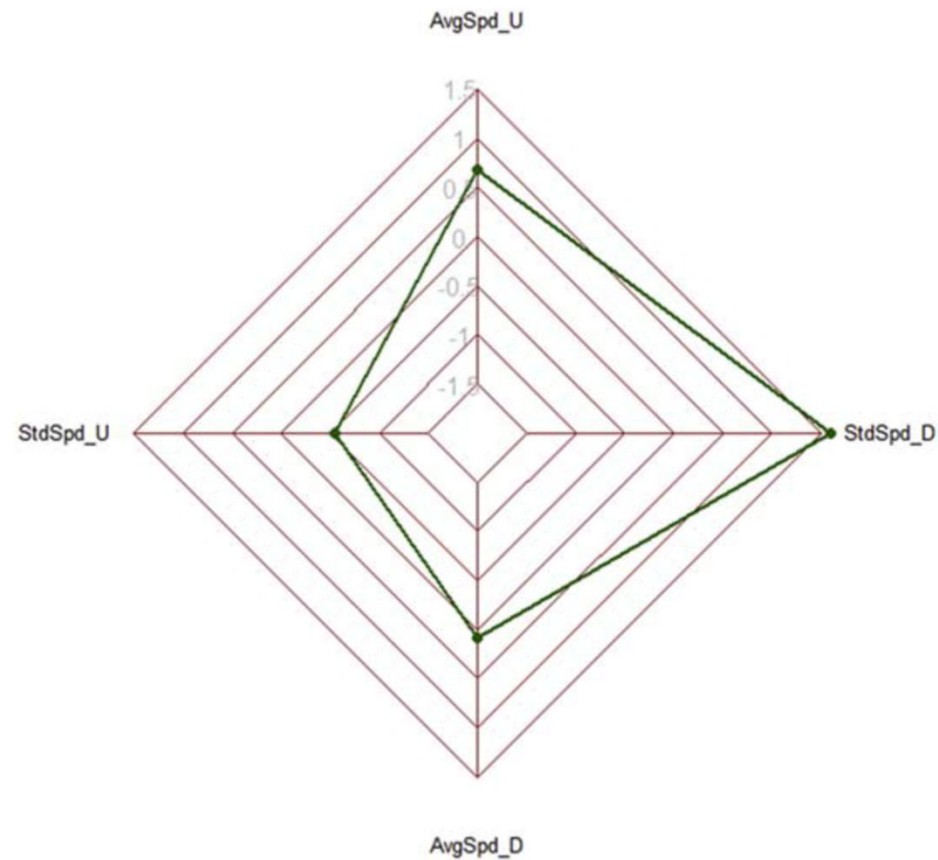


FIGURE 5.18 Relationship between the average speed with low-speed variation at upstream and average speed with high-speed variation at downstream of an urban freeway segment (see Exercise 10.1).

Radar Plots

Graphical Methods

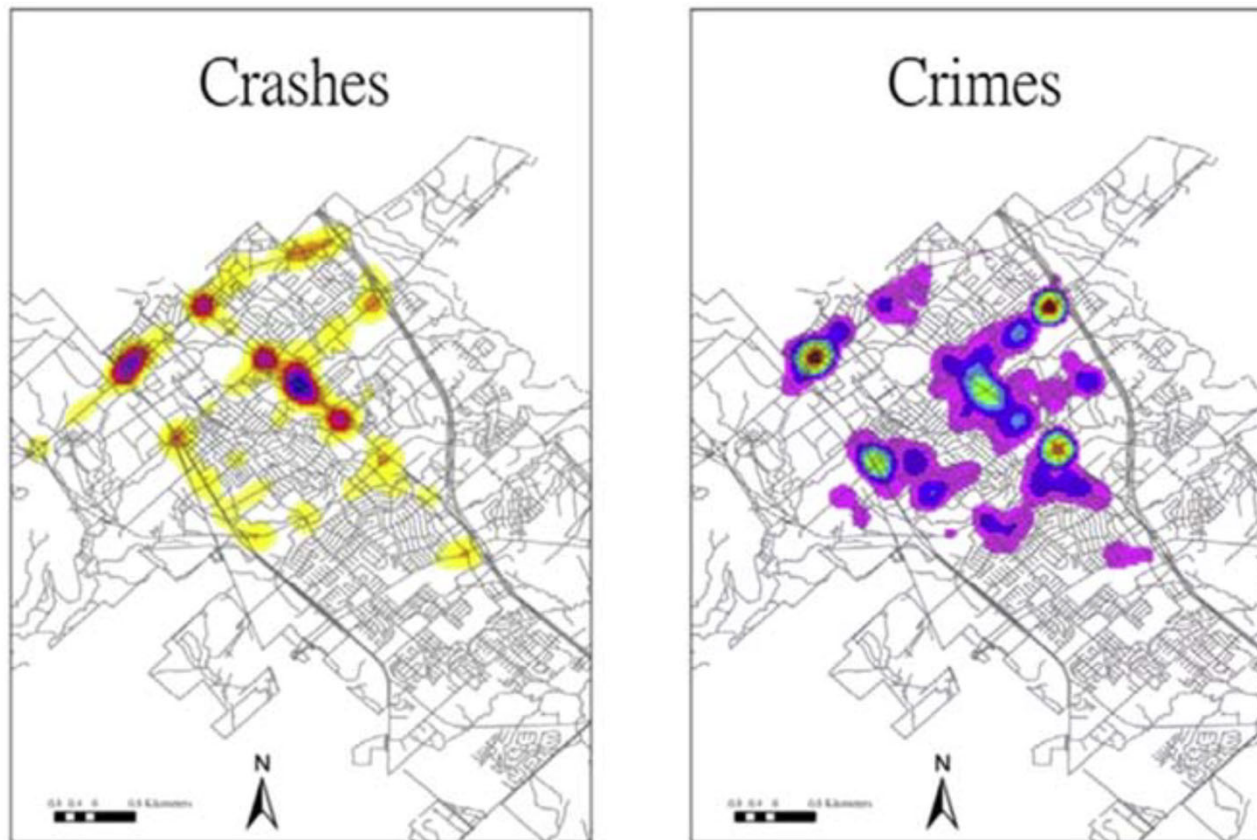
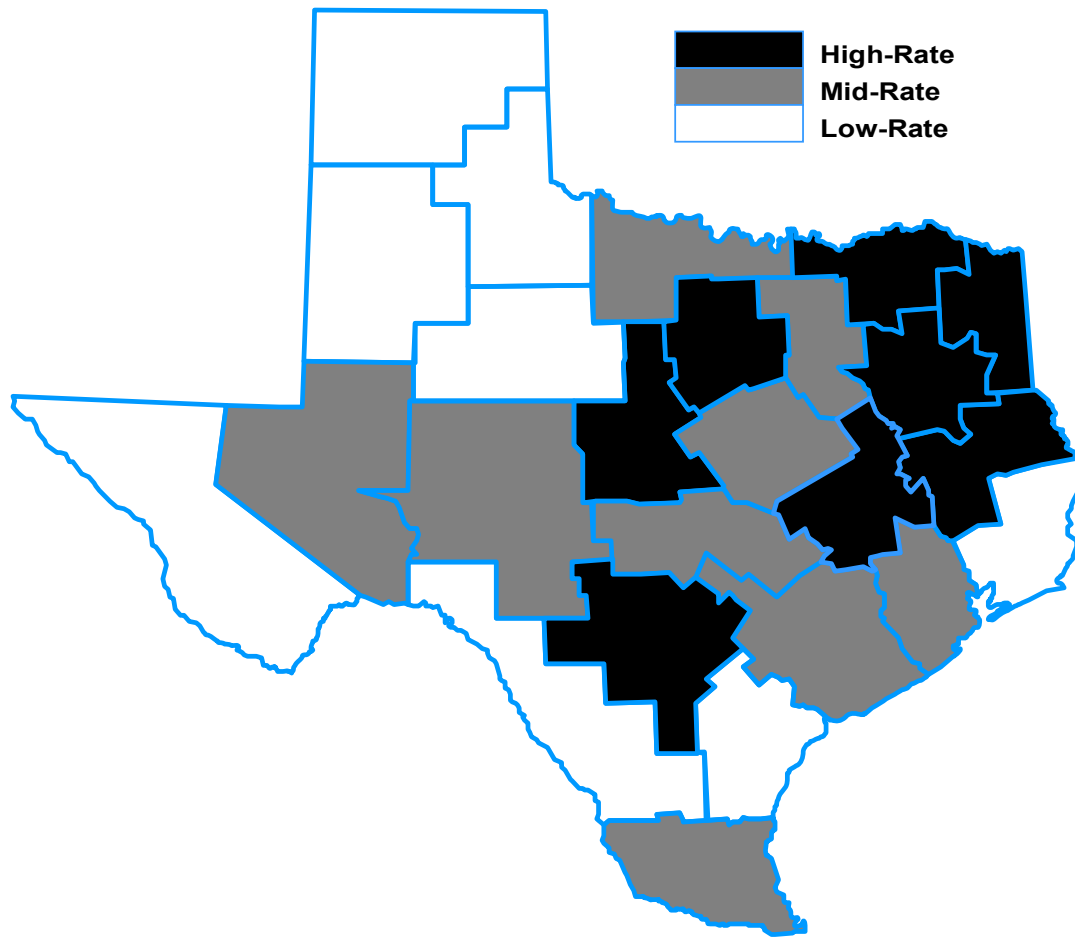


FIGURE 5.19 Heat map of high-risk locations for crashes and crimes. *From Kuo, P.-E., D. Lord, and T.D. Walden (2013) Using geographical information systems to organize police patrol routes effectively by grouping hot spots of crash and crime data. J. Transport Geogr., Vol. 30 (June), pp. 138–148.*

Heat Map

Graphical Methods



Maps



Maps - GIS Information

The screenshot displays the SafeRoadMaps website in a Windows Internet Explorer browser. The page features a navigation menu with options: Home, My Travel, My Community, My State, National Maps, and Analysis & Tools. The main content area is dominated by a map of the United States, overlaid with a color-coded heatmap representing crash density. A legend on the right side of the map indicates 'High Crash Density' in red and 'Low Crash Density' in blue. The map includes a scale bar (500 miles and 500 kilometers) and a 'Take Our Survey' button. Below the map, there are three columns of interactive elements:

- Step 1:** A search form with a text input field for 'Enter an Address, Neighborhood, City, or Zip Code', a 'Zoom to Location' button, a dropdown menu for 'Alabama', and a 'Zoom to State' button. Below this, instructions state: 'Start by either entering location information or use the dropdown menu to find a State.'
- Step 2:** A 'Search Radius' dropdown menu set to '1 Mile'. Below this, instructions state: 'Once you have zoomed into the area you wish to examine, click directly on any Heat Map Cluster on the map (Green or Yellow colored areas) for crash information. From there, you can click on any of the triangular yellow icons which appear on the map to obtain more detailed information relating to that event. If no triangular icons appear after clicking, that means that no accidents were found in that vicinity, so by clicking on another spot on the map. You can also control the Search Radius through the pull-down menu above.'
- Summer Hot Spots:** A section with several links: 'Rural Summer Hot Spots [On / Off]', 'Rural Summer Hot Spots Compiled [On / Off]', 'Urban Summer Hot Spots [On / Off]', 'Urban Summer Hot Spots Compiled [On / Off]', 'Heat Map [On / Off]', and 'Reload Page'. Below these, it states: 'You can also toggle On or Off various overlays for Rural and Urban areas through the links above.'
- All Year Hot Spots:** A link labeled 'Click Here'.
- PDFs:** A link labeled 'Download' with sub-links: 'List of States with the most Rural Summer Hot Spots' and 'List of States with the most Urban Summer Hot Spots'.

The browser's address bar shows the URL: <http://saferoadmaps.org/maps/index.htm#Fragment-5>. The status bar at the bottom indicates 'Internet | Protected Mode: Off' and '100%' zoom level.

<http://www.saferoadmaps.org/home/>

Graphical Methods

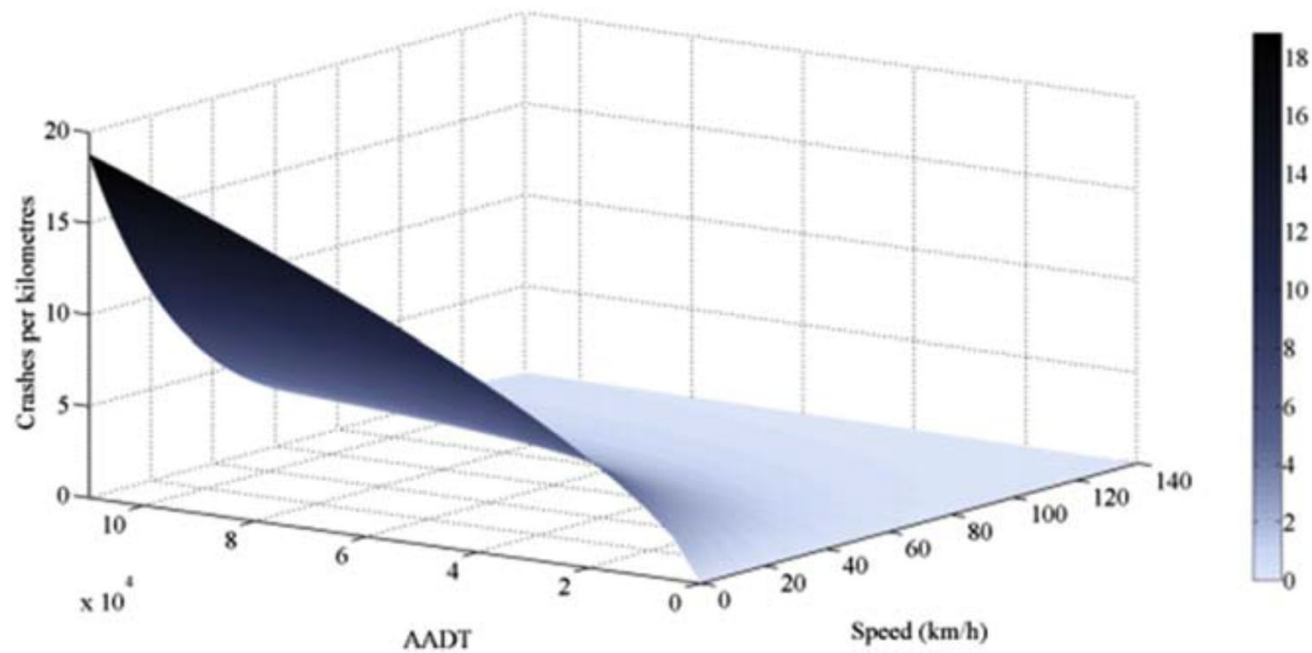
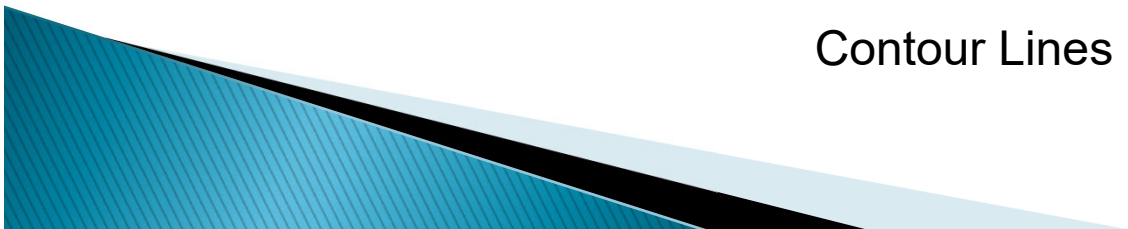


FIGURE 5.20 Contour plot of slight injury crashes. From Imprialou, M-I, M. Quddus, D. Pitfield, D. Lord (2016) *Re-visiting crash-speed relationships: a new perspective in crash modelling*. *Accid. Anal. Prev.*, Vol. 86, pp. 173–185.

Contour Lines



Graphical Methods

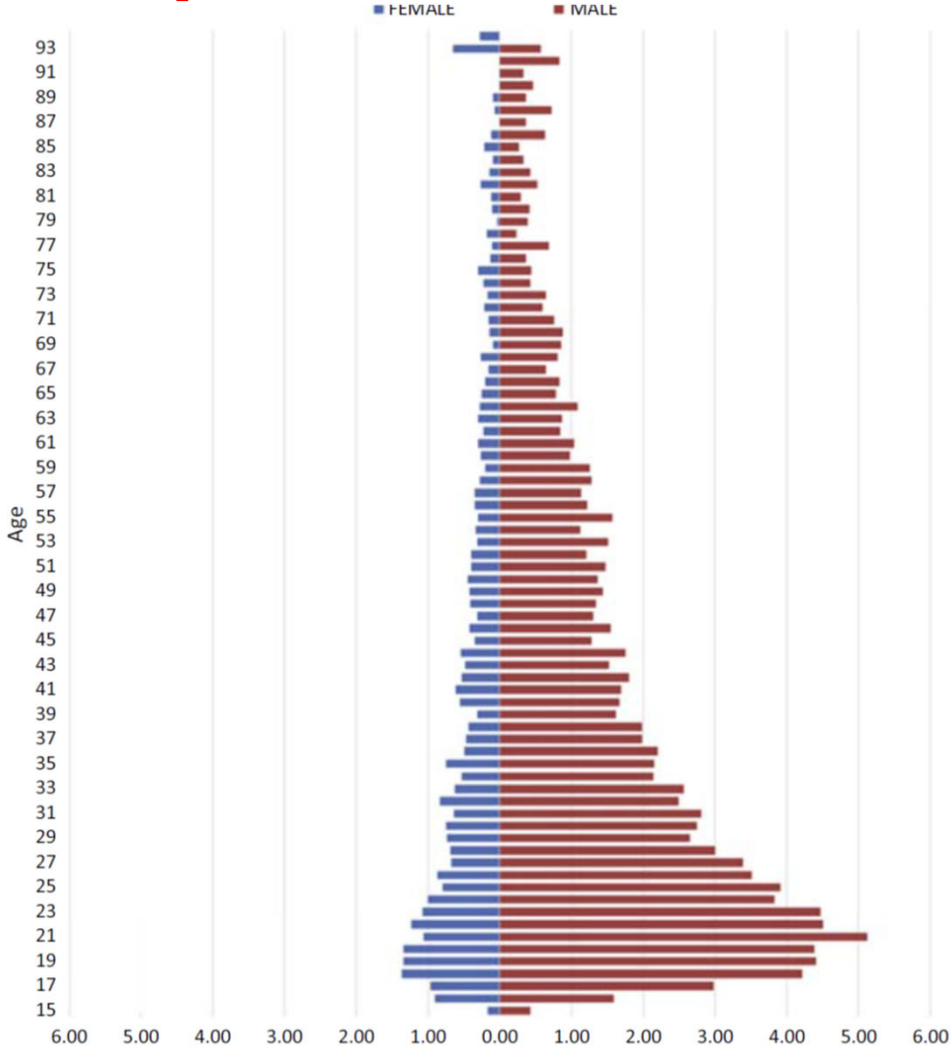


FIGURE 5.21 Population pyramid for fatal and serious injury speeding crashes. This figure is taken from the link: <https://www.texasshsp.com/wp-content/uploads/2019/02/SHSP-2019-v3.pdf>.

Pyramid