Exploratory Analyses of Safety Data

December 1, 2021

Instructor: Srinivas Geedipally
Texas A&M Transportation Institute
Srinivas-g@tti.tamu.edu

Objectives

- Understanding the data and identifying data issues such as errors and missing information,
- Detecting outliers whose values are significantly different from the other observations in the dataset.

The yearly cost of bad data is over $3 trillion annually in the US.

– *Harvard Business review*

Objectives

- Selecting the most important variables
  - Identifying possible relationships in terms of direction and magnitude between independent and outcome variables,
- Testing hypotheses and developing associated confidence intervals or margins of error,
- Examining underlying assumptions to know if the data follows a specific distribution, and
- Choosing a preliminary model that fits the data appropriately.

Techniques

Quantitative

- Calculation of summary statistics

Graphical

- Summarizing through graphs

Methods

Univariate

- One variable at a time

Multivariate

- Two or more variables

Quantitative Techniques

Measures of Central Tendency

- Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample mean

Sample size

- Median
  - A value that divides the dataset into two halves
  - If N is odd number, median = middle value
  - If N is even number, median = mean of middle values
- Mode
  - Observation that has highest number of occurrences in the dataset

Measures of Variability

- Range
  - Captures amount of variability
  - Difference between the largest and smallest observations
- Quartiles
  - Separate the dataset into four equal parts
  - Use percentiles
  - First quartile (Q1) – 25th percentile; second quartile (Q2) or median – 50th percentile; Third quartile (Q3) – 75th percentile
  - 85th percentile speed used for setting up speed limits
- Interquartile range
  - Captures data spread
  - Middle half of the data
  - IQR = Q3 – Q1

Measures of Variability

- Variance
  - Used to calculate dispersion in the data

    Sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

    Population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- Standard deviation
  - Square root of variance
- Standard error

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

- Coefficient of variation

$$CV = \frac{s}{\bar{x}}$$

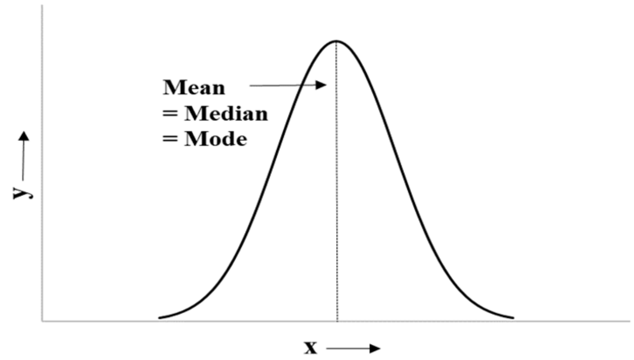
Summary Statistics

Table 1
Summary Statistics of datasets

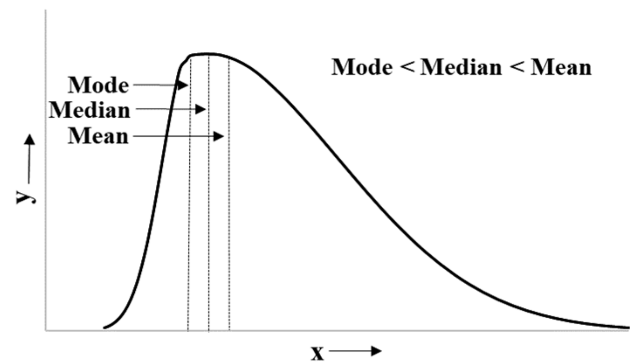
Dataset	Variables	Min	Max	Average	Standard Deviation
Texas	Number of crashes	0	15	0.86	1.65
	Average 5-years AADT (vpd)	43	1166	313.8	253
	Segment length (miles)	0.10	4.41	0.96	0.93
Virginia	Number of crashes	0	8	2.01	2.09
	Average 5-years AADT (vpd)	163	5180	694	625
	Segment length (miles)	0.13	5.67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

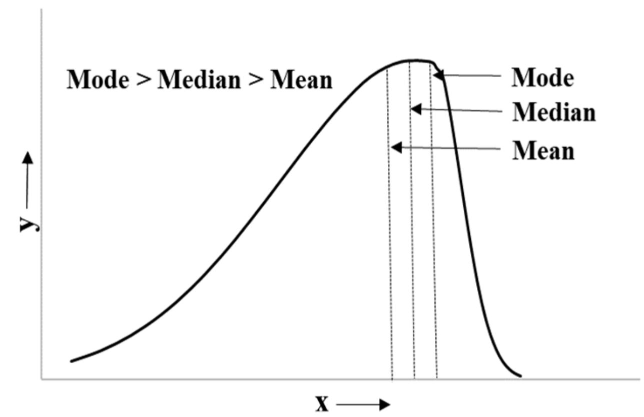
Measures of Variability

- Symmetrical



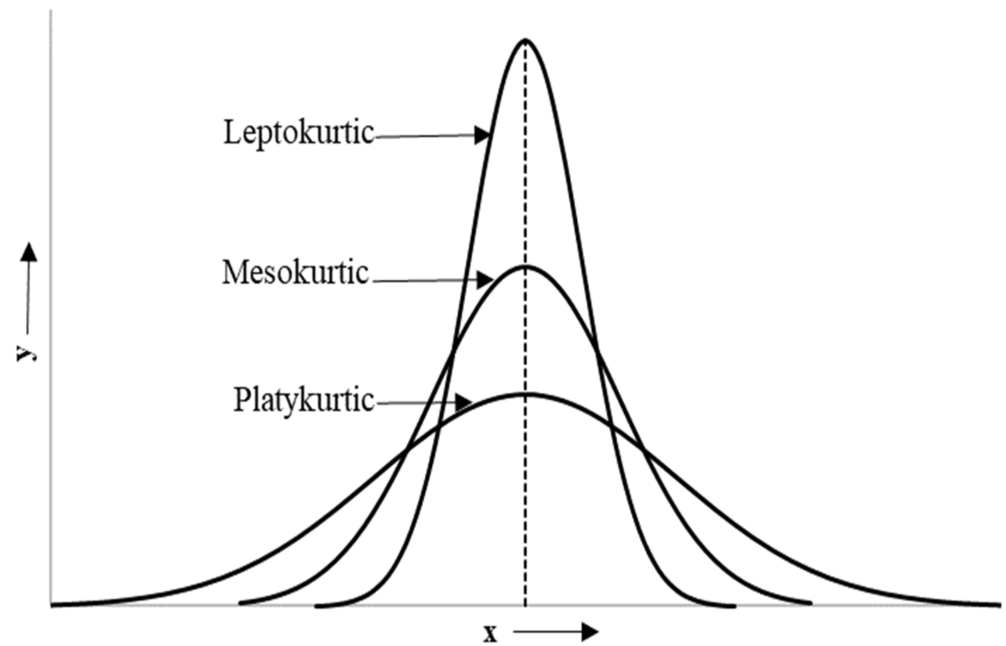
- Asymmetrical data





Measures of Variability

- Skewness, Kurtosis (measure of the sharpness of the peak of a frequency distribution).



Measures of Association

- Pearson's correlation coefficient
  - Used when variables are measured on an interval/ratio (continuous) scale
- Spearman rank-order correlation coefficient
  - Used when variables are measured on an ordinal/ranked (integer) scale
- Chi-square test for independence
  - Used when two sets of data are measured on the categorical scale

Measures of Association

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.*

Measures of Association

- Relative risk and odds ratio

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

$$RR = \frac{A/(A+B)}{C/(C+D)} \qquad OR = \frac{A/B}{C/D} = \frac{AD}{BC}$$

Measures of Association

- Is it dangerous to use cell phone while driving? (Example 5.3)

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

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

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

Confidence Intervals

- Statistics are usually calculated from samples
- 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

Confidence Intervals

Confidence intervals for unknown mean and known standard deviation

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

Confidence intervals for unknown mean and unknown standard deviation

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

Confidence intervals for proportions

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

Confidence Intervals

- Calculate the confidence interval for truck proportion. (Example 5.4)

A survey was conducted at two horizontal curves for a short period of time in Texas. At first horizontal curve, 296 passenger cars and 43 trucks, and at the second curve, 324 passenger cars and 72 trucks were observed.

The sample truck proportion in the traffic is
$(43 + 72)/(296 + 43 + 324 + 72) = 0.156.$

The z-value for the 95% level $\left(\text{significance level} = \frac{1-C}{2} = \frac{1-0.95}{2} = 0.025\right)$= 1.96.

The confidence interval for the truck proportion is obtained as

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

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 not 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.
- As the P-value becomes smaller, the evidence against the null hypothesis becomes stronger.

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 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 b. The probability of not committing a Type II error is called the Power of the test, and is denoted by 1- b.

Hypothesis Testing

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

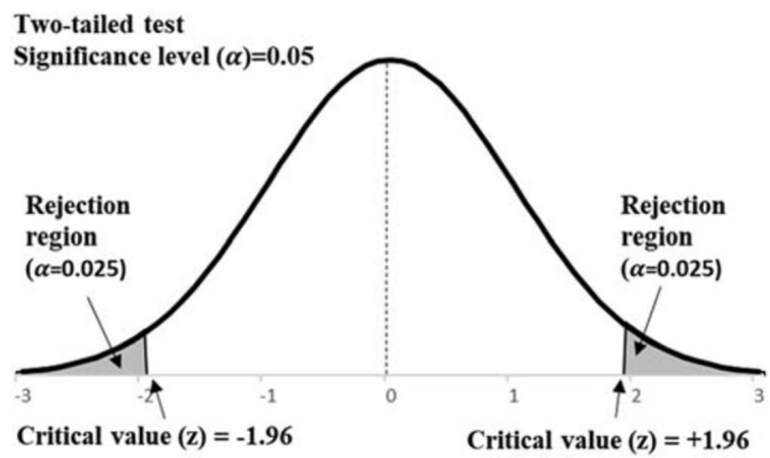


FIGURE 5.3 Critical values for a two-tailed (nondirectional) test.

One-tailed hypothesis test
A one-tailed test is a statistical test in which the rejection region is onesided of the sampling distribution.

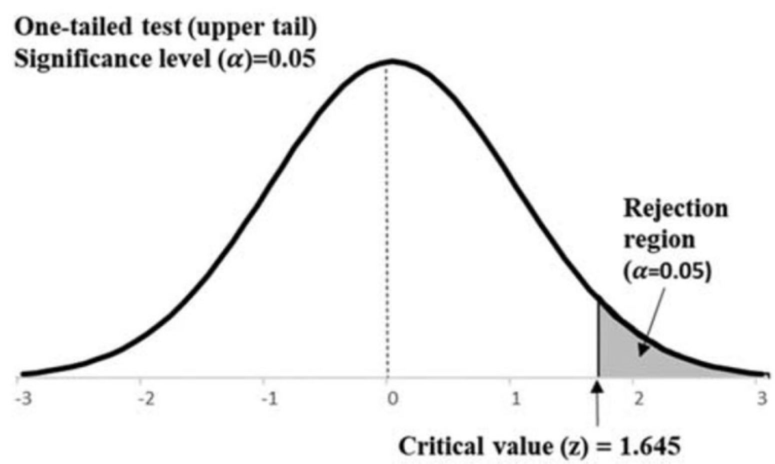


FIGURE 5.4 Critical values for a one-tailed (directional) test.

Hypothesis Testing

- Does drinking two beers cause driver impairment? (Example 5.6)

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
4.85	5.91	−1.06
. .	. .	. .
4.69	3.72	0.97
Mean difference ($\bar{d}$)		−0.5015
Standard deviation (s)		0.8686

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

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.

Hypothesis Testing

Hypothesis testing for one sample

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

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

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

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

Hypothesis Testing

Hypothesis testing for one sample

You compare the value with the one documented in the below table.

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

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.

Hypothesis Testing

Hypothesis testing for two samples

Dependent Samples

- Observations from one group are paired with observations in the other group, so it is called matched pairs.
- 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 previously with $\mu=0$:

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

Hypothesis Testing

Hypothesis testing for two samples

Independent Samples

- Observations selected from one group are completely independent from the observations selected in the second group.
- 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.

Graphical Techniques

Box Plot and Whiskers

- Box Plot displays the five-number summary of a set of data
- Whiskers show variability.

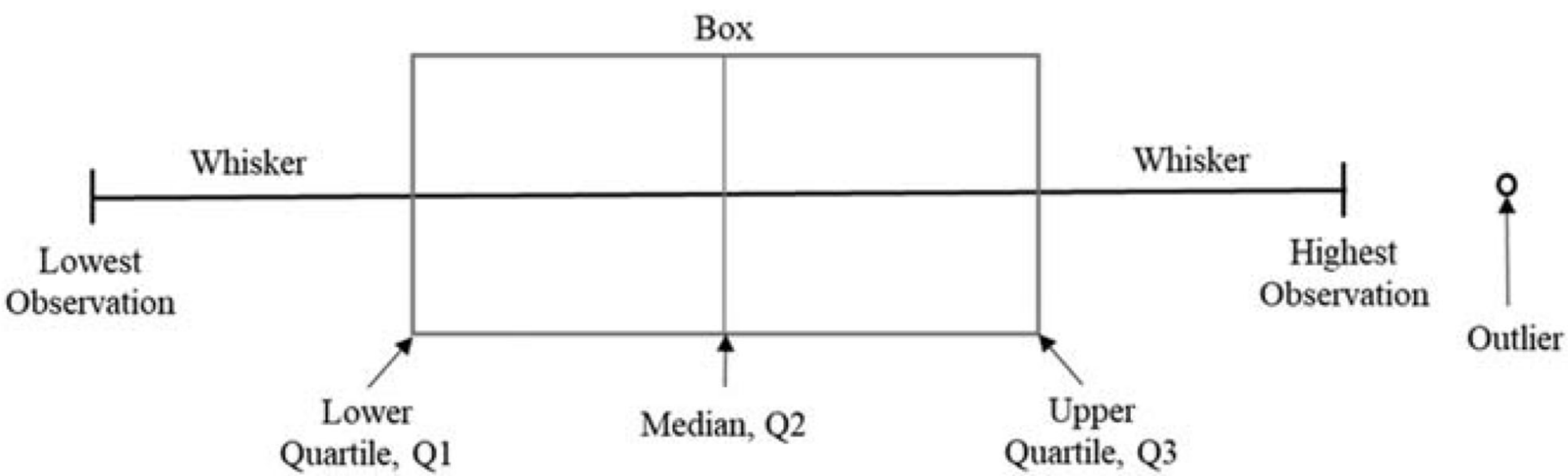


FIGURE 5.5 Box plot showing different measures.

Box Plot and Whiskers

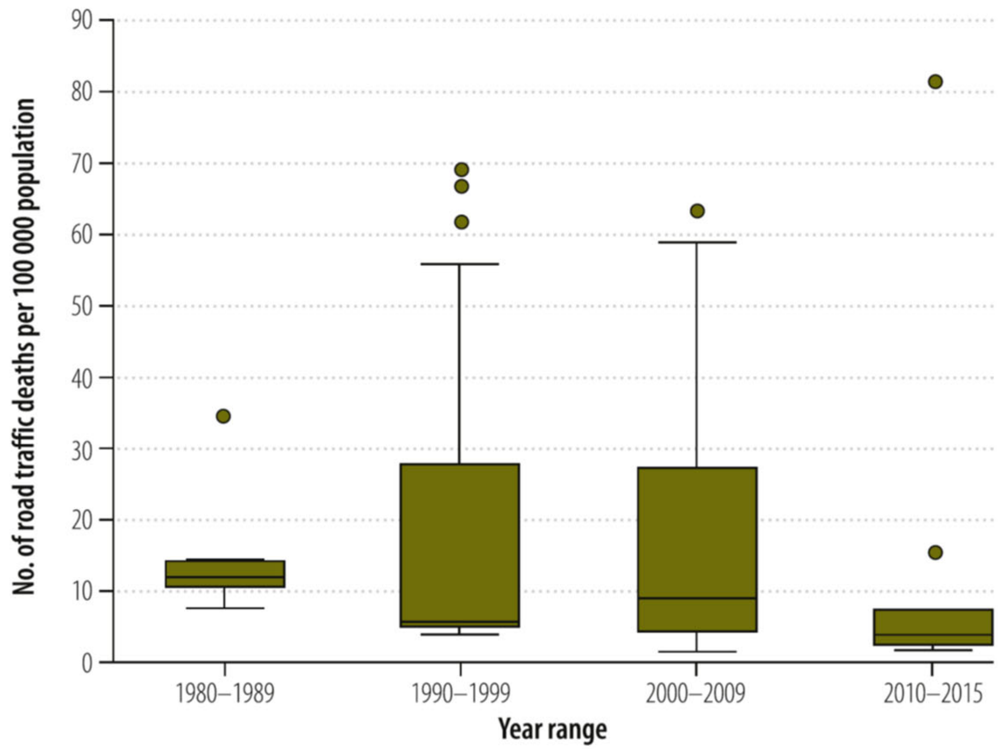


FIGURE 5.6 Box plot showing the traffic death rate in Africa. *From Adeloye D, Thompson JY, Akanbi MA, Azuh D, Samuel V, Omoregbe N, et al. The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and metaanalysis. Bull World Health Organ. 2016;94(7):510–21A.*

Histograms

- Shows the distribution of a continuous variable
- Bins must be adjacent without any gap

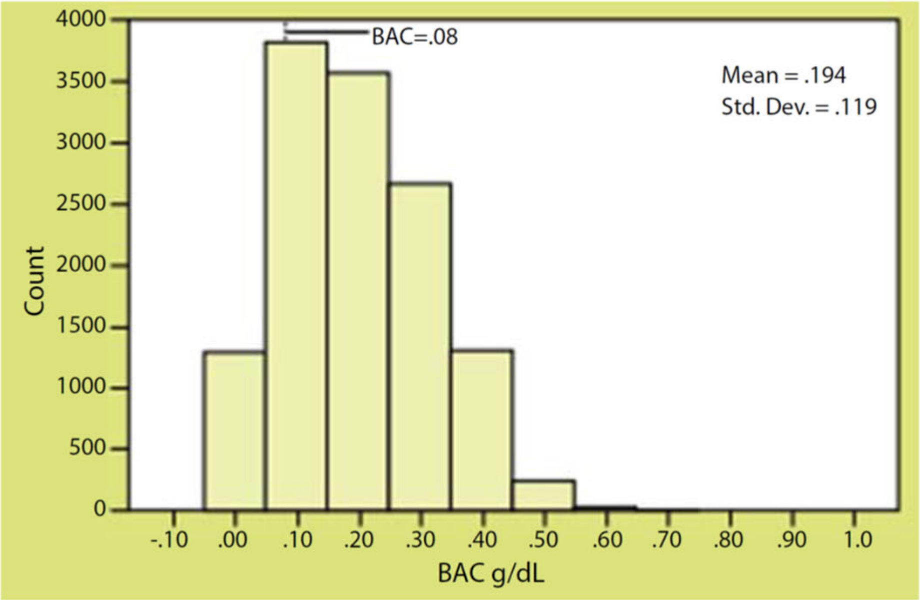


FIGURE 5.7 Histogram of passenger car driver BAC values. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.

Histograms

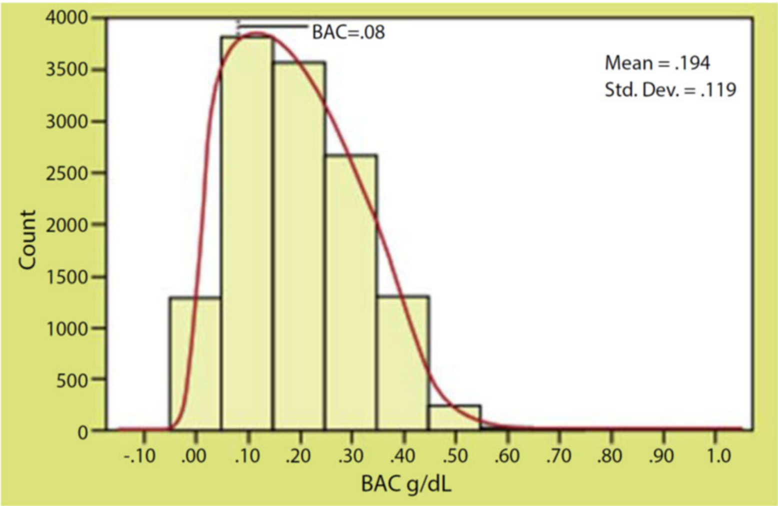


FIGURE 5.8 Histogram of passenger car driver BAC values with kernel density superimposed. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.

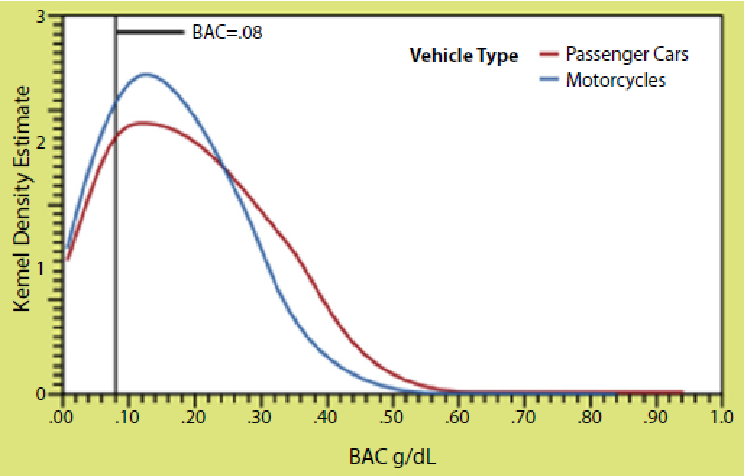


FIGURE 5.9 Kernel density plots: passenger car driver and motorcycle operator BAC values. *From National Highway Traffic Safety Administration, 2007. Traffic Safety Facts. Differences in Driver Alcohol Involvement by Age Group and Vehicle Type.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810754.

Bar Graphs

- Bar graph relates two variables
- It is used for categorical variables

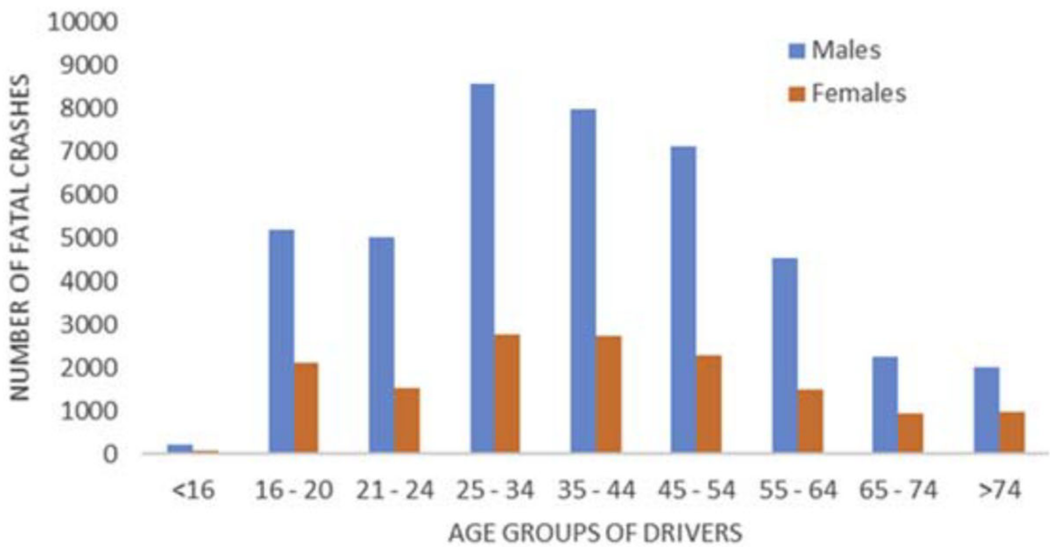
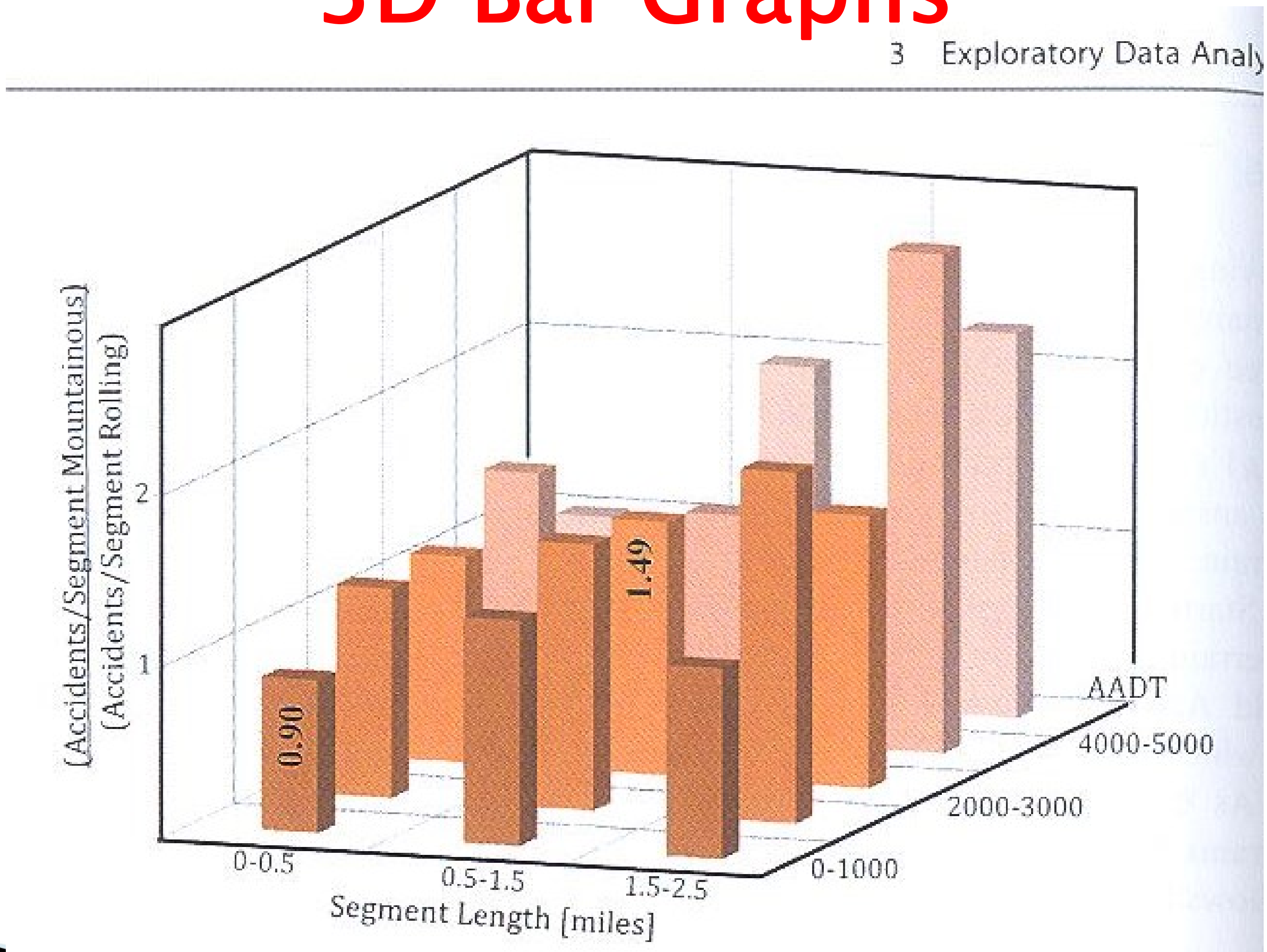


FIGURE 5.10 Male versus female drivers fatal crashes. *Based on data available at StatCrunch: https://www.statcrunch.com/5.0/viewreport.php?reportid=35500.*

3D Bar Graphs



Stacked Bar Graphs

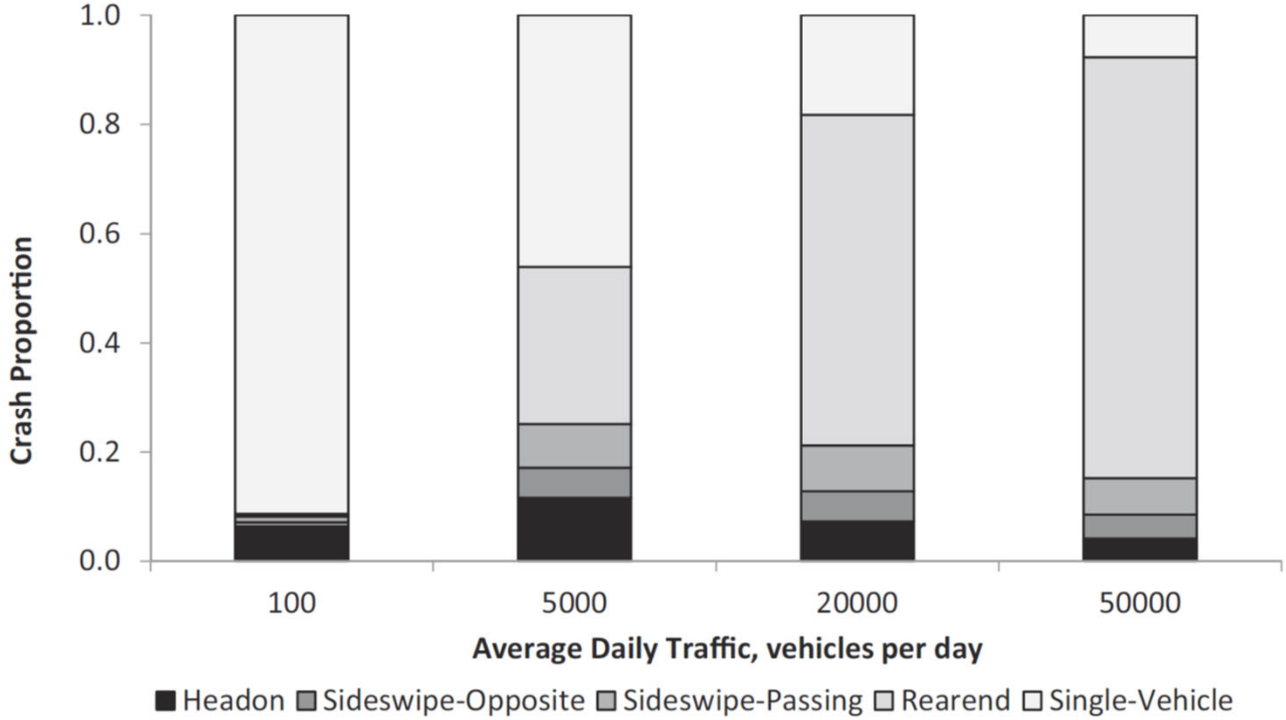


FIGURE 5.11 Crash proportion by collision type.

Mosaic Plot

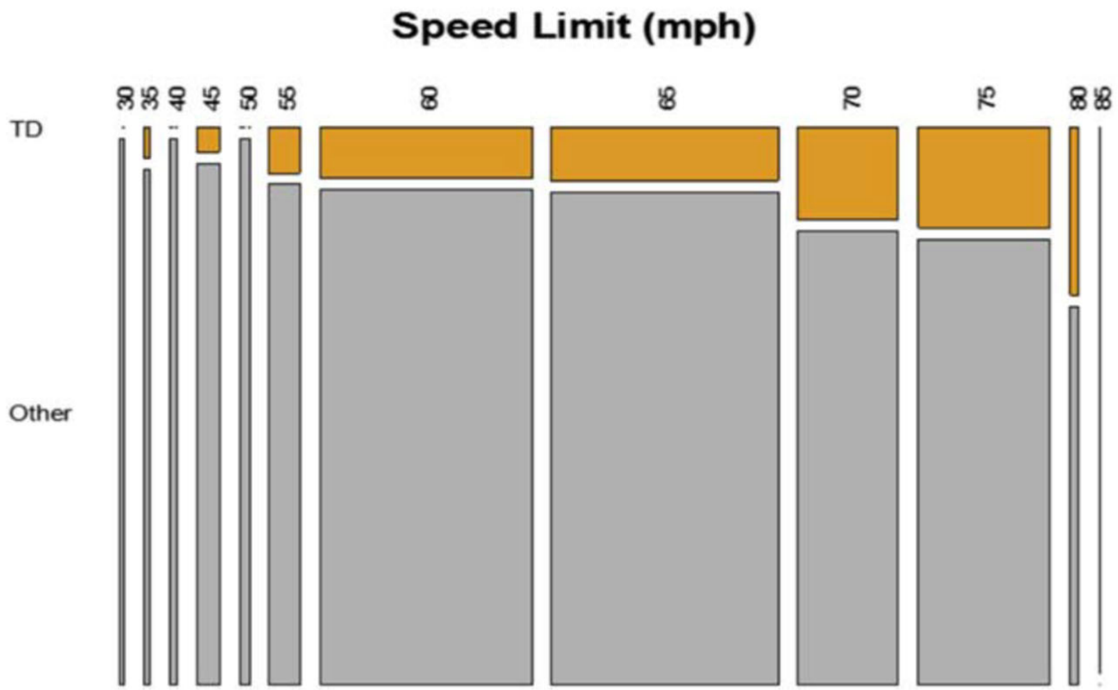


FIGURE 5.12 Crashes caused by tire debris by speed limit in Texas. *From Avelar, R.E., M.P. Pratt, J.D. Miles, T. Lindheimer, N. Trout, and J. Crawford (2017) report Develop Metrics of Tire Debris on Texas Highways: Technical Report. FHWA/TX-16/0-6860-1. Texas A&M Transportation Institute, College Station,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

Error Bars

- Indicates the uncertainty in the estimated measurement
- Used to show a confidence interval or the minimum and maximum values

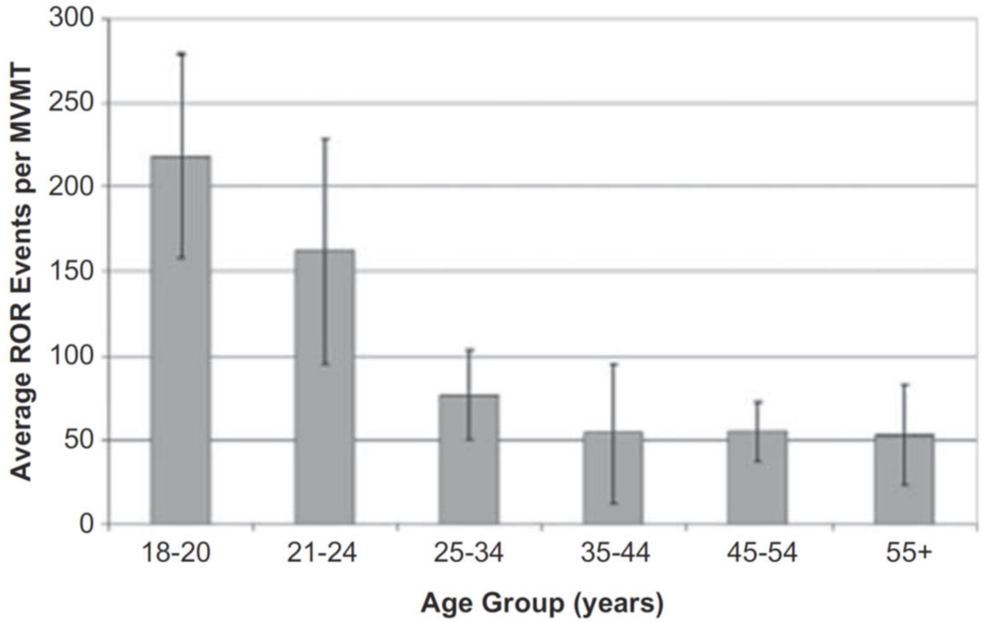


FIGURE 5.13 Average ROR events per MVMT by age group. *From McLaughlin, S.B., Hankey, J.M., Klauer, S.G., Dingus, T.A., 2009. report Contributing Factors to Run-Off-Road Crashes and Near-Crashes, National Highway Traffic Safety Administration, Report DOT HS 811 079.*

Pie Charts

- Used to show proportions of various categories

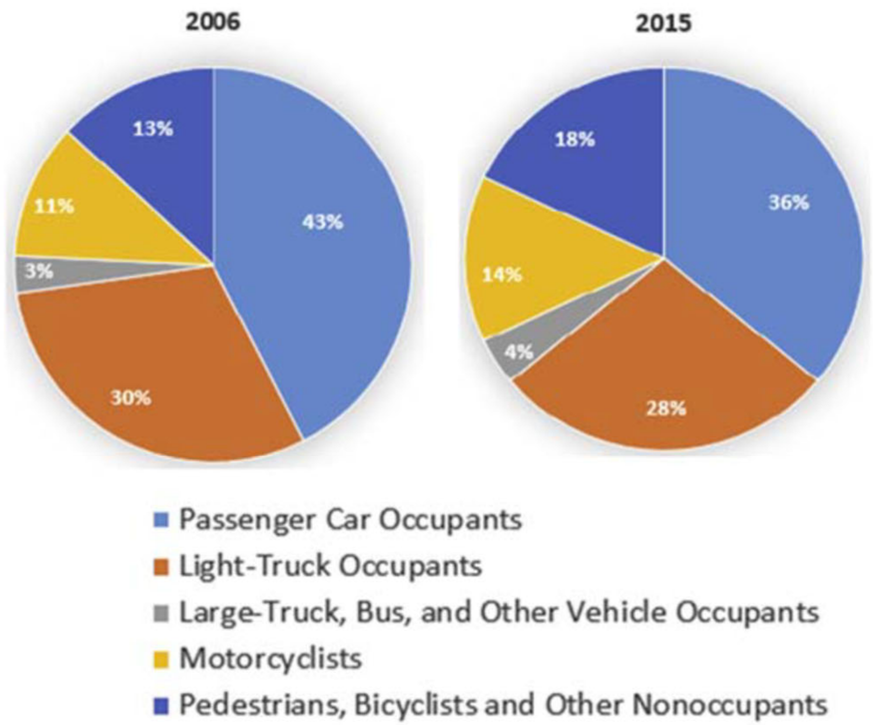


FIGURE 5.14 Fatality composition by vehicle type, 2006 and 2015. *Based on data available at: National Highway Traffic Safety Administration, report National Center for Statistics and Analysis. (2016, August). 2015 motor vehicle crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS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

Scatterplots

- Used to show the relationship between two variables

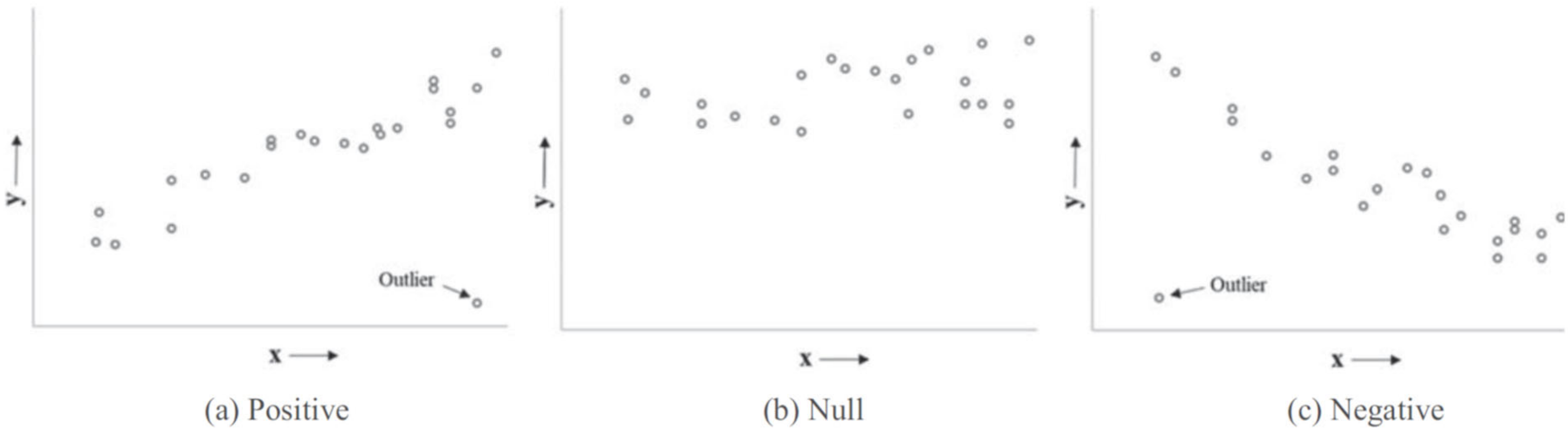
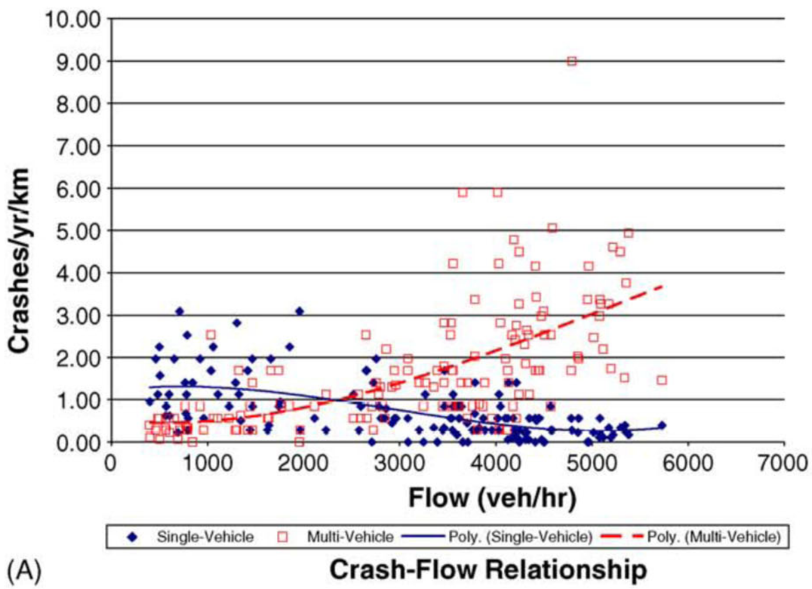
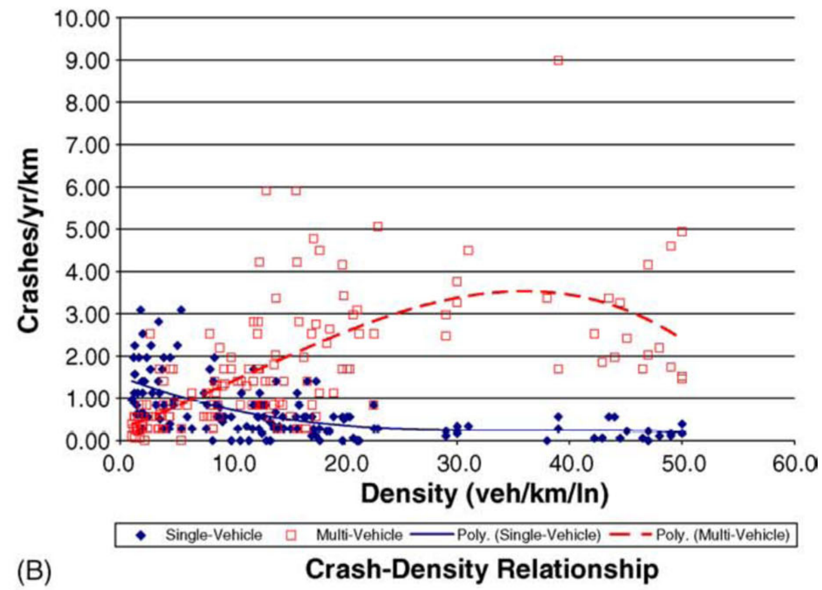
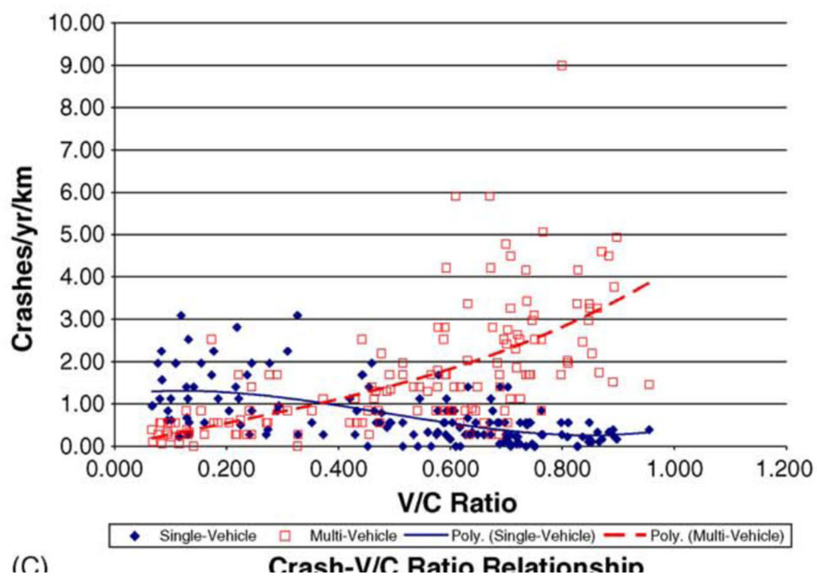


FIGURE 5.15 Scatterplots showing types of correlation.

Scatterplots







Bubble Charts

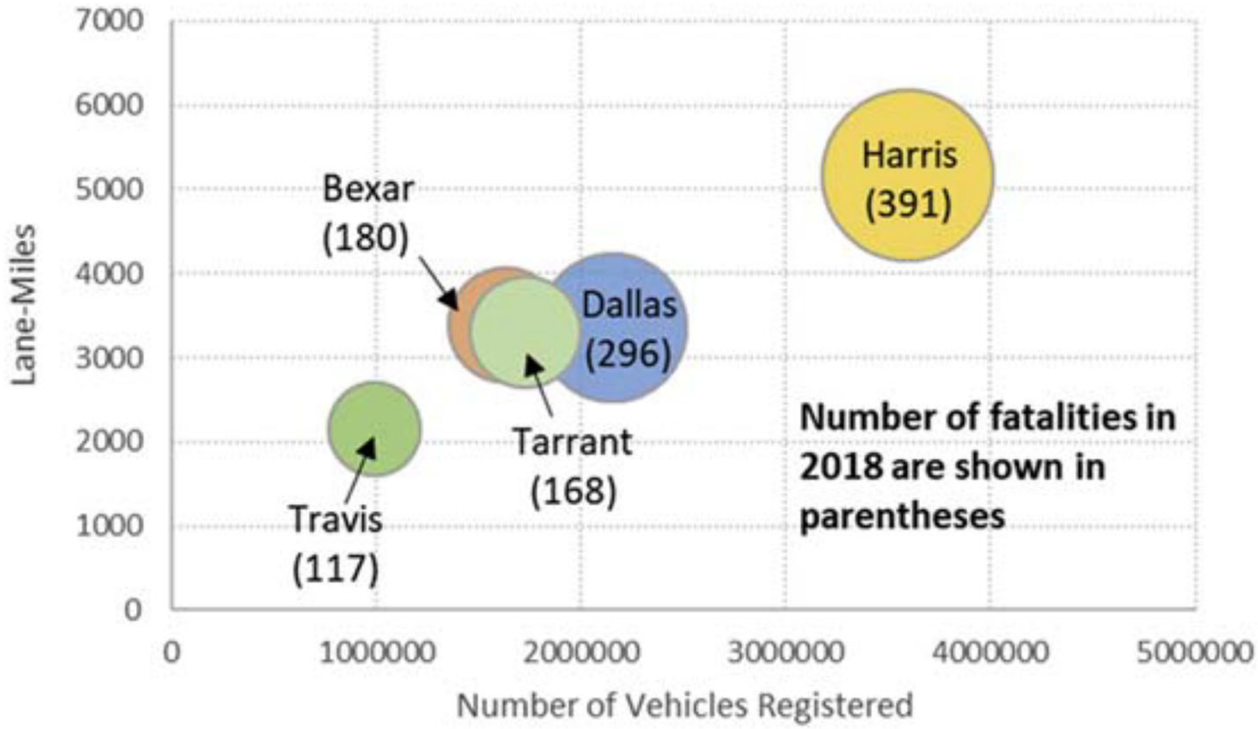


FIGURE 5.17 Bubble chart showing the relationship between fatalities, vehicles registered and lane-miles.

Radar/Web Plots

- A two-dimensional figure used for examining several variables at the same time or on the same plane for a single unit

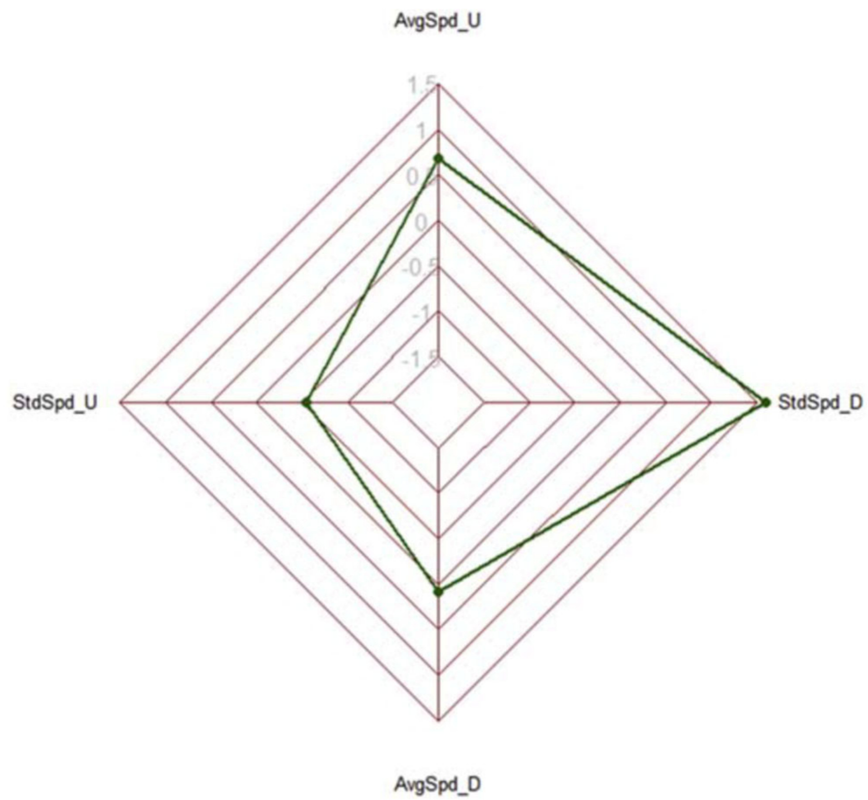


FIGURE 5.18 Relationship between the average speed with low-speed variation at upstream and average speed with high-speed variation at downstream of an urban freeway segment (see Exercise 10.1).

Heat Map

- Used to communicate relationships between data values and to explore large datasets

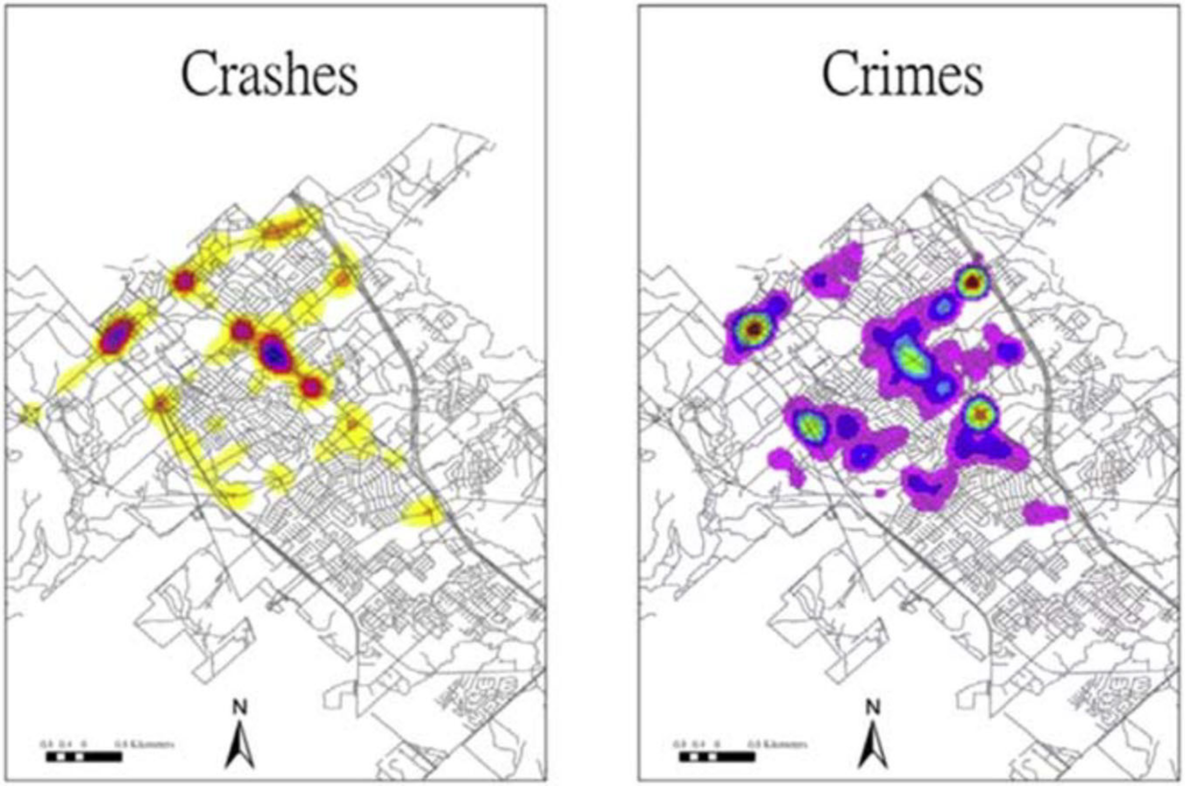


FIGURE 5.19 Heat map of high-risk locations for crashes and crimes. *From Kuo, P.-F., D. Lord, and T.D. Walden (2013) Using geographical information systems to organize police patrol routes effectively by grouping hot spots of crash and crime data. J. Transport Geogr., Vol. 30 (June), pp. 138–148.*

Contour Plots

- Uses constant z- slices, called contours, on a two-dimensional plane to show a three-dimensional surface

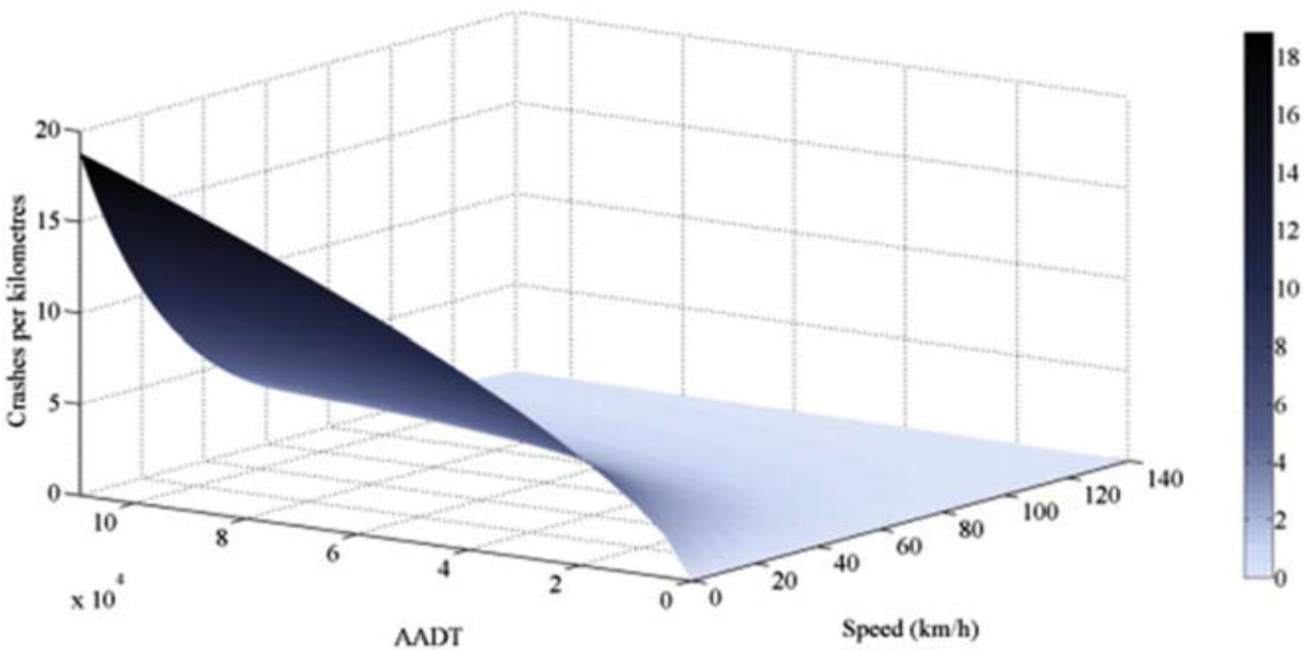


FIGURE 5.20 Contour plot of slight injury crashes. *From Imprialou, M-I, M. Quddus, D. Pitfield, D. Lord (2016) Re-visiting crash-speed relationships: a new perspective in crash modelling. Accid. Anal. Prev., Vol. 86, pp. 173–185.*

Population Pyramid

- Illustrates graphically the distribution of various age groups in a population by gender for a particular variable

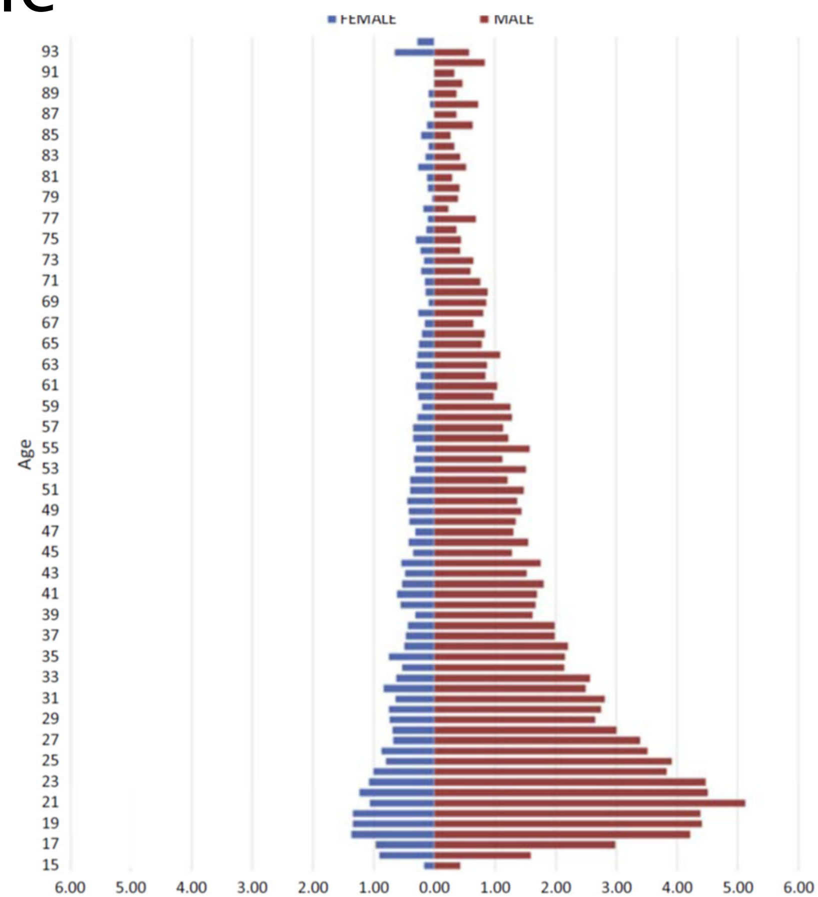


FIGURE 5.21 Population pyramid for fatal and serious injury speeding crashes. *This figure is taken from the link: https://www.texasshsp.com/wp-content/uploads/2019/02/SHSP-2019-v3.pdf.*