<u>Elementary Statistics</u> – by Mario F. Triola, Eighth Edition
DEFININITIONS, RULES AND THEOREMS

CHAPTER 1: INTRODUCTION TO STATISTICS

<u>Section 1- 2: The Nature of Data</u>

Statistics – a collections of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data. **(p. 4)**

Population – complete collection of all elements to be studied **(p. 4)**

Census - collection of data from *every* element in a population **(p. 4)**

Sample – a subcollection of elements drawn from a population **(p. 4)**

Parameter – a numerical measurement describing some characteristic of a *population* **(p. 5)**

Statistic – a numerical measurement describing some characteristic of a *sample* **(p. 5)**

Quantitative data – numbers representing counts or measurements
Ex: incomes of students **(p. 6)**

Qualitative data – can be separated into different categories that are distinguished by some nonnumeric characteristic
Ex: genders of students **(p. 6)**

Discrete data – number of possible values is either a finite number or a "countable" number, Ex: number of cartons of milk on a shelf **(p. 6)**

Continuous (numerical) data – infinitely many possible values on a continuous scale Ex: amounts of milk from a cow **(p. 6)**

Nominal level of measurement – data that consist of names, labels, or categories only, Ex: survey responses of yes, no and undecided **(p. 7)**

Ordinal level of measurement – can be arranged in some order, but differences between data values either cannot be determined or are meaningless
Ex: course grades of A, B, C, D, or F **(p. 7)**

Interval level of measurement – like ordinal level, with the additional property that the difference between any two data values is meaningful but no natural zero starting point. Ex: Body temperatures of 98.2 and 98.6 **(p. 8)**

Ratio level of measurement – the interval level modified to include the natural zero starting point. Ex: weights of diamond rings **(p. 9)**

<u>Section 1- 3: Uses and Abuses of Statistics</u>

Self-selected survey (voluntary response sample) – one in which the respondents themselves decide whether to be included **(p. 12)**

Section 1 - 4: Design of Experiments

Observational study – observe and measure specific characteristics, but we don't attempt to *modify* the subjects being studied **(p. 17)**

Experiment – some *treatment* is applied, then effects on the subjects are observed **(p. 17)**

Confounding – occurs in an experiment when the effects from two or more variables cannot be distinguished from each other **(p. 18)**

Random sample – members of population are selected in such a way that each has an *equal chance* of being selected **(p. 19)**

Simple random sample – of size *n* subjects is selected in such a way that every possible sample of size *n* has the same chance of being selected **(p. 19)**

Systematic sampling – some starting point is selected and than every *k*th element in the population is selected **(p. 20)**

Convenience sampling – simply use results that are readily available **(p. 20)**

Stratified sampling – subdivide population into at least 2 different subgroups (strata) that share the same characteristics, then draw a sample from each stratum **(p. 21)**

Cluster sampling – divide population area into sections (or clusters), then randomly select some of those clusters, and then choose *all* members from those selected clusters **(p. 21)**

Sampling error – the difference between a sample result and the true population result; such an error results from chance sample fluctuations **(p. 23)**

Nonsampling error – occurs when the sample data are incorrectly collected, recorded, or analyzed **(p. 23)**

CHAPTER 2: DESCRIBING, EXPLORING, AND COMPARING DATA

Section 2 - 2: Summarizing Data with Frequency Tables

Frequency table – lists classes (or categories) of values, along with frequencies (or counts) of the number of values that fall into each class **(p. 35)**

Lower class limits – smallest numbers that can belong to the different classes **(p. 35)**

Upper class limits – largest numbers that can belong to the different classes **(p. 35)**

Class boundaries – numbers used to separate classes, but without the gaps created by class limits. **(p. 35)**

Class midpoints – average of lower and upper class limits **(p. 36)**

Class width – difference between two consecutive lower class limits or two consecutive lower class boundaries **(p. 36)**

<u>Section 2 - 3: Pictures of Data</u>
Histogram – bar graph with horizontal scale of classes, vertical scale of frequencies **(p. 42)**

<u>Section 2 - 4: Measures of Center</u>
Measure of center – value at the center or middle of a data set **(p. 55)**

Arithmetic mean or just <u>mean</u> – sum of values divided by total number of values.
Notation: $\overline{x}$ *(pronounced x-bar)* **(p. 55)**

Median – middle value when the original data values are arrange in order from least to greatest. *Notation:* $\tilde{x}$ *(pronounced x-tilde)* **(p. 56)**

Mode – value that occurs most frequently **(p. 58)**

Bimodal – two modes **(p. 58)**

Multimodal – 3 or more modes **(p. 58)**

Midrange – value midway between the highest and lowest valued in the original data set, average of **(p. 59)**

Skewed – not symmetric, extends more to one side than the other **(p. 63)**

Symmetric – left half of its histogram is roughly a mirror image of its right half **(p. 63)**

<u>Section 2 - 5: Measures of Variation</u>

Standard deviation – a measure of variation of values about the mean
Notation: s = sample s.d.; σ *= population s.d.* **(p. 70)**

Variance – a measure of variation equal to the square of the standard deviation
Notation: s^2 *= sample variance;* σ^2 *= population variance* **(p. 74)**

Range Rule of Thumb (p. 77)

- **For estimation of standard deviation:** $s \approx$ <u>range/4</u>
- **For interpretation:** if the standard deviation *s* is known,
  Minimum "usual" value $\approx$ *(mean) – 2 x (standard deviation)*
  Maximum "usual" value $\approx$ *(mean) + 2 x (standard deviation)*

Empirical Rule for Data with a Bell-Shaped Distribution (p. 78)

- About 68% of all values fall within 1 standard deviation of the mean
- About 95% of all values fall within 2 standard deviations of the mean
- About 99.7% of all values fall within 3 standard deviations of the mean

Chebyshev's Theorem (p. 80)
The proportion of any set of data lying with *K* standard deviation of the mean is always *at least* $1\text{-}1/K^2$, where *K* is any positive number greater than 1. For K=2 and K=3, we get the following results:

- At least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean
- At least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean

Section 2 - 6: Measures of Position

Standard score, or **z score –** the number of standard deviations that a given value *x* is above or below the mean

Sample	Population
$z = \dfrac{x - \bar{x}}{s}$	$z = \dfrac{x - \mu}{\sigma}$

Section 2 - 7: Exploratory Data Analysis (EDA)

Exploratory data analysis - is the process of using statistical tools to investigate data sets in order to understand their important characteristics **(p. 94)**

5-number summary – minimum value; the first quartile, Q_1; the median, or second quartile, Q_2; the third quartile, Q_3; and the maximum value **(p. 96)**

Boxplot (or **box-and-whisker diagram) –** graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at Q_1; the median; and Q_3. **(p. 96)**

CHAPTER 3: PROBABILITY

Section 3 - 1: Overview

Rare Event Rule for Inferential Statistics (p. 114)
If under a given assumption (such as a lottery being fair), the probability of a particular observed event (such as five consecutive lottery wins) is extremely small, we conclude that the assumption is probably not correct.

Section 3 - 2: Fundamentals

Event – any collection of results or outcomes of a procedure **(p. 114)**

Simple event – outcome or event that cannot be further broken down inter simpler components **(p. 114)**

Sample space – all possible *simple* events for a procedure **(p. 114)**

Rule 1: Relative Frequency Approximation of Probability (p. 115)

$$P(A) = \frac{\text{number of times A occurred}}{\text{number of times trial was repeated}}$$

Rule 2: Classical Approach to Probability (Requires Equally Likely Outcomes) (p. 115)

$$P(A) = \frac{\text{number of ways A can occur}}{\text{number of difference simple events}} = \frac{s}{\sqrt{n}}$$

Rule 3: Subjective Probabilities (p. 115)
P(A), is found by simply guessing or estimating its value based on knowledge of the relevant circumstances.

Law of Large Numbers (p. 116)
As a procedure is repeated again and again, the relative frequency probability (from Rule 1) of an event tends to approach the actual probability.

Complement – of a, denoted by $\bar{A}$, consists of all outcomes in which event a does *not* occur **(p. 120)**

Actual odds against – ratio of event A not occurring to event A occurring:
$P(\bar{A}) / P(A)$ **(p. 121)**

Actual odds in favor – ratio or event A occurring to event A not occurring
$P(A) / P(\bar{A})$ **(p. 121)**

Payoff odds – ratio of net profit (if you win) to the amount bet **(p. 121)**

Section 3 - 3: Addition Rule

Compound event – any event combining two or more simple events **(p. 128)**

Formal Addition Rule (p. 128)
P(A or B) = P(A) + P(B) – P(A and B)

Intuitive Addition Rule (p. 128)
Find the sum of the number of ways event A can occur and the number of ways event B can occur*, adding in such a way that every outcome is counted only once*. P(A or B) is equal to that sum, divided by the total numbers of outcomes.

Mutually exclusive – cannot occur simultaneously **(p. 129)**

Section 3 - 4: Multiplication Rule: Basics

Independent – occurrence of one event does not affect the probability of the occurrence of the other **(p. 137)**

Formal Multiplication Rule (p. 138)
$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

Intuitive Multiplication Rule (p. 138)
Multiply the probability of event A by the probability of event B, but be sure that the probability of event B takes into account the previous occurrence of event A.

Section 3 - 5: Multiplication Rule: Complements and Conditional Probability

Conditional probability – (p. 145) $P(B \mid A) = \dfrac{P(A \text{ and } B)}{P(A)}$

Section 3 - 6: Probabilities Through Simulations
Simulation – process that behaves the same way as the procedure, so that similar results are produced **(p. 151)**

Section 3 - 7: Counting

Fundamental Counting Rule (p. 156)
For a sequence of two events in which the first event can occur *m* ways, the second *n* ways, the events together can occur a total of $m \cdot n$ ways

Factorial Rule (p. 158)
A collection of *n* different items can be arranged in order *n!* different ways

Permutations Rule (When Items Are All Different) (p. 158)
(without replacement, order matters) nPr = $\frac{n!}{(n-r)!}$

Permutations Rule (When Some Items Are Identical to Others) (p. 160)

$$\frac{n!}{n_1 n_2 \cdots n_k}$$

Combinations Rule (p. 161) (order does not matter)

nCr = $\frac{n!}{(n-r)!r!}$

CHAPTER 4: PROBABILITY DISTRIBUTIONS

SECTION 4 - 2: Random Variables
Random variable – a variable with a single numerical value, determined by chance, for each outcome of a procedure **(p. 181)**

Probability distribution – a graph, table or formula that gives the probability for each value of the random variable **(p. 181)**
1. $\sum P(x) = 1$ where x assumes all possible values
2. $0 \le P(x) \le 1$ for every value of x

Discrete random variable – finite or countable number of values **(p. 181)**

Continuous random variable – has infinitely many values, and those values can be associated with measurements on a continuous scale with no gaps or interruptions **(p. 181)**

Section 4 - 3: Binomial Probability Distributions
Binomial probability distribution – results from a procedure that meets all the following requirements: **(p. 194)**
1. The procedure has a *fixed number of trials.*
2. The trials must be *independent.*
3. Each trail must have all outcomes classified into *two categories.*
4. The probabilities must remain *constant* for each trial.

Section 4 - 5: The Poisson Distribution
Poisson distribution – a discrete probability distribution that applies to occurrences of some event *over a specified interval such as time, distance, area, or volume* **(p. 210)**

$P(x) = \frac{\mu^x * e^{-u}}{x!}$ where *e* = 2.71828

CHAPTER 5: NORMAL PROBABILITY DISTRIBUTIONS

Section 5 - 1: Overview

Normal distribution – a distribution with a graph that is symmetric and bell-shaped **(p. 226)**

Section 5 - 2: The Standard Normal Distribution

Uniform distribution – one of continuous random variable with values spread evenly over the range of possibilities and rectangular in shape **(p. 227)**

Density curve (or **probability density function) –** a graph of continuous probability distribution with **(p. 227)**

1. The total area under the curve equal to 1.
2. Every point on the curve must have a vertical height that is 0 or greater.

Standard normal distribution – a normal probability distribution that has a mean of 0 and a s.d. of 1 **(p, 229)**

Section 5 - 5: the Central Limit Theorem

Sampling distribution – of the mean is the probability distribution of sample means, with all samples having the same sample size *n*.**(p. 256)**

Central Limit Theorem (p. 257)
Given:

1. The random variable x has a distribution with mean μ and s.d σ.
2. Samples all of the same size n are randomly selected from the population of x values.

Conclusions:

1. The distribution of sample means $\bar{x}$ will approach a *normal* distribution, as the sample size increases.
2. The mean of the sample means will approach the population mean μ.
3. The standard deviation of the sample means will approach σ / n.

Section 5 - 6: Normal Distribution as approximation to Binomial Dist.

If $np \geq 5$ and $nq \geq 5$, then the binomial random variable is approximately normally distributed with the mean and s.d. given as **(p. 268)**

$$\mu = np \quad \sigma = \sqrt{npq}$$

Continuity correction - A single value x represented by the *interval* from x - 0.5 to x + 0.5 when the normal distribution (continuous) is used as an approximation to the binomial distribution (discrete) **(p. 272)**

Section 5 - 7: Determining Normality

Normal quantile plot – a graph of points (x, y), where each x value is from the original set of sample data, and each y value is a z score corresponding to a quantile value of the standard normal distribution.

CHAPTER 6: ESTIMATES AND SAMPLE SIZES

Section 6 - 2: Estimating a Population Mean: Large Samples

Estimator – a formula or process for using sample data to estimate a population parameter **(p. 297)**

Estimate – specific value or range of values used to approximate a population parameter **(p. 297)**

Point estimate – a single value (or point) used to approximate a population parameter, *the sample mean* $\bar{x}$ *being the best point estimate* **(p. 297)**

Confidence interval – a range (or interval) of values used to estimate the true value of a population parameter **(p. 298)**

Degree of confidence (or **level of confidence** or **confidence coefficient)–** the probability $1 - \alpha$ that is the relative frequency of times that the confidence interval actually does contain the population parameter **(p. 299)**

Critical value – the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur **(p. 301)** $Z_{a/2}$ is a critical value

Margin of error (*E*) – the maximum likely difference between the observed sample mean $\bar{x}$ and the true value of the population mean μ **(p. 302)**

$$E = Z_{a/2} \cdot \frac{\sigma}{n}$$

Note: If $n > 30$, replace σ by sample standard deviation s.
If $n < 30$, the population must have a normal distribution and we must know the value of σ to use this formula

Confidence interval limits – the two values $\bar{x} - E$ and $\bar{x} + E$ **(p. 303)**

Section 6 - 3: Estimating a Population Mean: Small Samples

Degrees of freedom – the number of sample values that vary after certain restrictions have been imposed on all data values **(p. 314)**

Margin of error (*E*) for the Estimate of μ when *n* < 30 and population is normal (p. 314)

$$E = t_{a/2} \cdot \frac{s}{n}$$ *where* $t_{a/2}$ has $n - 1$ degrees of freedom *Formula 6-2*

Confidence Interval for the Estimate of μ (p. 315)

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = t_{a/2} \cdot \frac{s}{n}$$

Section 6 – 4: Determining Sample Size Required to Estimate μ

Sample Size for Estimating Mean μ (p. 323)

$$n = \frac{z_{a/2}\sigma}{E}^{2}$$ *Formula 6-3*

Where $z_{a/2}$ = critical z score based on the desired degree of confidence
E = desired margin of error σ = population standard deviation

Section 6 - 5: Estimating a Population Proportion

Margin of Error of the Estimate of *p* (p, 331) $E = z_{a/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ *Formula 6-4*

Confidence Interval for the *p* (p, 331) $p - E < p < p + E$

where $E = z_{a/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Sample Size for Estimating Proportion *p* (p. 334)

When an estimate *p* is known: $n = (z_{a/2})^2 - \frac{\hat{p}\hat{q}}{E}$ *Formula 6-5*

When no estimate *p* is known $n = (z_{a/2})^2 - \frac{0.25}{E}$ *Formula 6-6*

Sectiion 6 - 7: Estimating a Population Variance

Chi-Square Distribution (p. 343) $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ *Formula 6-7*

where n = sample size, s^2 = *sample variance,* σ^2 = population variance

Confidence Interval for the Population Variance σ^2

$$\frac{(n-1)s^2}{X_R^2} < \sigma^2 < \frac{(n-1)s^2}{X_L^2}$$

CHAPTER 7: HYPOTHESIS TESTING

Section 7 - 1: Overview

Hypothesis – a claim or statement about a property of a population **(p. 366)**

Section 7 - 2: Fundamental of Hypothesis Testing

Test Statistic (p. 372) $z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$ where $n > 30$ *Formula 7-1*

Power - the probability $(1 - \beta)$ of rejecting a false null hypothesis **(p. 378)**

Section 7 - 3: Testing a Claim about a Mean: Large Samples

***P*-value –** probability of getting a value of the sample test statistic that is *at least as extreme* as the one found from the sample data, assuming that the null hypothesis is true **(p. 387)**

Section 7 - 4: Testing a Claim about a Mean: Small Samples

Test Statistic for Claims about μ when $n \leq 30$ and σ is Unknown (p. 400)

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

Test Statistic for Testing Hypotheses about σ or σ^2 (p. 418) Use *Formula 6-7*

CHAPTER 8: INFERENCES FROM TWO SAMPLES ($n_1 + n_2$)

Section 8 - 2: Inferences about 2 Means: Independent and Large Samples

Independent – if sample values selected from one population are not related to or somehow paired with sample values selected from other population **(p. 438)**

Dependent – if values in one sample are related to values in other sample often referred to as **matched pairs (p. 438)**

Test Statistic for Two Means: Independent and Large Samples (p. 439)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

σ_1 and σ_2: If σ_1 and σ_2 are not known use s_1 and s_2 in their places, provided that both samples are large.

P-value: Use the computed value of the test statistic *z*, and find the *P*-value by following the procedure summarized in Figure 7-8 (p. 388).

Critical values: Based on the significance level α, find critical values by using the procedures introduced in Section 7-2.

Confidence Interval Estimate of $\mu_1 - \mu_2$: (Independent and Large Samples)

$(\bar{x}_1 - x_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - x_2) + E$ **(p. 442**

$$\text{where } E = z_{a/2} \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

CALCULATOR: STAT, TESTS, 2-SampZTest

Section 8 - 3: Inferences about Two Means: Matched Pairs

Test Statistic for Matched Pairs of Sample Data (p. 450)

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where df = n - 1 d = mean value of the differences d

Critical values: If $n \leq 30$, critical values are found in Table A-3 (*t* distribution)
If n > 30, critical values are found in Table A-2 (*z* distribution)

Confidence Intervals $d - E < \mu_d < d - E$

where $E = t_{a/2} \frac{s_d}{\sqrt{n}}$ and degrees of freedom = n - 1

CALCULATOR: Enter data in L1 – L2 → L3, STAT, TESTS, T-Test, use Data, ENTER

Section 8 - 4: Inferences about Two Proportions

Pooled Estimate of p_1 and p_2 (p. 459)

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

Complement of $\bar{p}$ *is* $\bar{q}$, so $\bar{q} = 1 - \bar{p}$

Confidence Interval Estimate of p_1 and p_2 (p. 463)

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

Section 8 - 5: Comparing Variation in Two Samples

Test Statistic for Hypothesis Tests with Two Variances (p. 472)

$$F = \frac{s_1^2}{s_2^2}$$

Critical values: Using Table A-5, we obtain critical F values that are determined by the following three values:

1. The significance level α.
2. Numerator degrees of freedom = $n_1 - 1$
3. Denominator degrees of freedom = $n_2 - 1$

CALCULATOR: TESTS, 2-SampFTEST

Test Statistic (Small Samples with Equal Variances) (p. 481)

$$t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$ where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ and df = $n_1 + n_2 + 1$

Confidence Interval (Small Independent Samples and Equal Variances) (p. 481)

$$(x_1 - x_2) - E < (\mu_1 - \mu_2) < (x_1 - x_2) + E \qquad where E = t_{a/2}\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}$$

Test Statistic (Small Samples with Unequal Variances) (p. 484)

$$t = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$ where df = small of $n_1 – 1$ and $n_2 – 1$

Confidence Interval (Small Independent Samples and Unequal Variances) (p. 484)

$$(x_1 - x_2) - E < (\mu_1 - \mu_2) < (x_1 - x_2) + E \qquad where E = t_{a/2}\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}$$

and df = small of $n_1 – 1$ and $n_2 – 2$

CALCULATOR: TESTS, 2-SampTTEST (for a hypothesis test) **or 2-SampTInt** (for a confidence interval)

CHAPTER 9: CORRELATION AND REGRESSION

Section 9 - 2: Correlation

Correlation – exists between two variables when one of them is related to the other in some way **(p. 506)**

Scatterplot (or **scatter diagram) –** a graph in which the paired (*x, y*) sample data are plotted with a horizontal *x*-axis and a vertical *y*-axis. Each individual (*x, y*) pair is plotted as a single point. **(p. 507)**

Linear correlation coefficient *r* – measures the strength of the linear relationship between the paired *x*- and *y*-values in a *sample*.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2\ n(\Sigma y^2) - (\Sigma y)^2} \qquad -1 \le r \le 1 \qquad \textit{Formula 9-1}$$

Test Statistic *t* for Linear Correlation (p. 514)

$$T = \frac{r}{\sqrt{\frac{1 - \mathrm{r}^2}{\mathrm{n} - 1}}}$$ Critical values: Use Table A-3 with degrees of freedom = $n – 2$

Test Statistic *r* for Linear Correlation (p. 514) Critical values: Refer to Table A-6

Centroid – the point $(\bar{x}, \bar{y})$ of a collection of paired (x, y) data **(p. 517)**

CALCULATOR: Enter paired data in L1 and L2, STAT, TESTS, LinRegTTest. 2nd, Y=, Enter, Enter, Set the *X* list and *Y* list labels to L1 and L2, ZOOM, ZoomStat, Enter

Regression equation – algebraically describes the relationship between the two variables **(p. 525)** $y = b_o + b_1 x$

Regression line (or **line of best fit)** – graph of the regression equation **(p. 525)**
Only for linear relationships

Marginal change in a variable – amount that the regression equation changes when the other variable changes by exactly one unit **(p. 531)**

Outlier – point lying far away from the other data points in a scatterplot **(p. 531)**

Influential points – points that strongly affect the graph of the regression line **(p. 531)**

Residual – difference ($y – y$) between an observed sample y-value and the value of y, which is the value of y that is predicted by using the regression equation. **(p. 532)**
Least-squares property – satisfied by straight line if the sume of the squares of the residuals is the smallest sum possible **(p. 533)**

CALCULATOR: Enter data in lists L1 and L2, STAT, TESTS, LinRegTTest.

<u>Section 9 - 4: Variation and Prediction Intervals</u>
Total deviation - from the mean is the vertical distance $y - \hat{y}$ which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$ **(p. 539)**

Explained deviation – vertical distance $\hat{y}$ - $\bar{y}$, which is the distance between the predicted y-value and the horizontal line passing through the sample $\bar{y}$ **(p. 539)**

Unexplained deviation – vertical distance y - $\hat{y}$, which is the vertical distance between the point (x, y) and the regression line. **(p. 539)**

Coefficient of determination – the amount of variation in y that is explained by the regression line computed as $r^2 = \dfrac{\text{explained variation}}{\text{total variation}}$

Standard error of estimate – a measure of the differences (or distances) between the observed sample y-values and the predicted values y that are obtained using the regression equation give as **(p. 541)**

$$s_c = \sqrt{\frac{\Sigma(\mathrm{y} - \hat{\mathrm{y}})^2}{\mathrm{n} - 2}}$$

Prediction Interval for an Individual y (p. 543)
Given the fixed value $x_0, \hat{y} - E < y < \hat{y} + E$
Where the margin of error E is

$$E = t_{a/2} s_e \sqrt{\left(1 + \frac{1}{n} + \frac{n(x_o - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}\right)}$$ x_o represents the given value of x and $t_{a/2}$ has $n – 2$ df

CALCULATOR: Enter paired data in lists L1 and L2, STAT, TESTS, LinRegTTest.

Section 9 - 5: Multiple Regression

Multiple regression equation – expression of linear relationship between a dependent variable *y* and two or more independent variables (x_1, x_2, … x_k) **(p. 549)**

Adjusted coefficient of determination - the multiple coefficient of determination R^2 modified to account for the number of variables and the sample size calculated by *Formula 9-7* **(p. 552)**

$$Adjusted R^2 = 1 - \frac{(n-1)(1-R^2)}{[n-(k+1)]} \qquad \textit{Formula 9-7}$$

where *n = sample size and k* = number of independent (x) variables

Section 9 - 6: Modeling

CALCULATOR: 2ND CATALOG, choose DiagnosticOn, ENTER, ENTER, STAT, CALC, ENTER, enter L1, L2, ENTER

CHAPTER 10: MULTINOMIAL EXPERIMENTS AND CONTINGENCY TABLES

Section 10 - 2: Multinomial Experiments: Goodness-of-Fit

Multinomial experiment – an experiment that meets the following conditions:

1. The number of trials is fixed. **(p. 575)**
2. The trials are independent.
3. All outcomes of each trial must be classified into exactly one of several different categories.
4. The probabilities for the different categories remain constant for each trial.

Goodness-of-fit test – used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution **(p. 576)**

Test Statistic for Goodness-of-Fit Tests in Multinomial Experiments (p. 577)

$$X^2 = \Sigma \frac{(O-E)}{E}$$

where *O* represents the *observed frequency* of an outcome

Section 10 - 3: Contingency Tables: Independence and Homogeneity

Contingency table (or **two-way frequency table) –** a table in which frequencies correspond to two variables **(p. 589)**

Test of independence – tests the null hypothesis that the row variable and the column variable in a contingency table are not related **(p. 590)**

$$X^2 = \Sigma \frac{(O-E)}{E}$$

Critical values found in Table A-4 using **degrees of freedom = (*r* – 1) (*c* – 1)**

CALCULATOR: 2^{ND} X^{-1}, EDIT, ENTER, Enter MATRIX dimensions, STAT, TESTS, χ^2-Test, scroll down to Calculate, ENTER

CHAPTER 11: ANALYSIS OF VARIANCE

Section 11 - 1: Overview

Analysis of variance (ANOVA) – a method of testing the equality of three or more population means by analyzing sample variances **(p. 615)**

Section 11 - 2: One-Way ANOVA

Treatment (or **factor) –** a property, or characteristic, that allows us to distinguish the different populations from one another **(p. 618)**

Test Statistic for One-Way ANOVA (p. 620) $F = \dfrac{\text{var}\,iance between samples}{\text{var}\,iance within samples}$

Degrees of Freedom with k Samples of the Same Size n (p. 621)

numerator df = $k - 1$ denominator df = $k(n - 1)$

SS(total), or total sum of squares – a measure of the total variation (around x) in all of the sample data combined **(p. 622)** $SS(total) = \Sigma(x - \bar{\bar{x}})^2$ *Formula 11-1*

SS(treatment) – a measure of the variation between the sample means. **(p. 623)**

$$SS(treatment) = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2 = \Sigma n_i(\bar{x} - \bar{\bar{x}})^2$$ *Formula 11-3*

SS(error) – sum of squares representing the variability that is assumed to be common to all the populations being considered **(p. 623)**

$$\text{SS(error)} = (n_1 - 1)s^2_1 + (n_2 - 1)s^2_2 + \cdots + (n_k - 1)s^2_k = \Sigma\,(n_i - 1)s^2_i$$ *Formula 11-4*

MS(treatment) – a mean square for treatment **(p. 623)**

$$\text{MS(treatment)} = \frac{\text{SS(treatment)}}{k - 1}$$ *Formula 11-5*

MS(error) – mean square for error **(p. 624)**

$$\text{MS(error)} = \frac{\text{SS(total)}}{N - k}$$ *Formula 11-6*

MS(total) – mean square for the total variation **(p. 624)**

$$\text{MS(total)} = \frac{\text{SS(total)}}{N - 1}$$ *Formula 11-7*

Test Statistic for ANOVA with Unequal Sample Sizes (p. 624)

$$F = \frac{\text{MS(treatment)}}{\text{MS(error)}}$$ *Formula 11-8*

Has an F distribution (when the null hypothesis H_o is true) with degrees of freedom given by

numerator df = $k - 1$ denominator df = $N - k$

CALCULATOR: Enter data as lists in L1, L2, L3, STAT, TESTS, ANOVA, Enter the column labels (L1, L2, L3), ENTER

Section 11 - 3: Two-Way ANOVA

Interaction – between two factors exists if the effect of one of the factors changes for different categories of the other factor **(p. 632)**

CHAPTER 12: STATISTICAL PROCESS CONTROL

<u>**Section 12 - 2: Control Charts for Variation and Mean**</u>
Process data – data arranged according to some time sequence which are measurements of a characteristic of goods or services that results from some combination of equipment, people, materials, methods, and conditions **(p. 654)**

Run chart – sequential plot of *individual* data values with axis (usually vertical) used for data values, and the other axis (usually horizontal axis) used for the time sequence **(p. 655)**

Statically stable (or **within statistical control) –** a process is if it has only natural variation with no patterns, cycles or unusual points **(p. 656)**

Random variation – due to chance inherent in any process that is not capable of producing every good or service exactly the same way every time **(p. 658)**

Assignable variation – results from causes that can be identified (such factors as defective machinery, untrained employees, etc.) **(p. 658)**

CHAPTER 13: NONPARAMETRIC STATISTICS

<u>**Section 13 - 1: Overview**</u>
Parametric tests – require assumptions about the nature or shape of the populations involved **(p. 684)**

Nonparametric tests (or **distribution-free tests) –** don't require assumptions about the nature or shape of the populations involved **(p. 684)**

Rank – number assigned to an individual sample item according to its order in a sorted list, the 1st item is assigned rank of 1, the 2nd rank of 2 and so on **(p. 685)**

<u>**Section 13 - 2: Sign Test**</u>
Sign test – a nonparametric test that uses plus and minus signs to test different claims, including: **(p. 687)**

1. Claims involving matched pairs of sample data
2. Claims involved nominal data
3. Claims about the median of a single population

H_o: There is no difference
H_1: There is a difference.

Test Statistic for the Sign Test (p. 689)

For $n \le 25$: x (the number of times the less frequent sign occurs)

For $n > 25$: $$z = \frac{(x + 0.5)\frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

CALCULATOR: @nd, VARS, binomcdf, complete the entry of binomcdf(n,p,x) with *n* = total number of plus and minus signs, 0.5 for p, and *x* = the number of the less frequent sign, ENTER.

Section 13 - 3: Wilcoxon Signed-Ranks Test for Matched Pairs

Wilcoxon signed-ranks test - a nonparametric test uses ranks of sample data consisting of matched pairs **(p. 698)**

H_o: The two samples come from populations with the same distribution.

H_1: The two samples come from populations with different distributions.

Test Statistic for the Wilcoxon Signed-Ranks Test for Matched Pairs (p. 699)

For $n \le 30$: T For $n > 30$: $$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Where T = the smaller of the following two sums:

1. The sum of the absolute values of the negative ranks
2. The sum of the positive ranks

Section 13 - 4: Wilcoxon Rank-Sum Test for Two Independent Samples

Wilcoxon rank-sum test – a nonparametric test that uses ranks of sample data from two independent populations **(p. 703)**

H_o: The two samples come from populations with same distribution

H_1: The two samples come from populations with different distributions.

Test Statistic for the Wilcoxon Rank-Sum Test for 2 Independent Variables (p. 705)

$$z = \frac{R - \mu_R}{\sigma_R}, \quad \mu_2 = \frac{n_1(n_1 + n_2 + 1)}{2}, \quad \sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

n_1 = size of the sample from which the rank sum R is found

n_2 = size of the other sample R = sum of ranks of the sample with size n_1

Section 13 - 5: Kruskal-Wallis Test

Kruskal-Wallis Test (also called the ***H*** **test) –** nonparametric test using ranks of sample data from three or more independent populations to test **(p. 710)**

H_o: The samples come from populations with the same distribution.

H_1: The two samples come from populations with different distributions.

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k}\right)$$

Section 13 - 6: Rank Correlation

Rank correlation test (or **Spearman's rank correlation test) –** nonparametric test that uses ranks of sample data consisting of matched pairs to test **(p.719)**

H_o: $p_s = 0$ (There is *no* correlation between the two variables.)

H_1: $p_s \neq 0$ (There is a correlation between the two variables.)

Test Statistic for the Rank Correlation Coefficient (p. 720)

$$r_s = 1 - 6\Sigma d^2 / n(n^2 - 1)$$

where each value of d is a difference between the ranks for a pair of sample data.

1. $n \le 30$: critical values are found in Table A-9.
2. $n > 30$: critical values of r_s are found by using $t_s = \frac{\pm z}{\sqrt{n-1}}$ *Formula 13-1*

CALCULATOR: Enter data in L1 and L2, STAT, TESTS, LinRegTTest

Section 13 - 7: Runs Test for Randomness

Run – a sequence of data having the same characteristic; the sequence is preceded and followed by data with a different characteristic or by no data at all **(p. 729)**

Runs test – uses the number of runs in a sequence of sample data to test for randomness in the order of the data **(p. 729)**

5% Cutoff Criterion (p. 731)

Reject randomness if the number runs *G* is so small or so large i.e.

1. Less than or equal to the smaller entry in Table A-10
2. Or greater than or equal to the larger entry in Table A-10.
3.

Test Statistic for the Runs Test for Randomness (p. 733)

If $\alpha = 0.05$ and $n_1 \leq 20$ and $n_2 \leq 20$, the test statistic is G

If $\alpha \neq 0.05$ or $n_1 > 20$ or $n_2 > 20$, the test statistic is

$$Z = \frac{G - \mu_G}{\sigma_G}$$

Where $\mu_G = \dfrac{2n_1n_2}{n_1 + n_2} + 1$ *Formula 13-2*

Where $\sigma_G = \sqrt{\dfrac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1n_2)^2(n_1n_2 - 1)}}$ *Formula 13-3*

ROUND OFF RULES

- **Simple rule** – Carry one more decimal place than ;is present in the original set of values, **(p. 60)**
- **Rounding off probabilities** – either give the *exact* fraction or decimal or round off final decimal results to 3 significant digits. **(p. 120)**
- **For** $\mu, \sigma^2, and \sigma$ - round results by carrying one more decimal place than the number of decimal places used for random variable *x*. If the values of *x* are integers, round $\mu, \sigma^2, and \sigma$ to one decimal place. **(p. 186)**
- **Confidence intervals used to estimate μ (p. 304)**
  1. When using the *original set of data* to construct a confidence interval, round the confidence interval limits to one more decimal place than is used for the original set of data.
  2. When the original set of data is unknown and only the *summary statistics* $(n, \bar{x}, s)$ are used, round the confidence interval limits to the same number of decimal places used for the sample mean.
- **For sample size *n*** – if the used of Formula 6-3 does not result in a whole number, always *increase* the value of *n* to the next *larger* whole number. **(p. 324)**
- **Confidence interval estimates of *p* –** Round to 3 significant digits. **(p. 332)**
- **Determining sample size –** If the computed sample size is not a whole number, round it up to the next *higher* whole number. **(p. 334)**
- **Linear correlation coefficient –** round *r* to 3 decimal places. **(p. 510)**
- ***Y*-intercept b_o and Slope b_1 -** try to round each of these to 3 significant digits. **(p. 527)**