第一章 置信区间与假设检验

I. Descriptive Statistics (Methods and techniques for summarizing and interpreting data)

- **Numerical methods**: by estimating the value of an unknown population parameter using a single value (or point), one can make inference about a population.
- **Graphical methods**: graphical representation of data
- Relative to graphical methods, numerical methods provide precise and objectively determined values that can easily be manipulated, interpreted, and compared.
- **Measures of Relative Standing**

  - Percentile Value: 90th percentile: 90% of the observations have a lower magnitude

  - Quartiles: the percentage points that separate the data into quarters

- **Measures of Relative Standing**

  - First (lower) quarter, below which lies one quarter of the data, make it the 25th percentile

  - Second (middle/median) quarter, below which lies half of the data, make it the 50th percentile

  - Third (upper) quarter: 75th percentile

  - Interquartile range: the difference between the first and third quartiles (spread of the data)

- **Measures of Central Tendency**

  - **Median**: lies in the center of the data (50th percentile)

  - **Mean (arithmetic mean)**

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

Population Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Mean

- **Measures of Central Tendency**

  - **Mode**: the value that occurs most frequently in a set of observations. (can have multiple modes)

- **Measures of Variability**

  - Variability is a statistical term used to describe and quantify the spread or dispersion of data around the center.

  - **Interquatile range**

  - **Range**: the difference between the largest and the smallest observations in the data

  - **Variance**

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

Population Variance

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

Sample Variance

- **Measures of Variability**

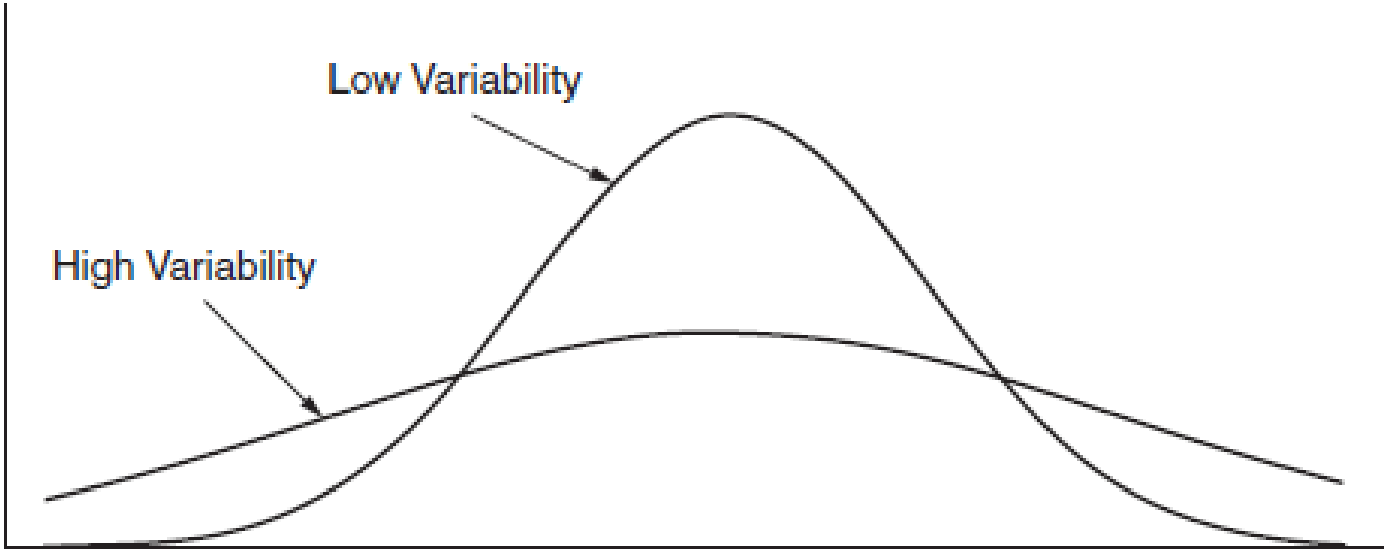
  - Standard Deviation

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

Population Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

Sample Standard Deviation



- Chebyshev Theorem: at least $(1-1/k^2)$ of all observations in a sample of population will lie within k standard deviations of the mean, where k is not necessarily an integer.

- For the approximately bell-shaped normal distribution of observations:

(-s, +s) contains approximately 68% of the measurements

(-2s, +2s) contains approximately 95% of the measurements

(-3s, +3s) contains approximately 99% of the measurements

- **Measures of Variability**

- The standard deviation is an absolute measure of dispersion which does not take into consideration the magnitude of the values in the population or sample.

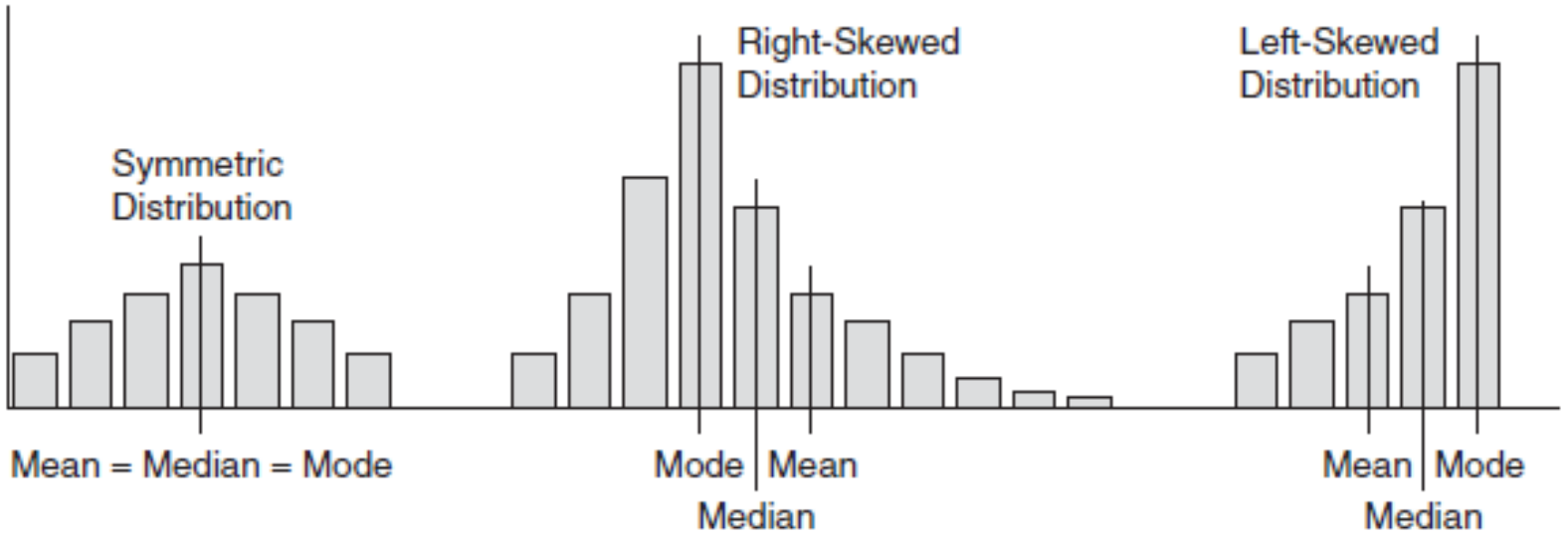
- The coefficient of variation accounts for the magnitudes of the observations. The coefficient of variation (CV) is given as:

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

- **Skewness**

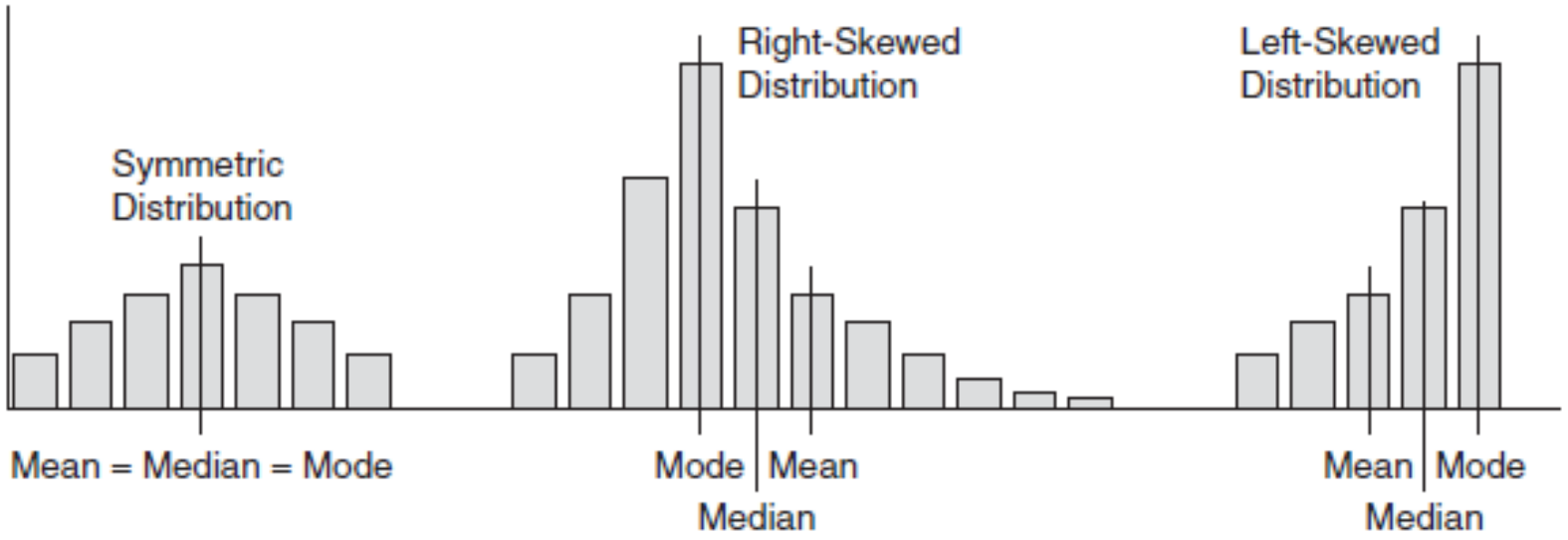
  - Skewness is a measure of the degree of asymmetry of a frequency distribution. It is given as the average value over the entire population (it is also called the third central moment)

  - When the distribution stretches to the right more than it does to the left, it can be said that the distribution is right-skewed, or positively skewed. When the distribution stretches to the left more than it does to the right, it can be said that the distribution is left-skewed, or negatively skewed.



$$g_1 = \frac{m_3}{(m_2\sqrt{m_2})}, \qquad m_3 = \sum_{i=1}^{n}\left(x_i - \overline{X}\right)^3 / n \qquad m_2 = \sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 / n$$

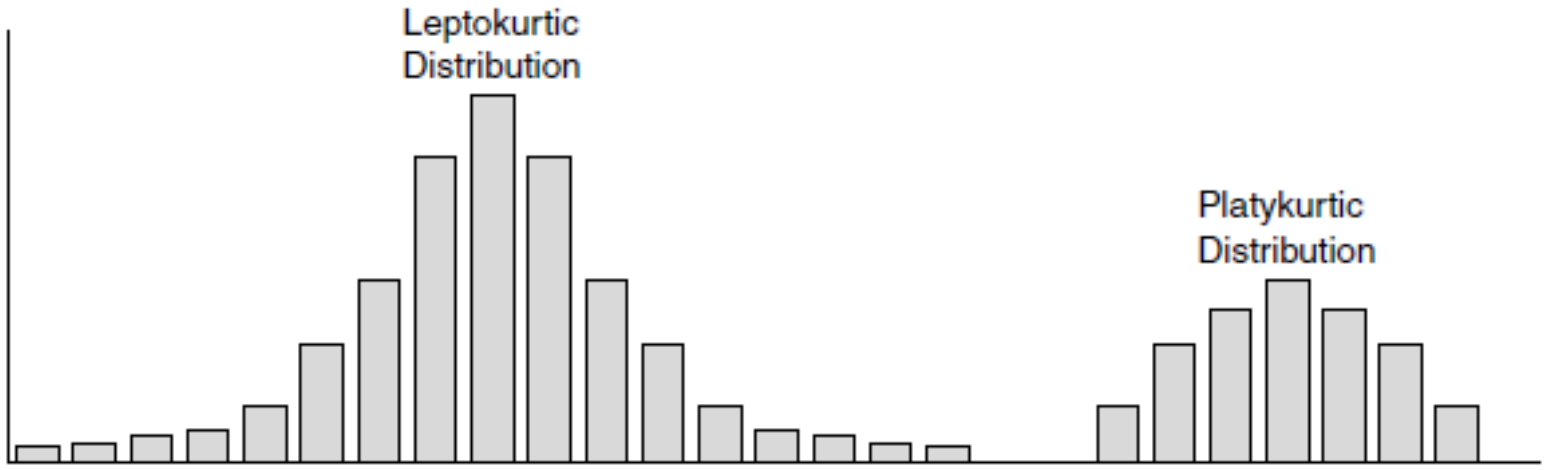
g=0 g>0 g<0



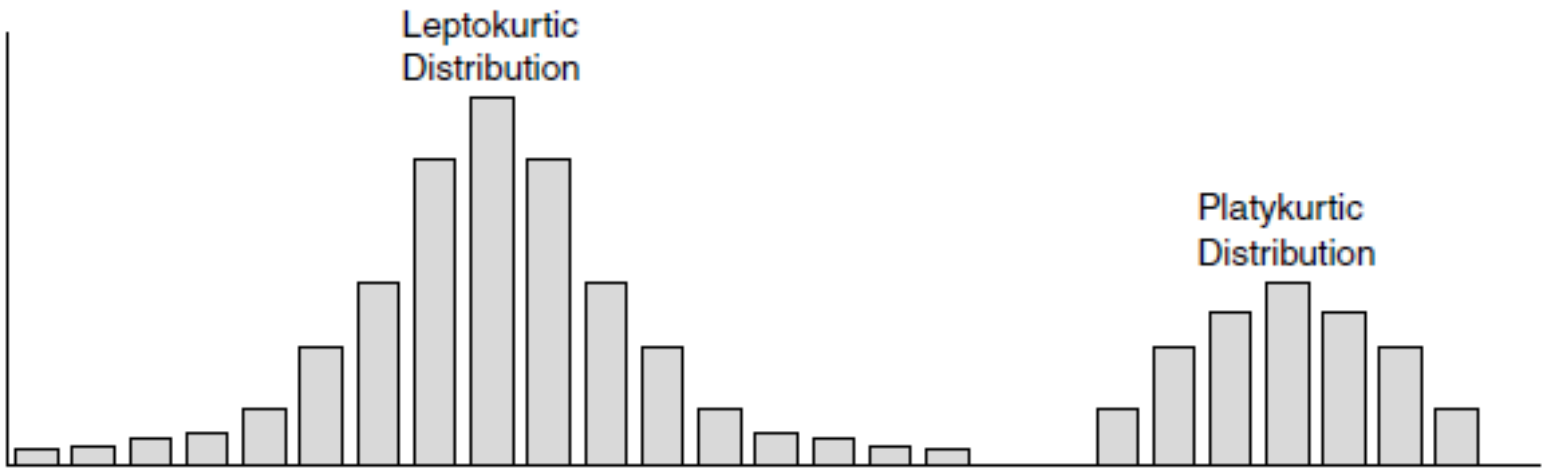
- **Kurtosis**

  - Kurtosis is a measure of the “flatness” (vs. peakedness) of a frequency distribution.

  - Kurtosis is often called the fourth moment around the mean or fourth central moment.



$$g_2 = \gamma_2 - 3 = \left(m_4 / m_2^2\right) - 3\,, \quad m_4 = \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^4 / n$$



Revisiting the speed data from Example 1.1, there is interest in determining the shape of the distributions for speeds on rural and urban Indiana roads. Results indicate that when all roads are examined together their skewness parameter is –0.05, whereas for rural roads the parameter has the value of 0.056 and for urban roads the value of –0.37. It appears that, at least on rural roads, whereas for urban roads the distribution is

Although the skewness parameter is similar for the two types of roads, the kurtosis parameter varies more widely. For rural roads the parameter has a value of 2.51, indicating a distribution close to normal, whereas for rural urban roads the parameter has a value of 0.26, indicating a relatively

Descriptive Statistics for Speeds on Rural vs. Urban Indiana Roads

Statistic	Rural Roads	Urban Roads
N (number of observations)	888	408
Mean	58.79	59.0
Std. deviation	4.60	3.98
Variance	21.19	15.87
CV	0.078	0.067
Maximum	72.5	68.2
Minimum	32.6	44.2
Upper quartile	60.7	62.2
Median	58.2	59.2
Lower quartile	56.4	56.15

- **Measures of Association**

  - The correlation between two random variables is a measure of the linear relationship between them.

  - Covariance: Consider two random variables, X and Y, both normally distributed with population means μ_x and μ_y, and population standard deviations δ_x and δ_y, respectively, the population and sample covariance between X and Y are defined, respectively, as follows:

$$COV_p(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Population Covariance

$$COV_s(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

Sample Covariance

> 0, the two variables move in the same direction
< 0, the two variables move in opposite directions
= 0, two variables are not linearly related.

- **Measures of Association**

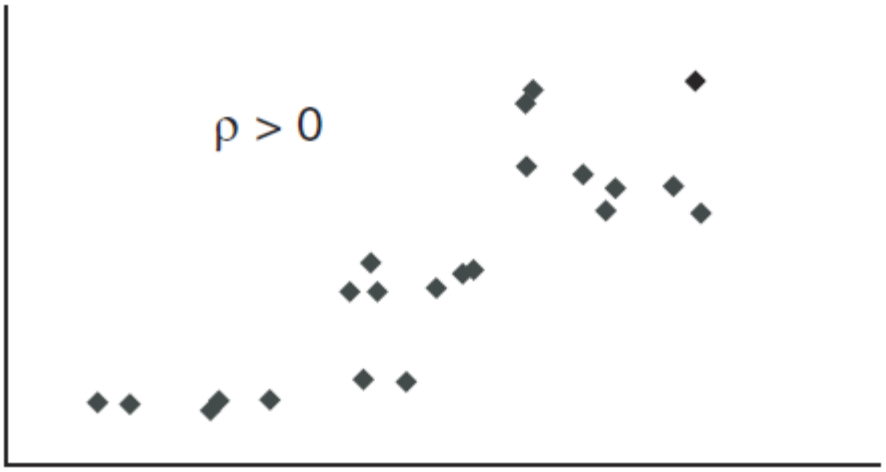
  - Pearson product-moment correlation parameter ρ (sample r) is a commonly used measure of linear correlation and gives a quantitative measure of how well two variables move together.

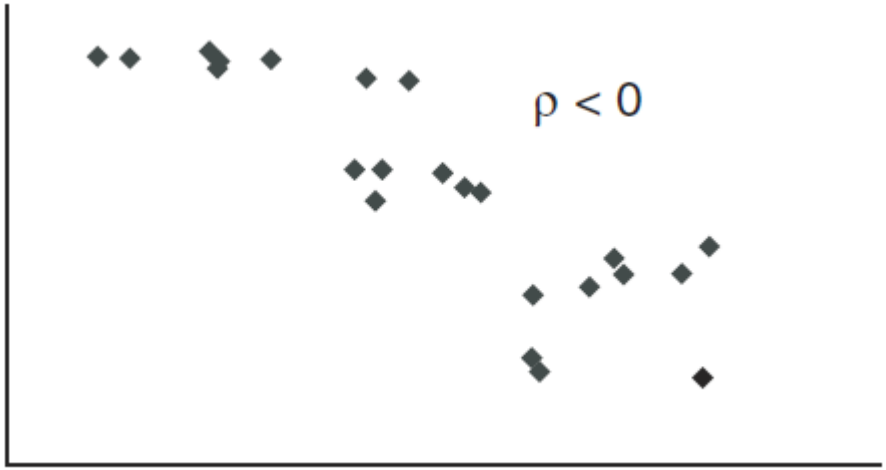
$$\rho = \frac{COV_p(X,Y)}{\sigma_X \sigma_Y} \qquad r = \frac{COV_s(X,Y)}{s_X s_Y}$$

  - The correlation parameter is always in the interval [-1, 1].

  - When ρ=0, there is no linear association, meaning that a linear relationship does not exit between the two variables examined;

  - When ρ>0, there is a positive linear relationship between the variables examined, such that when one of the variables increases the other variable also increases;


ρ > 0


ρ < 0

- **Measures of Association**

  - $\rho = 1$ perfect positively sloped straight-line relationship between two variables;

  - $\rho < 0$ there is a negative linear relationship between the two variables examined such that an increase in one variable is associated with a decrease in the value of the other variable;

  - $\rho = -1$ there is a perfect negatively sloped straight-line relationship between two variables

  - So far, the discussion on correlation has focused solely on continuous variables measured on the interval or ratio scale. In some situations, however, both of the variables may be measured on the ordinal scale. Or not normally distributed

- **Properties of Estimators**

  - "Good" statistical estimators of true population parameters satisfy four important properties: unbiasedness, efficiency, consistency, and sufficiency

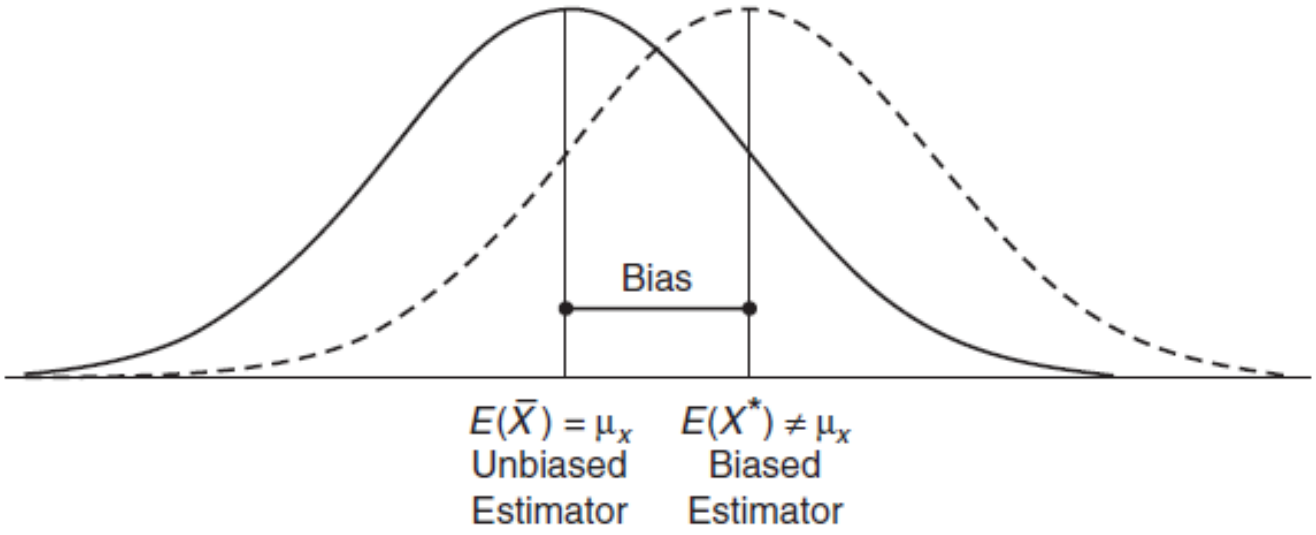
  - Unbiasedness: An estimator is said to be unbiased if its expected value is equal to the true population parameter it is meant to estimate.

  The sample average $\overline{X}$ is an unbiased estimator of μ_X if

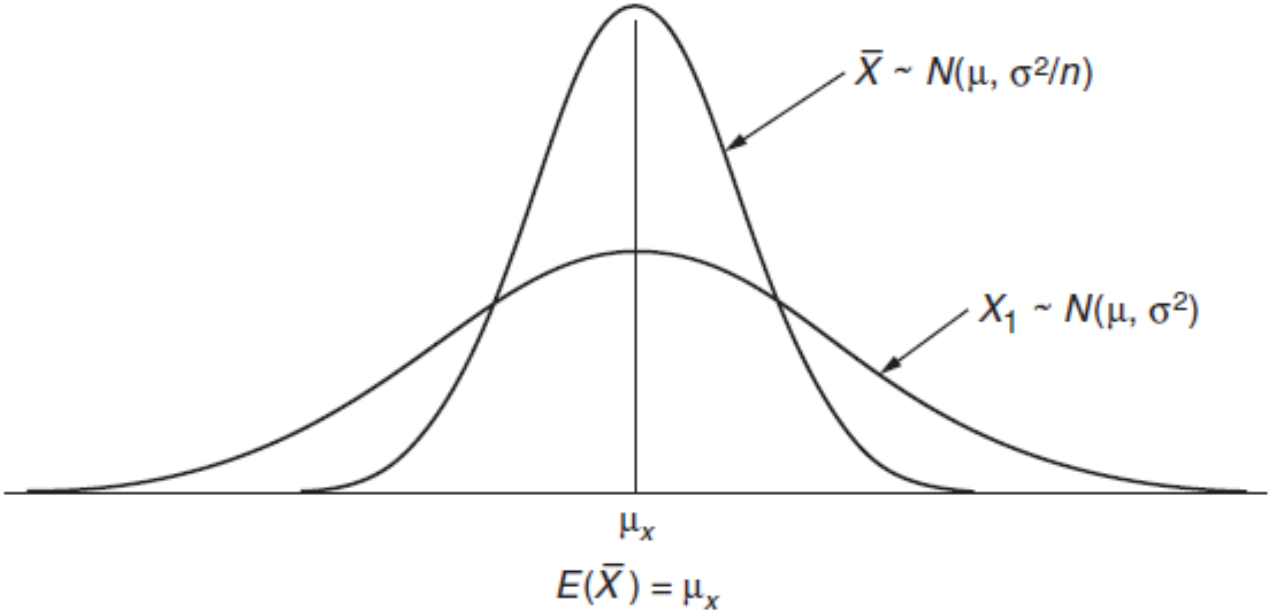
  $$E(\overline{X}) = \mu_X$$

  - Efficiency: Efficiency is a relative property in that an estimator is efficient relative to another, which means that an estimator has a smaller variance than an alternative estimator. An estimator with the smaller variance is more efficient.

  - Unbiasedness and efficiency hold true for any finite sample *n*, and when *n* approaches infinite they become asymptotic properties



The property of unbiasedness



The property of efficiency

- **Properties of Estimators**

  - Consistency: An estimator is said to be consistent if the probability of being closer to the true value of the parameter it estimates (θ) increases with increasing sample size.

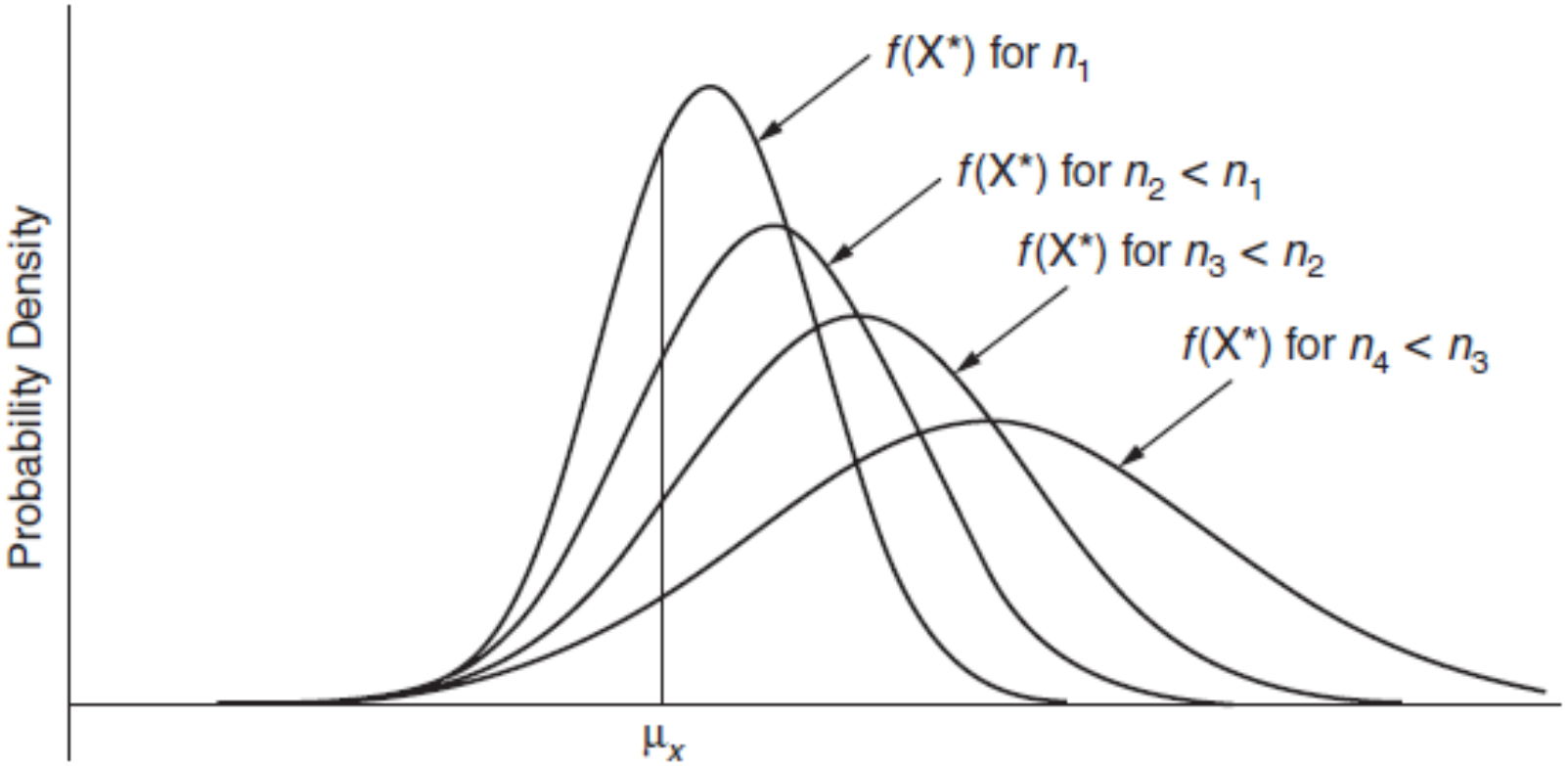
  - As $n \to \infty$, $\lim P[\left|\hat{\theta} - \theta\right| > c] = 0$ for any arbitrary constant c

  - A statistical estimator may not be an unbiased estimator; however it may be a consistent one. A sufficient condition for an estimator to be consistent is that it is asymptotically unbiased and that its variance tends to zero as $n \to \infty$

  - Sufficiency: An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates

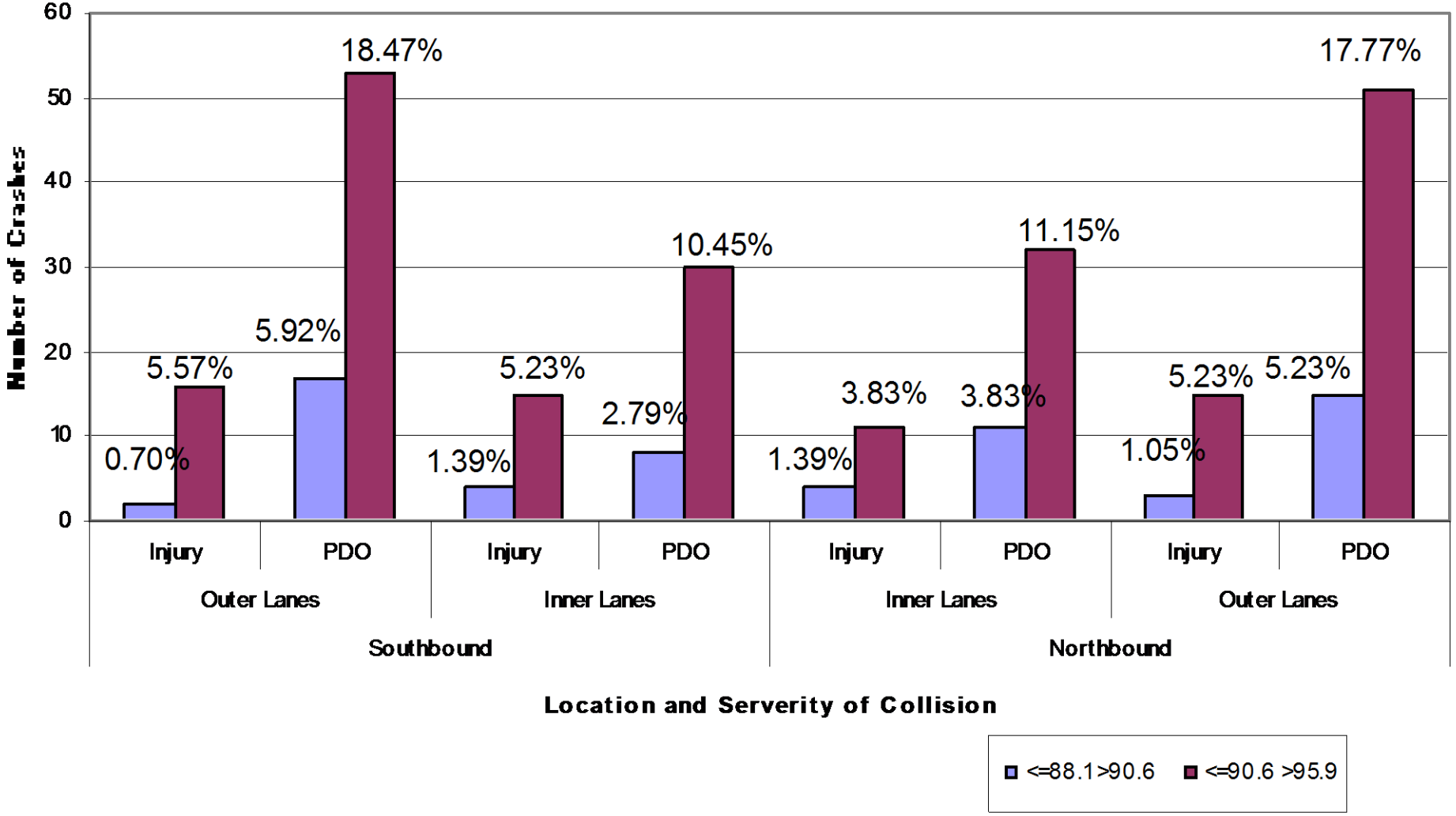
- **Methods of Displaying Data (Graphical methods)**

  - Histogram
  - Ogives
  - Box Plots
  - Scatter Diagrams
  - Bar and Line Charts

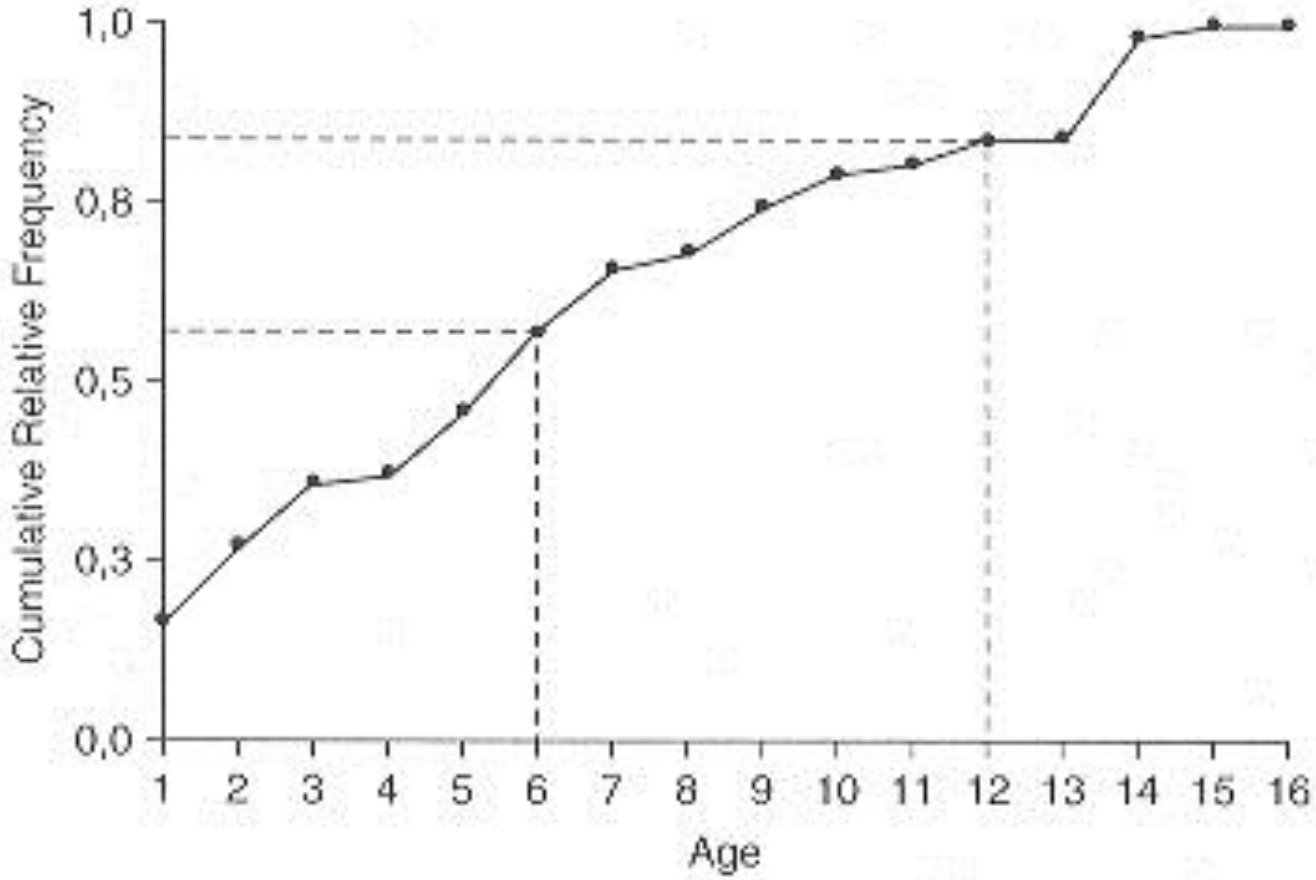


The property of consistency

Histograms

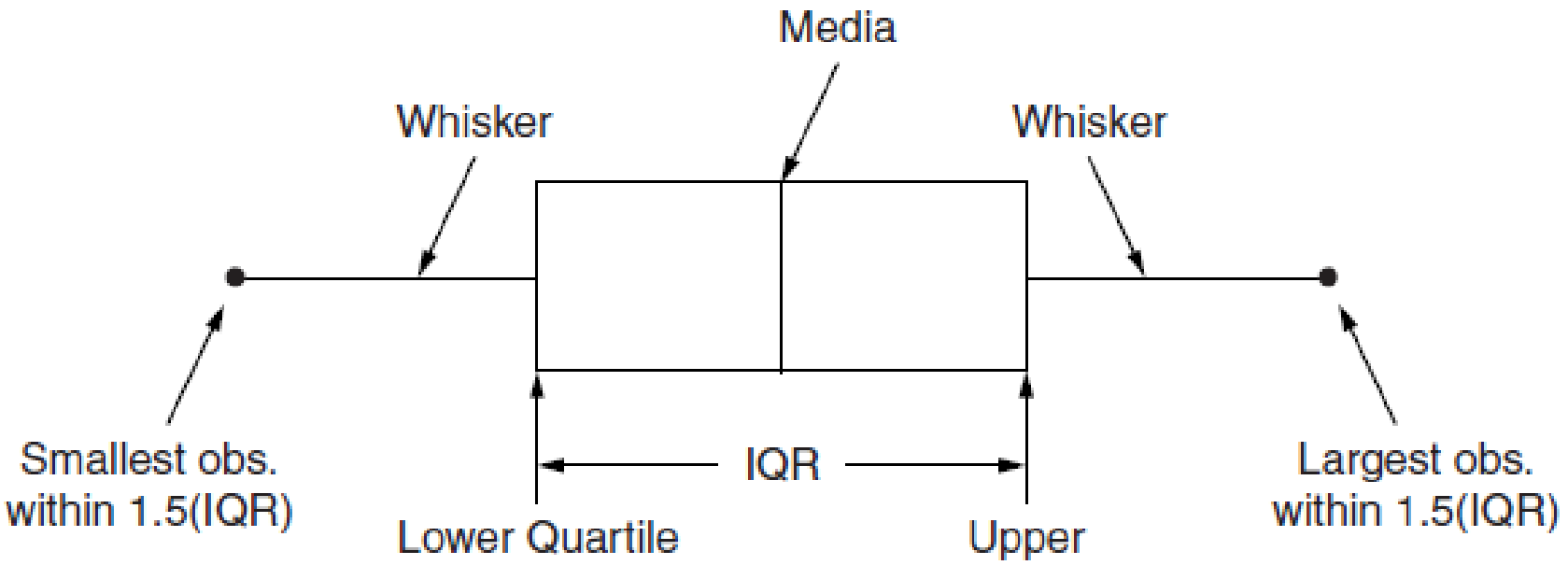

60
50
40
30
20
10
0
Number of Crashes
0.70%
5.57%
5.92%
18.47%
1.39%
5.23%
2.79%
10.45%
1.39%
3.83%
3.83%
11.15%
1.05%
5.23%
5.23%
17.77%
Injury
PDO
Injury
PDO
Injury
PDO
Injury
PDO
Outer Lanes
Inner Lanes
Inner Lanes
Outer Lanes
Southbound
Northbound
Location and Serverity of Collision
<=88.1>90.6
<=90.6 >95.9

Ogives

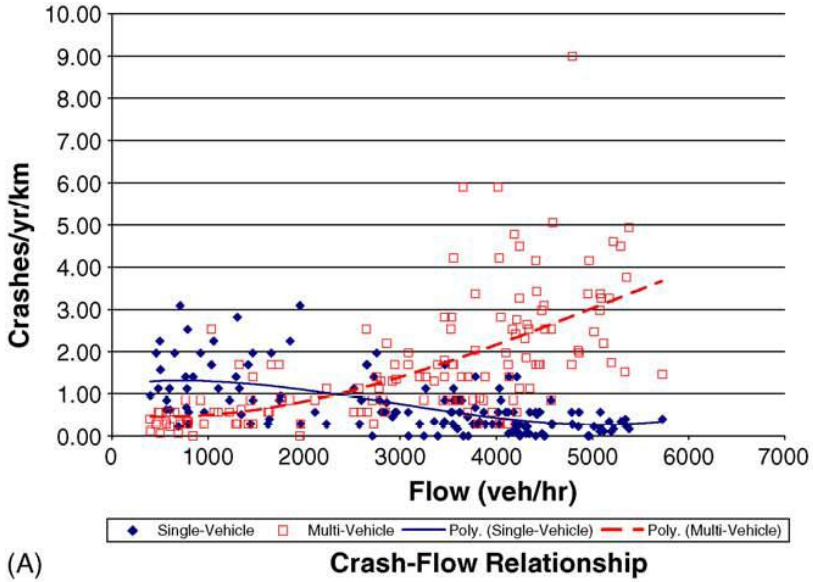


Source: Washington et al. (2003)

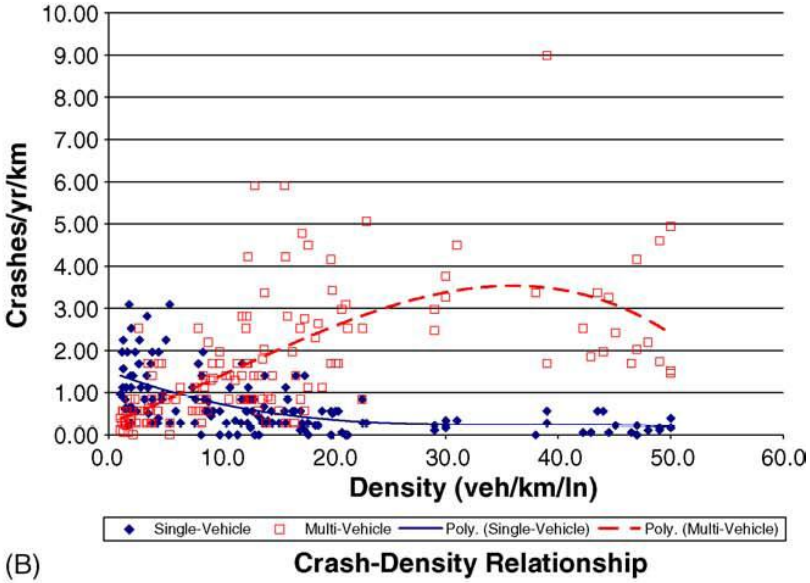
Box Plots


Media
Whisker
Whisker
Smallest obs.
within 1.5(IQR)
IQR
Largest obs.
within 1.5(IQR)
Lower Quartile
Upper

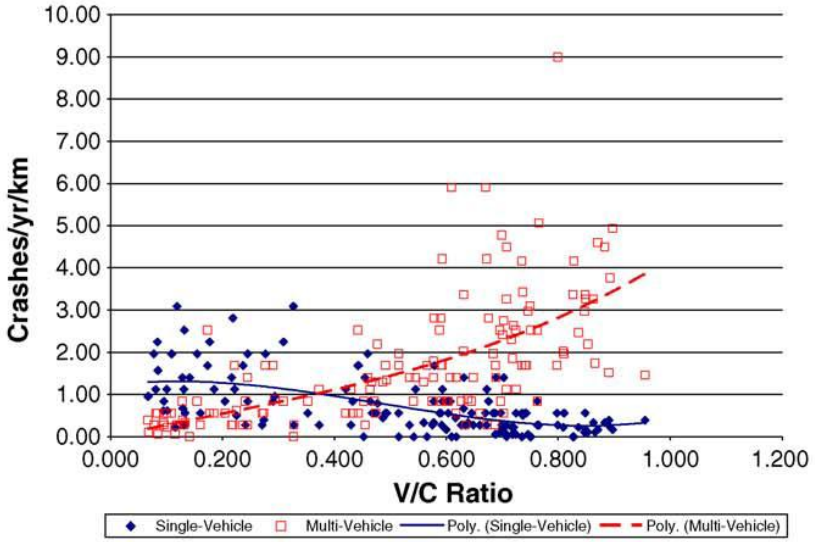
Scatter Diagrams



(A) Crash-Flow Relationship

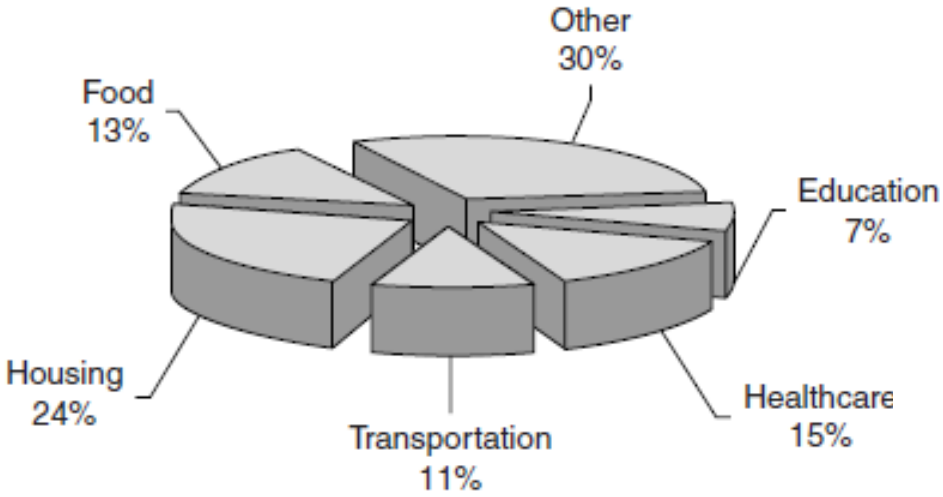


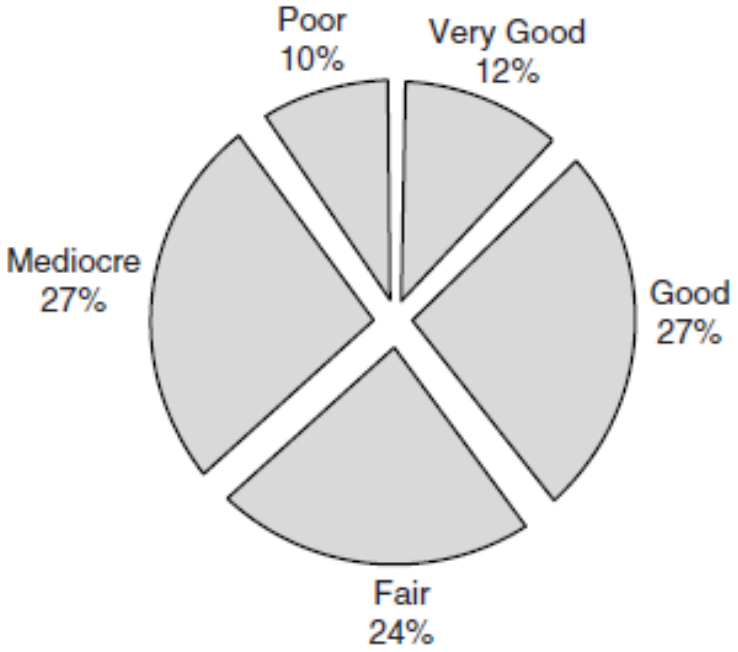
(B) Crash-Density Relationship



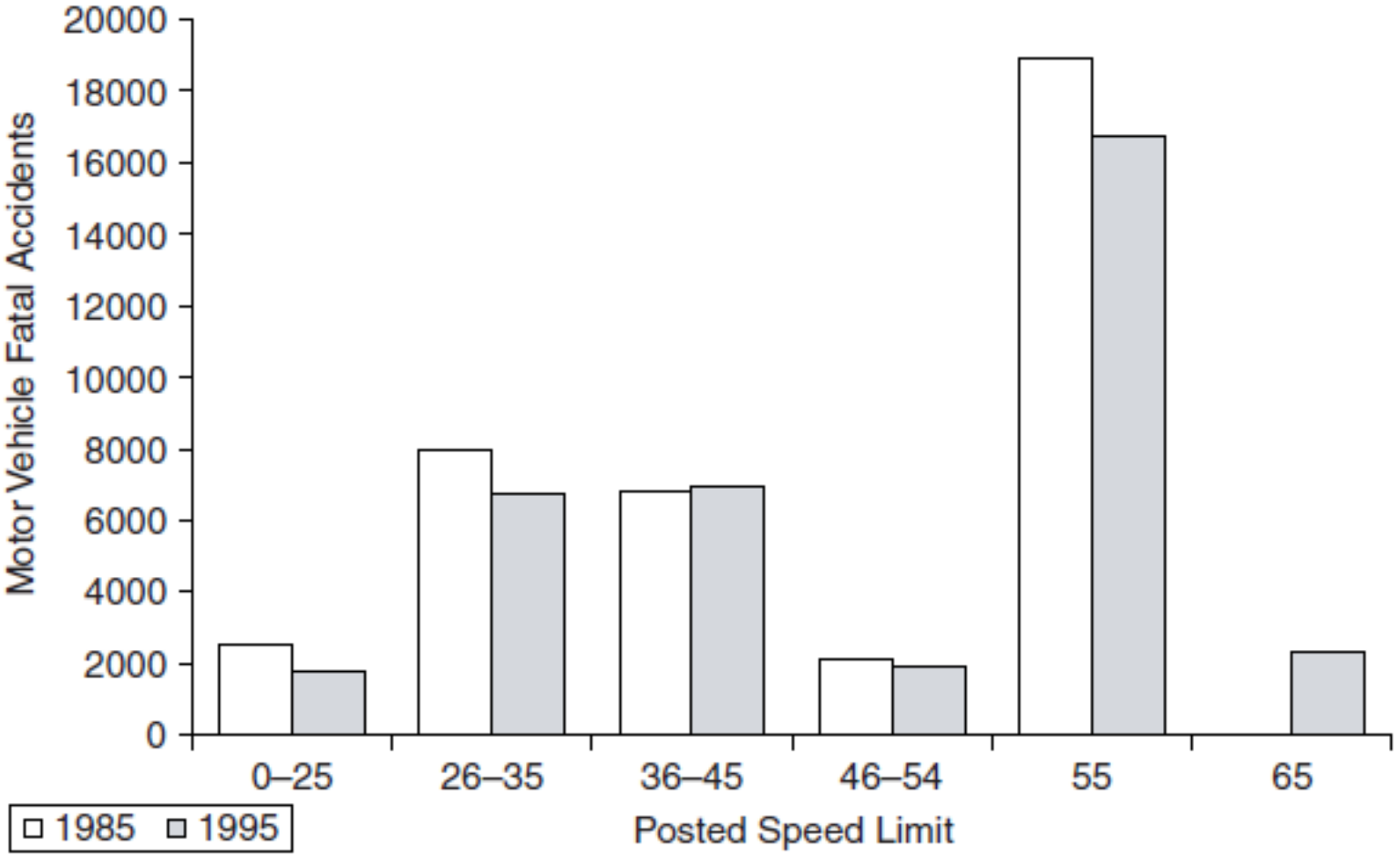
(C) Crash-V/C Ratio Relationship

Bar and Line Charts

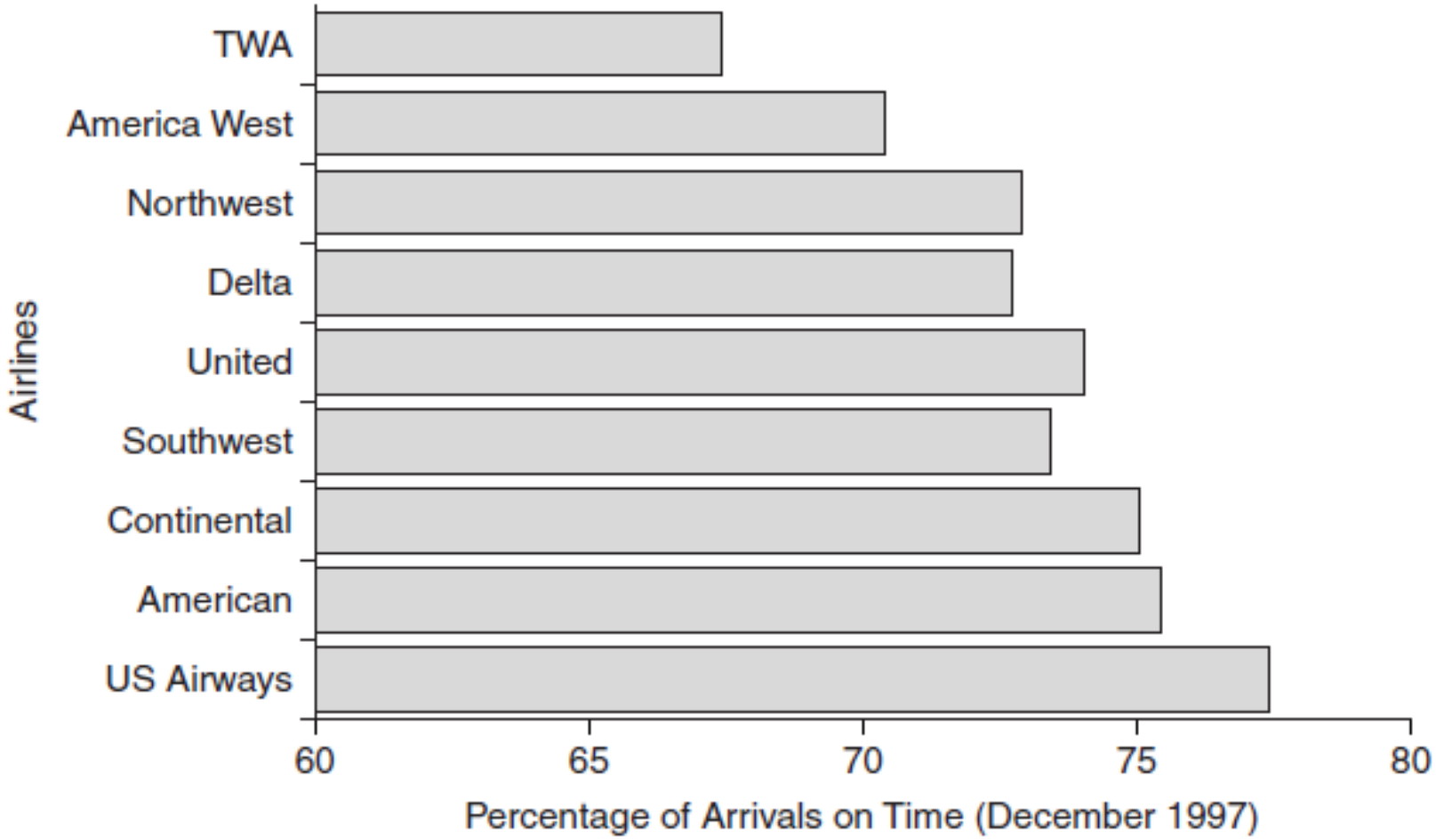




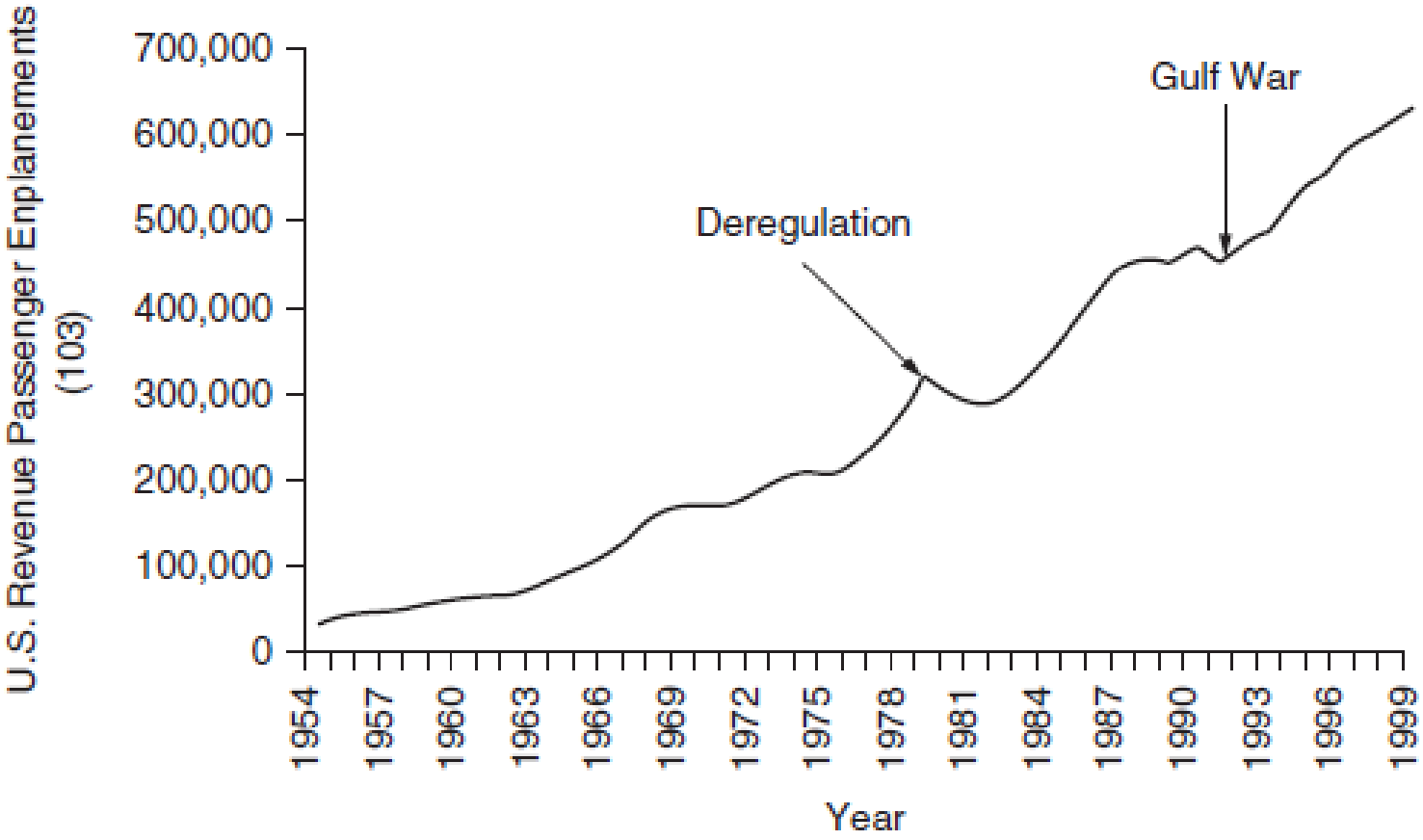
Bar and Line Charts



Bar and Line Charts



Bar and Line Charts



Two by Two Tables

Crash Severity / Flow Range	< 5,000	5,000-9,999	≥ 10,000
Fatal	10	12	15
Non-Fatal Injury	100	120	135
PDO	550	700	900

II. **Point Estimation and Interval Estimation**

- **Point estimators**: using a single parameter calculated from samples to estimate population parameters
- **Interval estimates**: allow inferences to be drawn about a population by providing an interval, a lower and upper boundary, within which an unknown parameter will lie with a prespecified level of confidence.

  - The lower value is called the lower confidence limit (LCL)

  - The upper value is called the upper confidence limit (UCL)
- **Useful Probability Distribution**

  - Two types of probability distribution: continuous and discrete

- Useful Probability Distribution

  - Continuous distributions: arise from variables that can take on any value within a range of values

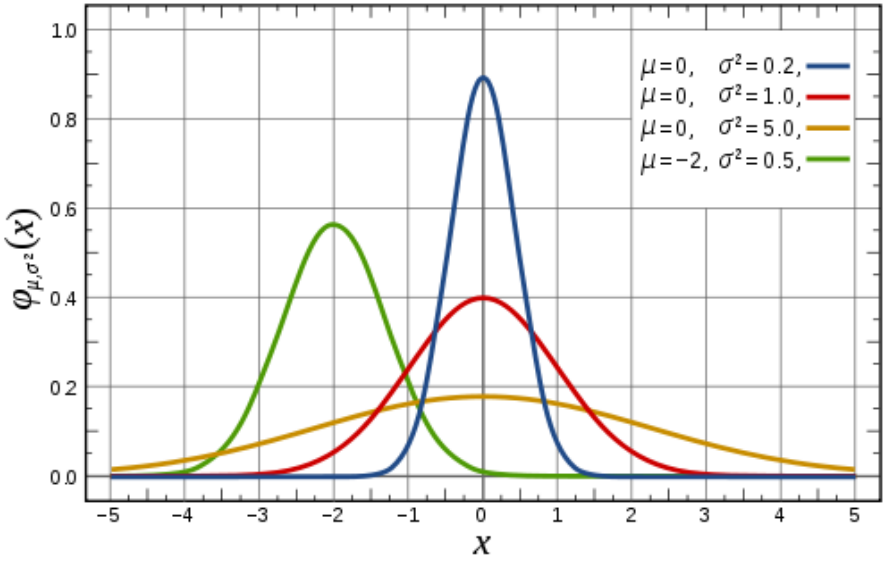
  - Discrete distributions: arise from ordinal data, or count data (data that are strictly continuous but can only take on integer values)
- Two types of random variables:
  - Discrete random variables
  - Continuous random variables

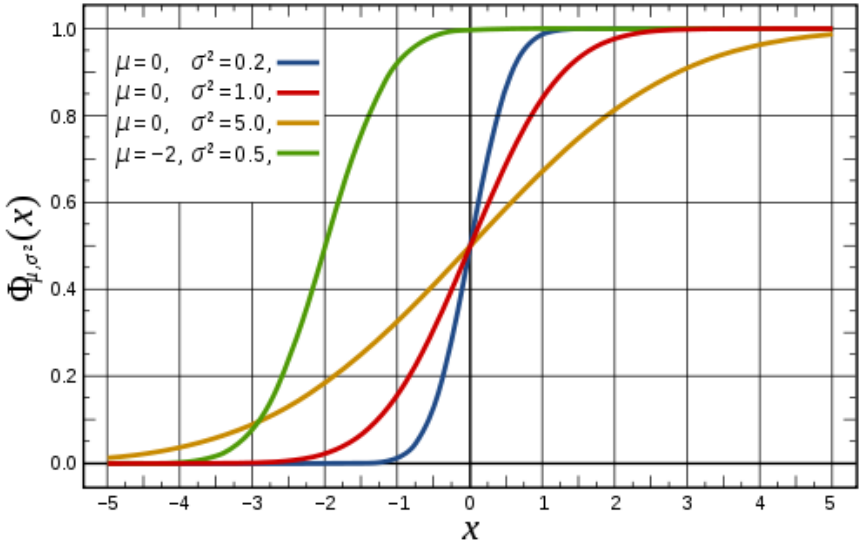
Normal Distribution

- Continuous probability distribution
- Probability Density Function

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Two parameters: μ *(mean) and* σ^2(variance)





- If $X \sim N(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim N(0, 1)$, (standard normal distribution)

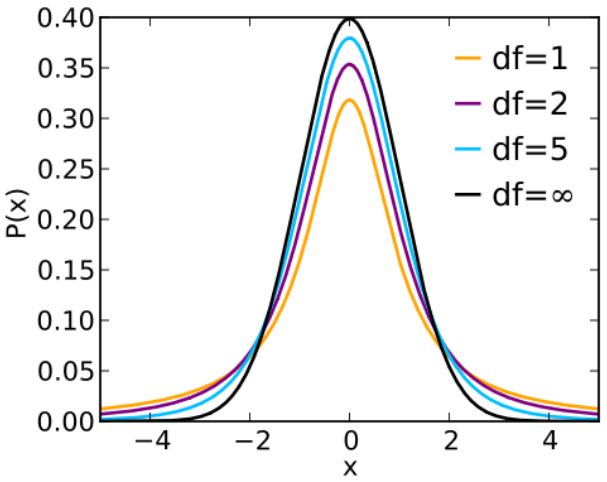
t Distribution

- Continuous probability distribution
- In most cases ∂^2 is unknown. It is natural to replace it with its unbiased estimator s^2. when this is done, a test statistic t* is obtained

$$t^* = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \approx t_\alpha (v = n - 1)$$

where t* is approximately t distributed with

n-1 degrees of freedom



- Continuous probability distribution
- The t distribution is similar to the standard normal distribution.
- As the sample size becomes large, the t distribution approaches the standard normal Z distribution.

Chi-square Distribution

- Statistical theory shows that the square of a standard normal variable Z is χ^2 distributed with 1 degree of freedom.
- Let $Z_1, Z_2, \ldots, Z_k$ be k independent standard normal random variables, if each of these variables is squared, their sum will follow a χ^2 distribution with k degrees of freedom

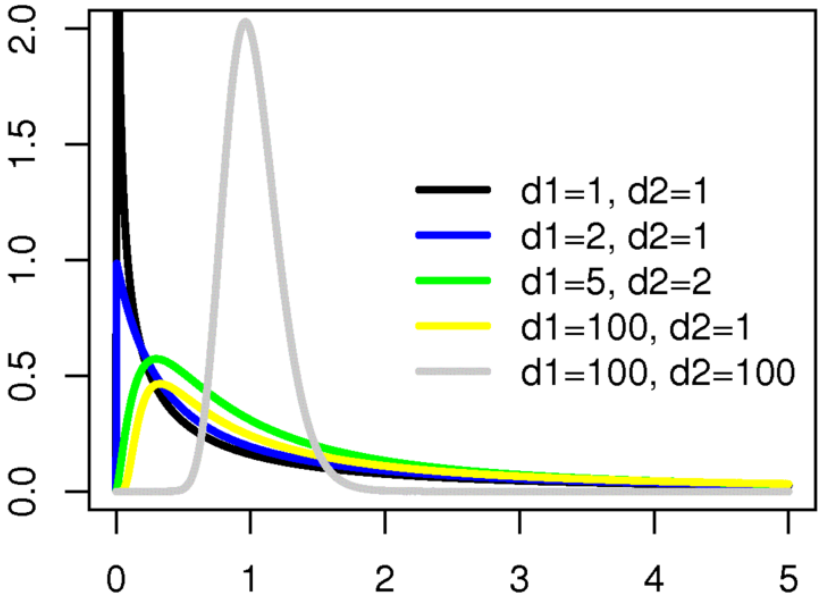
$$X^2 = \sum_{i=1}^{n} Z_i^2$$

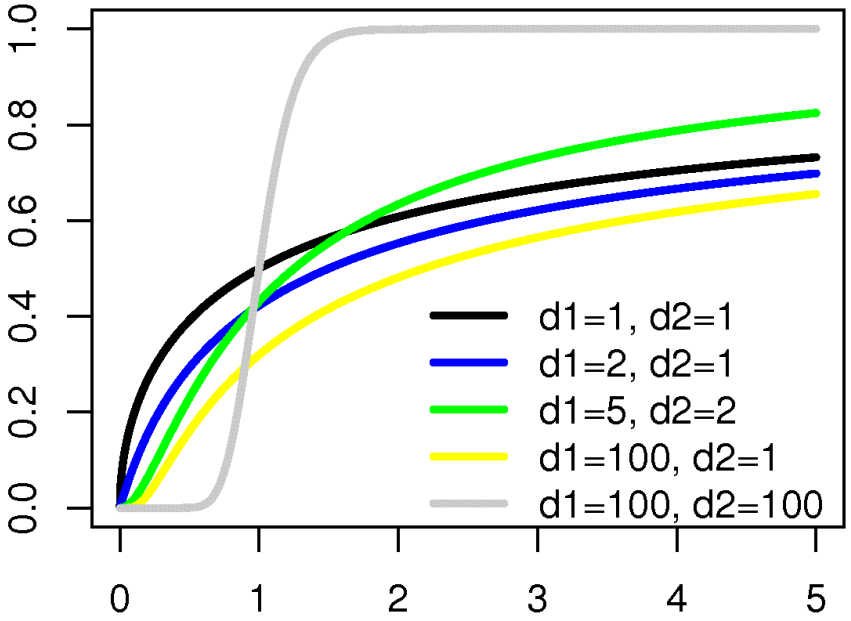
- The χ^2 distribution is skewed to the right
- As the degrees of freedom increase, the distribution approaches a normal distribution with mean equal to the degrees of freedom and variance equal to two times the degrees of freedom

F Distribution

- The F distribution is approximated by the ratio of two independent random variables, each of which is divided by its own degrees of freedom.

$$F^* = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$





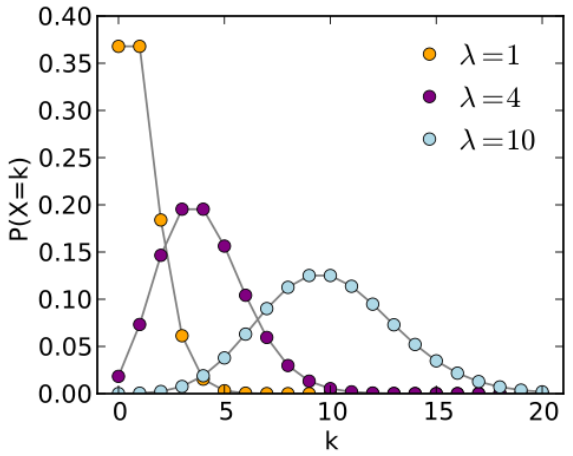
Poisson Distribution

- Discrete probability distribution
- It expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate, and are independent of the time since the last event
- The probability that there are exactly k occurrences (k being a non-negative integer, $k = 0, 1, 2, ...$) is:

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

where λ is a positive number, equal to the expected number of crashes that occur during the given interval

- One parameter: λ
- For Poisson distribution, the mean equals variance



- **Confidence Interval**

  - The logical behind an interval estimate is that an interval calculated using sample data contains the true population parameter with some level of confidence

  - The level of confidence is long run proportion of times that the true population parameter is contained in the interval

- **Confidence Interval for μ with known σ**

  - The central limit theorem (CLT) suggests that whenever a sufficiently large random sample is drawn from any population with mean μ and standard deviation σ, the sample mean $\bar{X}$ is approximately normally distributed with mean μ and standard deviation σ/ $\sqrt{n}$

$$Z^* = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx Z_\alpha$$

- **Confidence Interval for μ with known σ**

  - A standard normal distribution variable Z* has a 0.95 probability of being between the range of values [-1.96, 1.96]. Thus:

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- **Confidence Interval for μ with known σ**

  - With a large number of intervals computed from different random samples drawn from the population, the proportion of values of $\bar{X}$ for which the interval

$$(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

  captures is 0.95. This interval is called the 95% confidence interval of μ.

  - The (1-α)100% confidence interval estimator of can be written as:

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  Where is the value of Z such that the area in each of the tails under the standard normal curve is (α/2)

  Example: A 90% confidence interval is: $\bar{X} \pm 1.645\frac{\sigma}{\sqrt{n}}$

- **Confidence Interval for μ with known σ**

A 95% confidence interval is desired for the mean vehicular speed on Indiana roads (see Example 1.1 for more details). First, the assumption of normality is checked; if this assumption is satisfied we can proceed with the analysis. The sample size is $n = 1296$, and the sample mean is $\overline{X} = 58.86$. Suppose a long history of prior studies has shown the population standard deviation as $\sigma = 5.5$. Using Equation 2.4, the confidence interval can be obtained:

z	$f(z)$	$F(z)$	$1-F(z)$
1.50	0.1295	0.9332	0.0668
1.51	0.1276	0.9345	0.0655
1.52	0.1257	0.9357	0.0643
1.53	0.1238	0.9370	0.0630
1.54	0.1219	0.9382	0.0618
1.55	0.1200	0.9394	0.0606
1.56	0.1182	0.9406	0.0594
1.57	0.1163	0.9418	0.0582
1.58	0.1145	0.9429	0.0570
1.59	0.1127	0.9441	0.0559
1.60	0.1109	0.9452	0.0548
1.61	0.1091	0.9463	0.0537
1.62	0.1074	0.9474	0.0526
1.63	0.1057	0.9485	0.0515
1.64	0.1040	0.9495	0.0505
1.65	0.1023	0.9505	0.0495
1.66	0.1006	0.9515	0.0485
1.67	0.0989	0.9525	0.0475
1.68	0.0973	0.9535	0.0465
1.69	0.0957	0.9545	0.0455
1.90	0.0656	0.9713	0.0287
1.91	0.0644	0.9719	0.0281
1.92	0.0632	0.9726	0.0274
1.93	0.0619	0.9732	0.0268
1.94	0.0608	0.9738	0.0262
1.95	0.0596	0.9744	0.0256
1.96	0.0584	0.9750	0.0250
1.97	0.0573	0.9756	0.0244
1.98	0.0562	0.9761	0.0238
1.99	0.0551	0.9767	0.0233

z	$f(z)$	$F(z)$	$1-F(z)$
2.00	0.0540	0.9772	0.0227
2.01	0.0529	0.9778	0.0222
2.02	0.0519	0.9783	0.0217
2.03	0.0508	0.9788	0.0212
2.04	0.0498	0.9793	0.0207
2.05	0.0488	0.9798	0.0202
2.06	0.0478	0.9803	0.0197
2.07	0.0468	0.9808	0.0192
2.08	0.0459	0.9812	0.0188
2.09	0.0449	0.9817	0.0183
2.10	0.0440	0.9821	0.0179
2.11	0.0431	0.9826	0.0174
2.12	0.0422	0.9830	0.0170
2.13	0.0413	0.9834	0.0166
2.14	0.0404	0.9838	0.0162
2.15	0.0396	0.9842	0.0158
2.16	0.0387	0.9846	0.0154
2.17	0.0379	0.9850	0.0150
2.18	0.0371	0.9854	0.0146
2.19	0.0363	0.9857	0.0143
2.20	0.0355	0.9861	0.0139
2.21	0.0347	0.9865	0.0135
2.22	0.0339	0.9868	0.0132
2.23	0.0332	0.9871	0.0129
2.24	0.0325	0.9875	0.0126
2.25	0.0317	0.9878	0.0122
2.26	0.0310	0.9881	0.0119
2.27	0.0303	0.9884	0.0116
2.28	0.0296	0.9887	0.0113
2.29	0.0290	0.9890	0.0110
2.30	0.0283	0.9893	0.0107
2.31	0.0277	0.9896	0.0104
2.32	0.0271	0.9898	0.0102
2.33	0.0264	0.9901	0.0099
2.34	0.0258	0.9904	0.0096
2.35	0.0252	0.9906	0.0094
2.36	0.0246	0.9909	0.0091
2.37	0.0241	0.9911	0.0089
2.38	0.0235	0.9913	0.0087

- **Confidence Interval for μ with unknown variance**

  - In the majority of practical sampling situations, the population variance is rarely known and the population is normally distributed, a (1-α)100% confidence interval for is given by:

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

  where $t_{\alpha/2}$ is the value of the t distribution with n-1 degrees of freedom

- **Confidence Interval for a population proportion**

  - For qualitative variable, rather than a quantitative variable, there might be interest in the relative frequency of some characteristic in a population.

- **Confidence Interval for a population proportion**

  - In such cases, an estimate of the population proportion, p, whose estimator has an approximately normal distribution provided that n is sufficiently large ($np \geq 5$ and $nq \geq 5$, where q=1-p). The mean of the sampling distribution is the population proportion p and the standard deviation is pq/n

  - A large sample (1-α)100% confidence interval for the population proportion, p is given by

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Where the estimated sample proportion, $\hat{p}$, is equal to the number of "success" in the sample divided by the sample size, n, and $\hat{q} = 1 - \hat{p}$

- **Confidence Interval for a population proportion**

A transit planning agency wants to estimate, at a 95% confidence level, the share of transit users in the daily commute "market" (that is, the percentage of commuters using transit). A random sample of 100 commuters is obtained and it is found that 28 people in the sample are transit users. By using Equation 2.6, a 95% confidence interval for p is calculated as

z	$f(z)$	$F(z)$	$1-F(z)$
1.50	0.1295	0.9332	0.0668
1.51	0.1276	0.9345	0.0655
1.52	0.1257	0.9357	0.0643
1.53	0.1238	0.9370	0.0630
1.54	0.1219	0.9382	0.0618
1.55	0.1200	0.9394	0.0606
1.56	0.1182	0.9406	0.0594
1.57	0.1163	0.9418	0.0582
1.58	0.1145	0.9429	0.0570
1.59	0.1127	0.9441	0.0559
1.60	0.1109	0.9452	0.0548
1.61	0.1091	0.9463	0.0537
1.62	0.1074	0.9474	0.0526
1.63	0.1057	0.9485	0.0515
1.64	0.1040	0.9495	0.0505
1.65	0.1023	0.9505	0.0495
1.66	0.1006	0.9515	0.0485
1.67	0.0989	0.9525	0.0475
1.68	0.0973	0.9535	0.0465
1.69	0.0957	0.9545	0.0455
1.90	0.0656	0.9713	0.0287
1.91	0.0644	0.9719	0.0281
1.92	0.0632	0.9726	0.0274
1.93	0.0619	0.9732	0.0268
1.94	0.0608	0.9738	0.0262
1.95	0.0596	0.9744	0.0256
1.96	0.0584	0.9750	0.0250
1.97	0.0573	0.9756	0.0244
1.98	0.0562	0.9761	0.0238
1.99	0.0551	0.9767	0.0233

z	$f(z)$	$F(z)$	$1-F(z)$
2.00	0.0540	0.9772	0.0227
2.01	0.0529	0.9778	0.0222
2.02	0.0519	0.9783	0.0217
2.03	0.0508	0.9788	0.0212
2.04	0.0498	0.9793	0.0207
2.05	0.0488	0.9798	0.0202
2.06	0.0478	0.9803	0.0197
2.07	0.0468	0.9808	0.0192
2.08	0.0459	0.9812	0.0188
2.09	0.0449	0.9817	0.0183
2.10	0.0440	0.9821	0.0179
2.11	0.0431	0.9826	0.0174
2.12	0.0422	0.9830	0.0170
2.13	0.0413	0.9834	0.0166
2.14	0.0404	0.9838	0.0162
2.15	0.0396	0.9842	0.0158
2.16	0.0387	0.9846	0.0154
2.17	0.0379	0.9850	0.0150
2.18	0.0371	0.9854	0.0146
2.19	0.0363	0.9857	0.0143
2.20	0.0355	0.9861	0.0139
2.21	0.0347	0.9865	0.0135
2.22	0.0339	0.9868	0.0132
2.23	0.0332	0.9871	0.0129
2.24	0.0325	0.9875	0.0126
2.25	0.0317	0.9878	0.0122
2.26	0.0310	0.9881	0.0119
2.27	0.0303	0.9884	0.0116
2.28	0.0296	0.9887	0.0113
2.29	0.0290	0.9890	0.0110
2.30	0.0283	0.9893	0.0107
2.31	0.0277	0.9896	0.0104
2.32	0.0271	0.9898	0.0102
2.33	0.0264	0.9901	0.0099
2.34	0.0258	0.9904	0.0096
2.35	0.0252	0.9906	0.0094
2.36	0.0246	0.9909	0.0091
2.37	0.0241	0.9911	0.0089
2.38	0.0235	0.9913	0.0087

- **Confidence Interval for the population variance**

  - In many situations, interest center on the population variance (or a related measure such as the population standard deviation)

  - A (1-α)100% confidence interval for , assuming the population is normally distributed, is given by:

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

  where $\chi^2_{\alpha/2}$ is the value of the χ^2 distribution with n-1 degrees of freedom.

	α							
v	.9999	.9995	.999	.995	.99	.975	.95	.90
41	15.48	17.54	18.58	21.42	22.91	25.21	27.33	29.91
42	16.07	18.19	19.24	22.14	23.65	26.00	28.14	30.77
43	16.68	18.83	19.91	22.86	24.40	26.79	28.96	31.63
44	17.28	19.48	20.58	23.58	25.15	27.57	29.79	32.49
45	17.89	20.14	21.25	24.31	25.90	28.37	30.61	33.35
46	18.51	20.79	21.93	25.04	26.66	29.16	31.44	34.22
47	19.13	21.46	22.61	25.77	27.42	29.96	32.27	35.08
48	19.75	22.12	23.29	26.51	28.18	30.75	33.10	35.95
49	20.38	22.79	23.98	27.25	28.94	31.55	33.93	36.82
50	21.01	23.46	24.67	27.99	29.71	32.36	34.76	37.69
60	27.50	30.34	31.74	35.53	37.48	40.48	43.19	46.46
70	34.26	37.47	39.04	43.28	45.44	48.76	51.74	55.33
80	41.24	44.79	46.52	51.17	53.54	57.15	60.39	64.28
90	48.41	52.28	54.16	59.20	61.75	65.65	69.13	73.29
100	55.72	59.90	61.92	67.33	70.06	74.22	77.93	82.36

	α							
v	.10	.05	.025	.01	.005	.001	.0005	.0001
60	74.40	79.08	83.30	88.38	91.95	99.61	102.69	109.50
70	85.53	90.53	95.02	100.43	104.21	112.32	115.58	122.75
80	96.58	101.88	106.63	112.33	116.32	124.84	128.26	135.78
90	107.57	113.15	118.14	124.12	128.30	137.21	140.78	148.63
100	118.50	124.34	129.56	135.81	140.17	149.45	153.17	161.32

III. Hypothesis Testing

- Hypothesis tests are used to assess the evidence on whether a difference in a population parameter (a mean, variance, proportion, etc.) between two or more groups is likely to have arisen by chance or whether some other factor is responsible for the difference
- Statistical distributions are employed in hypothesis testing to estimate probabilities of observing the sample data, given an assumption about what “should have” occurred
- When observed results are extremely unlikely to have occurred under assumed conditions, then the assumed conditions are considered unlikely
- In statistical terms, the hypothesis test provides the following probability: P(data\true null hypothesis), which is the probability of obtaining or observing the sample data conditional upon a true null hypothesis.

- **Mechanics of Hypothesis Testing**

  - To formulate questions about transportation phenomena a researcher must pose two competing statistical hypothesis: a null hypothesis (the hypothesis to be nullified) and an alternative Hypothesis

  - The null hypothesis (H_o) is an assertion about one or more population parameters that are assumed to be true until there is sufficient statistical evidence to conclude otherwise

  - Need to specify hypotheses correctly

  - Generally, should state the hypothesis in the form of whatever it is hoped or believed will be rejected

  - The alternative hypothesis (H_a) is the assertion of all situations not covered by the null hypothesis

  - The purpose of hypothesis testing is to determine whether it is appropriate to reject or not to reject the null hypothesis

- **Mechanics of Hypothesis Testing**

  - The test statistic is the sample statistic upon which the decision to reject, or **fail to reject**, the null hypothesis is based (Z, t, F, and χ^2)

  - If the observed data reflected through the test statistic falls into the rejection or critical region, the null hypothesis is rejected

  - If the test statistic falls into the acceptance region, the null hypothesis cannot be rejected.

  - Whenever a decision is based on the result of a hypothesis test, there is a chance that it will be incorrect. Four possible results of the hypothesis testing procedure

	True State	**True State**
	H_0	**H_1**
Accept H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

- **Mechanics of Hypothesis Testing**

  - A Type I error occurs when we reject the null hypothesis when it is in fact true

  - A Type II error occurs when we accept the null hypothesis when in fact it is false

  - When the null hypothesis is true, there is α percent chance of rejecting it (Type I error) (significance level)

  - When the null hypothesis is false, there is β percent chance of accepting it (Type II error)

  - The value of $1-\beta$ is called the power of the test

  - Sufficiently large sample sizes will help to minimize the probability of making Type I and Type II errors

- **Formulating one- and Two-tailed Hypothesis Tests**

  - The decision of whether the null hypothesis is rejected or not is based on the rejection region

  - In testing the equality between two average values, we can write:

$$H_0: \mu = \mu_0$$

  - Any one of the following alternative hypotheses may be considered against the null hypothesis for comparison:

$$H_1: m > m_0$$

$$H_1: m < m_0$$

$$H_1: m \neq m_0$$

  - The first two alternative hypotheses would result in a one-tailed test, while the last would yield a two-tailed test

- **The *p*-value of a Hypothesis Test**

  - The *p*-value is the smallest level of significance α that leads to rejection of the null hypothesis.

  - The value quantifies the amount of statistical evidence that supports the alternative hypothesis.

  - The more evidence that exits to reject the null hypothesis in favor of the alternative hypothesis, the larger the test statistic and the smaller is the *p*-value.

  - The *p*-value provides a convenient way to determine the outcome of a statistical test base on any specified Type I error rate α; if the *p*-value is less than or equal to α, then the null hypothesis will be rejected.

  - **For example:** a *p*-value of 0.031 suggests that a null hypothesis will be rejected at $\alpha=0.05$, but will not be rejected at $\alpha=0.01$.

- **Inference regarding a single population**
  - Testing the population mean with unknown variance
  - Testing the population variance
  - Testing for a population proportion
- **Comparing two populations**
  - Testing differences between two means: independent samples
  - Testing differences between two means: Paired observations
  - Testing differences between two population proportions
  - Testing the equality of two population variances

Testing the population mean with unknown variance

$$Z^* = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

$$t^* = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

Testing for a population proportion

$$Z^* = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

Testing the population variance

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

Testing the population mean with unknown variance

Using the data from Example 2.1, a test is conducted to assess whether the mean speed on Indiana roads is 59.5 mph at the 5% significance level. The sample size is $n = 1296$, and the sample mean is $\mu = 58.86$. Suppose that numerous past studies have revealed the population standard deviation to be $\sigma = 5.5$. The parameter of interest is the population mean, and the hypotheses to be tested are

$$H_0: \mu = 59.5$$

$$H_a: \mu \neq 59.5$$

$$Z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{58.86 - 59.5}{5.5/\sqrt{1296}} = -3.27$$

Testing the population mean with unknown variance

$$p\text{-value}\left(Z^{*}=3.27\right)=p\left[Z\le -3.27 \text{ and } Z\ge 3.27\right]$$

$$=2p\left[Z\ge 3.27\right]=2\left[1-p\left[Z\le 3.27\right]\right].$$

$$=2\left[1-.99946\right]=.001$$

Testing differences between two means: independent samples

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

$$Z^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Testing differences between two means: Paired observations

$$H_0: \mu_d = 0$$

$$H_a: \mu_d \neq 0$$

$$t^* = \frac{\overline{X}_d - \mu_d}{s_d / \sqrt{n_d}}$$

Testing differences between two population proportion

$$H_0: p_1 - p_2 = 0$$

$$H_a: p_1 - p_2 \neq 0$$

$$Z^* = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Testing the equality of two population variances

$$H_o: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

$$F^*_{(n_1-1, n_2-1)} = \frac{s_1^2}{s_2^2}$$

IV. Nonparametric Methods

- Statistical procedures discussed previously have focused on making inferences about specific population parameters and have relied upon specific assumptions about the data being satisfied
- One assumption is that the samples examined are approximately normally distributed
- For the means and variances tests discussed, data are required to be measured on a ratio or interval scale
- Finally, sample sizes are required to be sufficiently large
- Nonparametric methods typically require fewer stringent assumptions than do their parametric alternatives, and they use less information contained in the data
- If nonparametric techniques are applied to data that meet conditions suitable for parametric tests, then the likelihood of committing a Type II error increases

- Thus, parametric methods are more powerful and will be more likely to lead correctly to rejection of a false null hypothesis
- It is often said that a parametric procedure is an exact solution to an approximate problem, while a nonparametric procedure is an approximate solution to an exact problem
- A nonparametric technique should be considered under the following conditions:
  - The sample data are frequency counts and a parametric test is not available.
  - The sample data are measured on the ordinal scale.
  - The research hypotheses are not concerned with specific population parameters such as μ and σ^2
  - Requirements of parametric tests such as approximate normality, large sample sizes, and interval or ratio scale data, are grossly violated.
  - There is moderate violation of parametric test requirements, as well as a test result of marginal statistical significance

- **The Chi-Square Test**
  - The Chi-Square test sees widespread use in a variety of transportation analyses.
  - The data used in Chi-Square tests are either counts or frequencies measured across categories that may be measured on any scale.
  - A common five-step process for Chi-Square tests:

  a. Competing hypotheses for a population are stated

  b. Frequencies of occurrence of the events expected under the null are computed. This provides expected counts or frequencies based on some "statistical model," which may be a theoretical distribution, an empirical distribution, an independence model, etc.

  c. Observed counts of data falling in the different cells are noted.

  d. The difference between the observed and the expected counts are computed and summed. The difference leads to a computed value of the Chi-Square test statistic.

  e. The test statistic is compared to the critical points of the Chi-Square distribution and a decision on the null hypothesis is made.

- **The Chi-Square Test**

  - The Chi-Square test statistic is equal to the squared difference between the observed count and the expected count in each cell divided by the expected count and summed over all cells. If the data are grouped into k cells, let the observed count in cell i be O_i and the expected count (expected under H_0) be E_i. The summation is over all cells i = 1, 2,…, k. The test statistic is:

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

  - With increasing sample size and for a fixed number of k cells, the distribution of the Chi-Square test statistic approaches the Chi-Square distribution with k-1 degrees of freedom provided that the expected frequencies are 5 or more for all categories

  - A Chi-Square goodness-of-fit test assesses how well the sample distribution supports an assumption about the population distribution.

- **The Chi-Square Test**
  - For example, in statistical analyses it is often assumed that samples follow the normal distribution; the Chi-Square test can be used to assess the validity of such an assumption.
  - The Chi-Square distribution is useful for other applications besides goodness-of-fit tests. Contingency tables can be helpful in determining whether two classification criteria, such as age and satisfaction with transit services, are independent of each other.

TABLE

General Layout of a Contingency Table

Second Classification Category	First Classification Category			
	1	.	j	Total
1	C_{11}	.	.	R_1
i	.	.	C_{ij}	R_i
Total	C_1	.	C_j	n

H_0: The two classification variables are statistically independent.

H_a: The two classification variables are not statistically independent.

- **The Chi-Square Test**

- The test statistic for a two-way contingency table is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where the differences between observed and expected frequencies are summed over all rows and columns (r and c, respectively). The test statistic is approximately Chi-Square distributed with degrees of freedom df = (r – 1)(c – 1). Finally, the expected count in cell (i, j), where R_j and C_i are the row and column totals, respectively, is

$$E_{ij} = \frac{R_i C_j}{n}$$