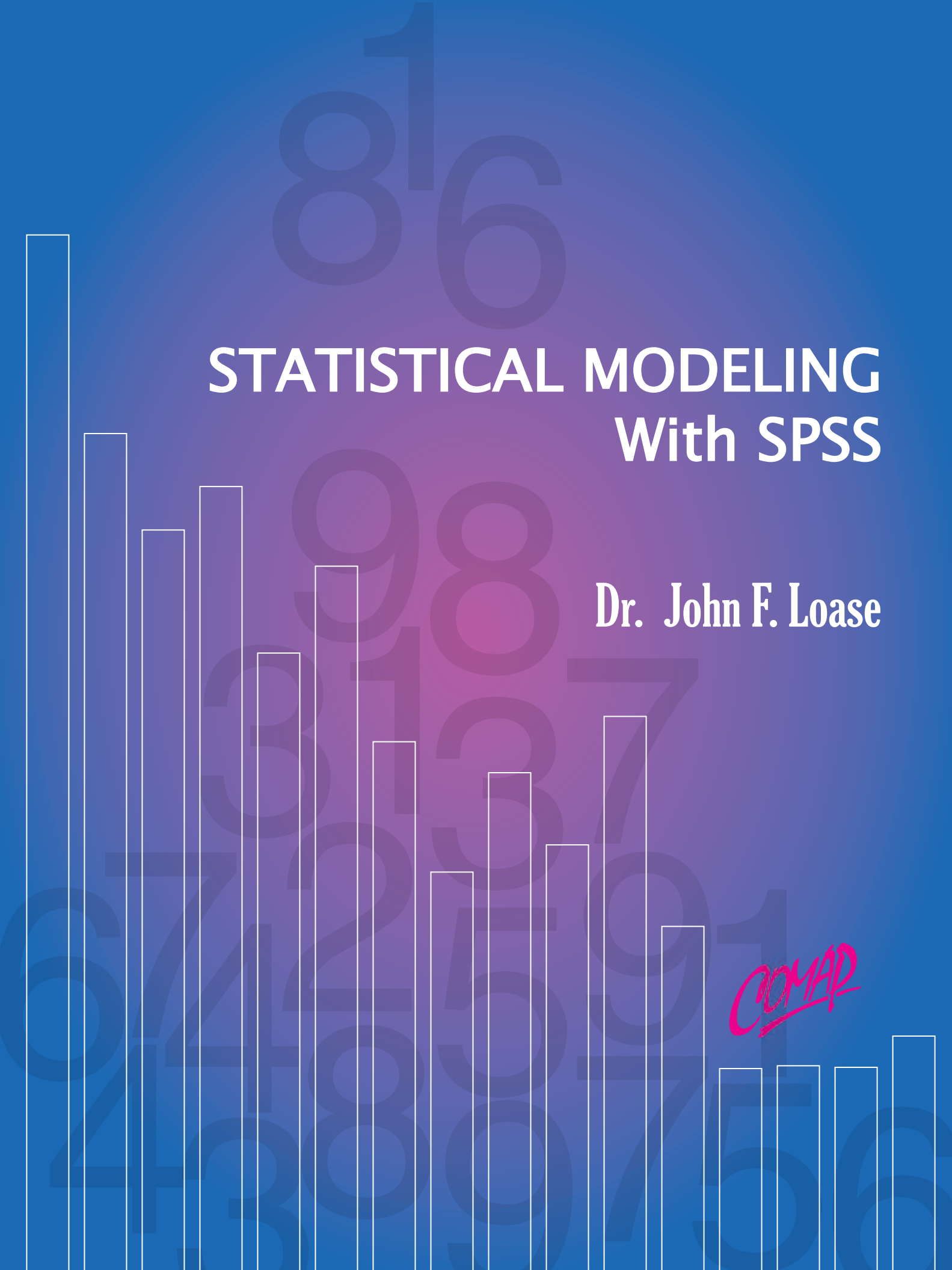


STATISTICAL MODELING With SPSS

Dr. John F. Loase

COMAP

STATISTICAL MODELING
With SPSS

Dr. John F. Loase

COMAP

STATISTICAL MODELING WITH SPSS

by Dr. John F. Loase
Concordia College, New York

Published and distributed by

COMAP, Inc.
The Consortium for Mathematics and Its Applications
170 Middlesex Turnpike, Suite 3B
Bedford, MA 01730

Support for this book was provided by:

1) This book was prepared with the support of National Science Foundation Grant # 9155966 - Mathematics Modeling. However, any opinions, findings, conclusions, and / or recemmendations herein are those of the authors and do not necessarily reflect the views of the NSF.

2) Concordia College - New York Sabbatical 2012

ISBN 1 933223 78 2

DEDICATED

TO

1) Dr. Ben Fusaro, who founded the International Contest in Mathematical Modeling and led the Advisory Council developing this book.

and

2) My late friend Burdette Graham, who lost his life in the Vietnam War.

May Mathematical Modeling unite the world's people in concern, cooperation and connection.

PREFACE

PHILOSOPHY

Statistical Modeling with SPSS is the result of over twenty years of teaching Elementary and Intermediate Statistics on the undergraduate level and Advanced Statistics and Mathematical Modeling at the graduate level.

This text has been used to prepare students for the International Contest in Mathematical Modeling and for mini-courses for college and university faculty interested in innovating mathematical modeling programs.

Statistical Modeling with SPSS was sponsored by the National Science Foundation. A distinguished advisory council and team of editors assisted with concepts and editorial suggestions throughout the book's development. They will be acknowledged at the conclusion of the preface.

PREREQUISITES

Statistical Modeling with SPSS is written as a senior level/graduate level text for mathematics, statistics, computer science or engineering majors. It reviews elementary statistics in Chapter One. The rest of the text assumes that the student has completed three semesters of Calculus, Calculus-Based Probability and Statistics, and at least one course in computer programming.

The text has been used to train students for the International Contest in Mathematical Modeling. In its early development, this book was focused on graduate level mathematical

modeling (with a statistical focus) and for advanced mathematics students preparing for the contest in modeling.

TECHNOLOGY

Statistical Modeling with SPSS makes extensive use of SPSS to test student initiated hypotheses from a set of real data included with the test. The data set is the result of coding the 104 responses (variables) of 542 undergraduates at Concordia College - NY and Iona College to the Marketing and Sigfluence Survey, included in Appendix A.

For students who need more extensive review of elementary statistics, an extensive TI-83 based Primer is included in Appendix B.

ORGANIZATION

Chapter One - Selected Topics from Elementary Statistics

A review of hypothesis testing, confidence intervals, correlation, and single variable and multiple regression analysis. An extensive review of these topics is included in Appendix B, for the interested student, geared to the TI-83 calculator.

At the conclusion of Chapter One, the student can immediately test hypotheses and perform multiple regression analyses with the enclosed set of data. Step by step instructions on the proper use of SPSS for testing hypotheses and performing regression analysis are featured in the book.

Chapter Two - Selected Topics from Calculus-Based Statistics and Probability

A review of the essential topics from Calculus Based Probability and Statistics that form the foundation of Statistical Modeling.

Chapter Three - Input Probability Distributions

Goodness of fit tests using the Poisson, normal, uniform, and exponential density functions.

The chapter concludes with SPSS exercises to test the included data set for exponential, normal, and uniform density functions.

Chapter Four - Random Number Generators

Linear congruential number theory and current research in irrational numbers as sources of random numbers.

Chapter Five - Generating Random Variables

The inverse transform method with discrete and continuous modeling examples.

A current research approach, validating multiple regression results with a statistical model, is presented together with myriad research possibilities for the student in Appendix F.

Chapter Six – Application from Linear Algebra

An applied problem from Linear Algebra solved using elementary matrix operations.

Chapter Seven - Two Modeling Exercises

Chapter Eight – Two Problems and Oustanding Solutions from the International Contest in Mathematical Modeling

Two outstanding papers are reprinted with permission of COMAP.

SPSS

One of the principal features of this book is the opportunity for students to use SPSS to analyze a 50 variable by 542 row (respondent matrix). The student should take the Marketing and Sigfluence Survey early in the course and then explore new insights into our college

students' beliefs about money and meaning. It took two years for my two graduate students, Teresa Osadnik and Grace Dickson, to enter the 104 responses for each of the 542 undergraduates who completed the survey. For the next year we deleted variables as a result of data mining and correlational methods and arrived at the 50 variable data set.

The 50 variables were all mapped to the interval (0, 1) to further explore graphical and subtle relationships. A 50 variable set has virtually unlimited potential for statistical insights. For example, there are 50 C 5 = 2,118,760 combinations of five variables we could isolate for multiple regression.

SIGFLUENCE

My doctorate was the first awarded in Mathematics (emphasis Statistics) and Psychology (Measurement, Research and Evaluation in Psychology and Education) from Columbia University Teachers College. In 1984 I invented the new word "sigfluence" to define significant, long-term, positive influence. My 8th book, *The Sigfluence Generation: Our Young People's Potential to Transform America*, is free for you to download from my website sigfluence.com. It took over 20 years of Statistical Modeling to discover that our 18-25 year olds reported dramatically high potential and need to effect sigfluence. As the book develops, you can use the data set to discover "Golden Nuggets" of significant relationships. For example, it took 18 months for my wonderful graduate students Teresa Osadnik and Grace Dickson to enter the data. Then in one afternoon, I was able to peruse 10,000 correlations that over time led to the exciting discovery that our young people can positively transform the world if we Baby Boomers serve as mentors and guides.

Statistical Modeling is foundational to recognizing and remedying real world problems. Without Statistical Modeling we rely on appearance and convenience, forever spinning our wheels in futile attempts at making the world a better place.

ACKNOWLEDGMENTS

Thanks are due Dr. Henry Ricardo (CUNY) and Professor Rowan Lindley (SUNY), who edited the test and to Professor Joyce McQuade, who completed the Solution Set. Professor Louis Rotando (SUNY) served as my mentor, department chair, and valued colleague for eighteen years.

I am especially grateful to Dr. Catherine Ricardo (Chair - Graduate Computer Science at Iona College), who offered me my first course in graduate Mathematical Modeling in 1988. We are very thankful for the leadership furthering mathematical modeling and the award of our National Science Foundation grant (1992-1996). Our distinguished advisory council provided encouragement, invaluable suggestions, and were partners in our mini-courses and lectures, which were outgrowths of the NSF grant.

Also, special recognition and deep gratitude is due Mrs. Barbara Boyce for her painstaking attention to detail and consistent loyalty for three decades of typing and editing of this, our eleventh book.

Dr. Catherine Ricardo - Iona College

Dr. Henry Ricardo - CUNY - Medgar Evars

TABLE OF CONTENTS

OVERVIEW

The reader has now reviewed elementary statistics. If more basic review of elementary statistics is useful (with specific instructions on use of the TI-83 calculator), the reader should proceed to Appendix B - a review of the essentials of a first course in statistics, TI-83 calculator based.

If the student has reviewed this material, he/she should next use SPSS to analyze the enclosed data set and perform correlational/regression analyses. SPSS (15.0) is recommended for purchase to accompany this text.

The first recommended activity is described in detail in Appendix C and uses SPSS to analyze original and current data included with your book. You should have a data set with 542 rows and 50 variables (columns). If you have reviewed your elementary statistics and understand the basics of multiple regression, proceed to Appendix C.

CHAPTER ONE

SELECTED TOPICS FROM PROBABILITY AND STATISTICS

1.1 Population and Sample

In statistics, we try to make inferences or judgments about an entire ***population*** based upon a part of the population we can observe and collect information from, called the ***sample***. For example, if we could call every registered voter in the country prior to the upcoming presidential election and ask whom the respondent was voting for, we could get a near perfect prediction of who our next President would be.

However, the phone bill would be too expensive and the number of callers too extensive, so we rely upon statistics to predict the winner. We ***sample*** a relatively small number of people who are representative of the total population. We make sure that every special characteristic like region, gender, and socioeconomic group is proportionately represented, and the results of our study will never be certain. So we indicate the confidence that we have in our statistics.

For the purpose of this text, we will use the standard sample and population definitions which include:

1) Random Variable

A random variable is a function which assigns a number (probability) to each number in a sample space; e.g., in the binomial event of tossing a coin, the probability of three consecutive heads - P (3) = 1/8. The random variable or function is the binomial formula,

$$P(X) = \binom{n}{x} P^x (1 - P)^{n-x}$$

2) Sample Mean

Let $x_1, x_2, ..., x_n$ represent the random variables for a sample of size n. The mean of the sample is given by:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

If $y_1, y_2, ... y_n$ are values obtained in a particular sample, then

$$\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n}$$

3) Sample Variance

If we have a sample of n observations $x_1 + x_2 + ... + x_n$ with mean x, we define the sample variance as

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

This is slightly different from the definition of population variance which would equal

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

However, dividing by (n - 1) results in an unbiased estimate of σ^2, the population variance. We will show that in a later section. The standard deviation of a population, σ, is the square root of the variance: $\sigma = \sqrt{\text{variance}}$. For example, if the variance = 100, the standard deviation $= \sqrt{100} = 10$. The sample standard deviation is $\sqrt{s^2}$, written as s.

EXERCISES 1.1:

1. Find the sample mean and standard deviation for a set of customer arrival times that follow, where 8 AM = 0, 4 PM = 8, noon = 4, etc.

   8 A.M., 8:30 A.M., 9 A.M., 10 A.M., 11 A.M., 2 P.M.,

   3 P.M., 3:30 P.M., 4 P.M.

2. Show the equivalence of the two formulas for variance.

$$S^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} = \frac{n \Sigma (x^2) - (\Sigma x)^2}{n(n - 1)}$$

3. How would you obtain a large representative sample of American high school students to provide norms (scores for comparison) for Sternberg's (Yale University) innovative multi-factor intelligence test?

4. Why are there two formulas for variance?

1.2 Confidence Intervals for Means

We rarely, if ever, know the mean of the population. If we did, we wouldn't need statistics to estimate it.

Confidence intervals allow us to form an interval based upon our sample mean and the level of confidence that we select. Putting these two together results in an interval that contains the population mean with a corresponding probability or confidence level.

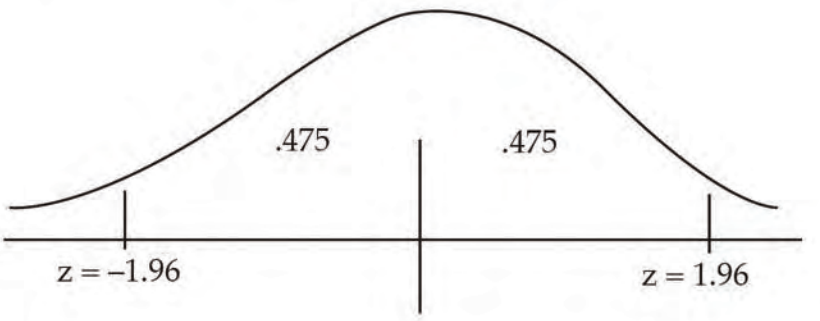
For example, your annual salary after completing college will be between $15,000 and $150,000 with 99% confidence. Perchance, a few of you will live off the land in the Adirondacks and require less than $15,000 for sustenance; some of you may have a Rockefeller for a relative and could expect to start higher than $150,000.

Two formulas that are used to compute confidence intervals for means are:

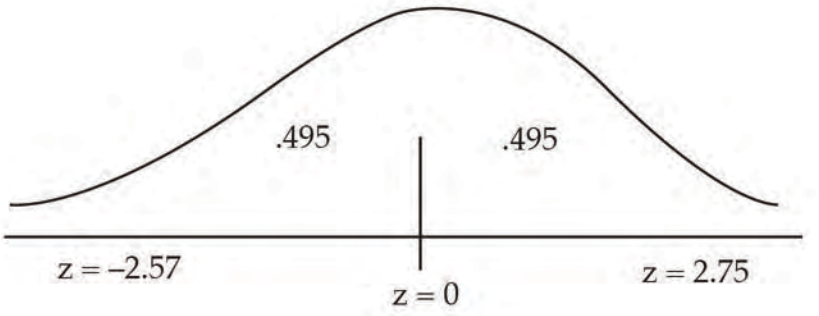
1. Large samples ($n \geq 30$)

$$\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}}$$

z_c refers to the z value that you obtain from the table of standard normal z distribution values. For a 95% confidence interval, you would use the z value of 1.96 since the probability value of .475 corresponds to this number. See the graph below.



The two probabilities of .475 together yield an interval with 95% confidence. For 99% confidence, you use a z value of 2.57. See the graph below.



The symmetry of the standard normal curve allows us to add the two probabilities to obtain our 99% confidence intervals by using $z_c = 2.57$.

We could write this as $z_{.495} = 2.57$. Sometimes it is written as $z_{.005} = 2.57$, where .005 refers to the probability in the extreme right tail.

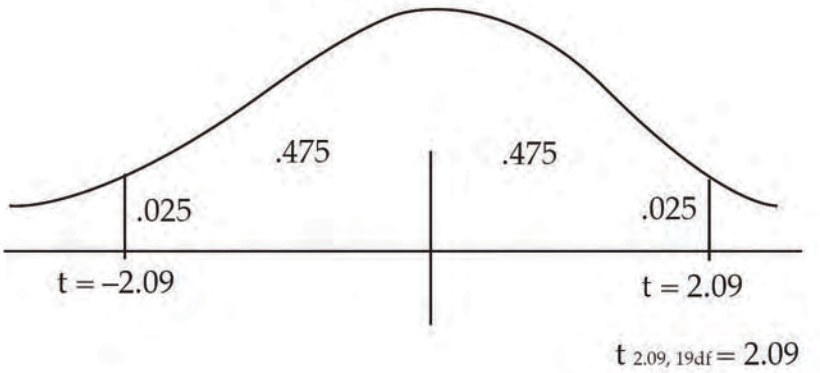
2. Small samples ($n < 30$)

$$x \pm t_c \cdot \frac{s}{\sqrt{n}} \quad \text{use } (n - 1) \text{ df} \quad (\text{df refers to degrees of freedom})$$

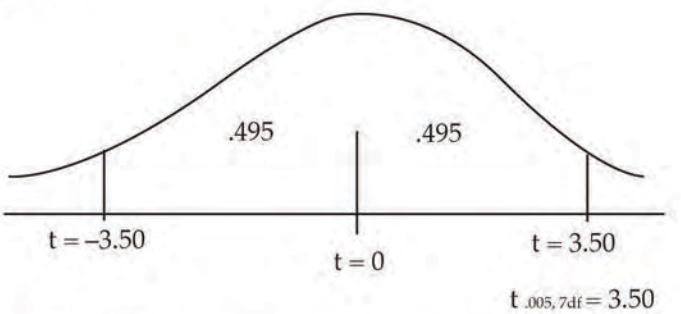
t_c represents the t value obtained from the table of student's t distribution. The t distribution is used with small samples. When the number in the sample is 30 or greater, the z and t values are equal.

The table requires the user to calculate the degrees of freedom for the individual problem. For the small sample, confidence interval df = (n - 1), where n refers to sample size.

To illustrate, if a sample of 20 were used, df = (n - 1) = 20 – 1 = 19. To determine t_c for a 95% confidence interval with 19 df, look up $t_{.025,19df}$. The value .025 is the probability remaining in the extreme left and right hand tails of the confidence interval. Consider the graph below.



As another example, consider a sample size of 8 and a specified 99% confidence interval. Look up the table value of t corresponding to df = 8 – 1 = 7 and probability of .005 in the extreme left and right tail. This will yield probabilities of .495 + .495 = 99% confidence. Consider the graph below.



Frequently the t table gives values corresponding to the extreme tail of .005 rather than the .495 value which is half of the desired confidence level. Be careful when you use the table as to which probability measure you are using. Even better, read the section on α level or Type I error carefully so that you understand the notation. For example, 95% confidence

yields a Type I error or α value of 5%. Therefore, $\frac{(1 - \alpha)}{2} = .475$ and this

could be the way that your table reports t values. Shortly we will discuss the meaning of α or Type I error.

Example: Compute a 99% confidence interval corresponding to the mean IQ of readers of this book, if a sample of 50 yielded an average IQ of 120 with standard deviation 10. [$z_{.495} = 2.57$].

$$99\% \text{ C.I.} = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

$$99\% \text{ C.I.} = 120 \pm 2.57 \cdot \frac{*10}{\sqrt{50}}$$

*We replace σ by s - the sample standard deviation

$$99\% \text{ C.I.} = (116.37,\ 123.63)$$

We can say with 99% confidence that the average IQ of the readers of this book is between 116.37 and 123.63. This makes you a superior group of young researchers.

EXERCISE: An efficiency expert visits a local bank and observes the average waiting time in line of 25 customers. She finds that the mean is 5 minutes with variance of 3 minutes. Compute the 95% confidence interval for average waiting time at the local bank. [$t_{.025,24df} = 2.06$].

$$95\% \text{ C.I.} = x \pm t_{.025,24df} \frac{s}{\sqrt{n}}$$

$$95\% \text{ C.I.} = 5 \pm 2.06 \frac{\sqrt{3}}{\sqrt{25}} \qquad [\text{Note } s = \sqrt{3}]$$

$$95\% \text{ C.I.} = [4.29,\ 5.71]$$

The efficiency expert can conclude with 95% confidence that customers wait an average of 4.29 to 5.71 minutes on line at the local bank.

EXERCISES 1.2:

1. The average weight of 100 male college freshmen at a certain event is 165 lbs. with standard deviation 16. Estimate the average weight of male college freshmen with 99% confidence.

2. Consider the table below of times when customers enter a service station for full service.

7:30	8:31	10:10	11:06
7:30	8:45	10:28	11:09
7:35	9:11	10:43	11:41
8:05	9:36	10:51	11:53

Let 7:30 = 0 12 noon = 4.5

a) Find the mean of the above data.

b) Find the standard deviation for the above data.

c) Compute the 95% confidence interval for the mean of the population.

3. What would be some shortcomings of the results of the efficiency expert in relation to the time of her visit and hours that she actually spent observing in the bank?

4. Show how formula (1) can be derived from the random variable $z = \frac{x - u}{\sigma / \sqrt{n}}$

5. For airplane takeoffs, the following times in queue were observed:

2.2	4.7	5.6
3.2	3.6	4.8
5.4	2.8	10.6
5.8	6.0	4.8
7.6	6.6	9.6

Find a 95% confidence interval for population waiting times at the airport. The times are reported in minutes.

*6. If $x_1, x_2, \ldots x_n$ are dependent random variables, how would this affect formula (1)?

*indicates research level problems

1.3 Confidence Intervals for Difference of Population Means

Testing whether one procedure is more effective than another leads us to subtract the means of two samples and employ standard tests. For example, to determine whether a drug to combat AIDS is effective, we could subtract the average longevity of groups taking a placebo from the average longevity of the group taking the new drug. This leads to confidence interval for differences of population means which is given by:

$$[n_1 \geq 30, \quad n_2 \geq 30]$$

$$\overline{x_1} - \overline{x_2} \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where x_1, x_2, σ_1, σ_2, n_1, n_2 are the means, standard deviations and sizes of the two samples taken from the population.

Example 1: Find a 95% confidence interval for the difference of mean waiting time before and after a new queuing system was set up in the bank, if the sample statistics are:

$$x_1 = 5.6$$

$$x_2 = 4.5$$

$$s_1^2 = 1.4$$

$$s_2^2 = 2.1$$

$$n_1 = 100$$

$$n_2 = 225$$

The 95% confidence interval is

$$\overline{x_1} - \overline{x_2} \pm z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$z_{.475} = 1.96$

We replace σ_1 by s_1 and replace σ_2 by s_2. This gives us a 95% confidence interval which is calculated as follows:

$$95\%\ \text{C.I.} = 5.6 - 4.5 \pm 1.96 \sqrt{\frac{1.4}{100} + \frac{2.1}{225}}$$

$$95\%\ \text{C.I.} = 1.1 \pm .30 = (.80,\ 1.41)$$

We leave to the next section whether this innovation was statistically significant.

Confidence Interval for Small Sample Difference of Population Means

If either or both sample(s) are fewer than 30, adjust the previous confidence interval formula to the following:

$$\overline{x_1} - \overline{x_2} \pm \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where t is found from a standard t table with $n_1 + n_2 - 2$ degrees of freedom.

$$S_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Example 2: If the waiting times at a bank for two different tellers are studied with the following results, find a 90% confidence interval for the difference in mean waiting time between the two tellers.

$\overline{x_1} = 3.2$

$\overline{x_2} = 1.9$

$s_1^2 = .64$

$s_2^2 = .81$

$n_1 = 36$

$n_2 = 24$

$$S_p^2 = \frac{35\ (.64) + 23\ (.81)}{36 + 24 - 2} = .7074$$

$$90\%\ \text{C.I.} = 3.2 - 1.9 \pm 1.66\ \sqrt{.7074\ (1/36 + 1/24)}$$

$$90\%\ \text{C.I.} = 1.3 \pm 1.66\ (.2216)$$

$$90\%\ \text{C.I.} = 1.3 \pm .37 = (.93,\ 1.67)$$

EXERCISES 1.3

1. Find a 99% confidence interval for the difference of means for the data from Example 1.

2. Find a 95% confidence interval for the difference of means for the statistics taken from Example 2.

3. The following data represents the output of simulated rejection rates for a part that is inspected by way of (A) a conveyor belt in a factory or (B) by an alternate procedure. Each output represents a percentage of rejected parts for a simulation of 1000 inspected parts. Find a 95% confidence interval for the overall difference of means in the percentage of rejected parts.

   A - .175, .216, .165, .129, .163, .312, .326, and .096

   B - .136, .123, .121, .321, .236, .315, and .102

1.4 Confidence Intervals for Proportions

Suppose our statistic of interest is the proportion p of "success" in a population if a sample size $n \geq 30$ is drawn from a binomial population. We let p represent the percentage of success that we calculate from our sample.

The 99% confidence limits for p - the population proportion is:

$$\bar{p} \pm z_{.495}\sqrt{\frac{p(1-p)}{n}} \qquad \bar{p} = \frac{\text{the number of successes}}{\text{total number of trials}}$$

Consider the following example. We traveled to Maine to survey 100 voters over our handling of foreign policy in the Arab states. Suppose we find that 64 agree and 36 disagree. Provided that our sample is independent, random, and representative of Maine, the estimate of the proportion of Maine voters favoring our policy is:

$$\bar{p} \pm z_{.495}\sqrt{\frac{p(1-p)}{n}} \qquad (z_{.495} = 2.57)$$

$$= .64 \pm 2.57\sqrt{\frac{.64(.36)}{100}} = (.52, .76)$$

EXERCISES 1.4

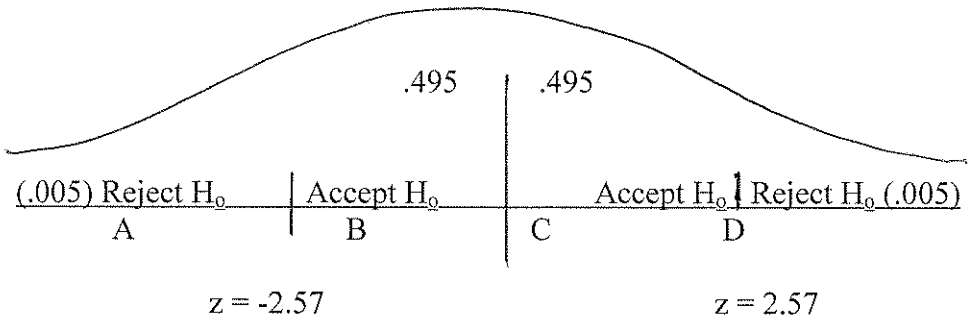
1. Find a 90% confidence interval for the proportion of Maine voters against our foreign policy.
2. Find a 95% confidence interval for the proportion of heads obtained in flipping a fair coin 1000 times.
3. Find the probability that in 200 tosses of a fair coin, the proportion of heads will be between 45 and 65%.
4. In two computer simulations of factory inspection of parts, 15% of 1000 were rejected in the first experiment and 200 of 2000 were rejected in the second. Assuming independence and binomially distributed data, find a 95% confidence interval for parts rejected at the factory based on the sample data. [Pool both samples into one sample to compute the appropriate statistics.]

1.5 Tests of Hypothesis

A. Type I and Type II Errors

When we test a hypothesis, we always run the risk of error with our conclusion. If we reject a hypothesis that should be accepted, this is a Type I error. Or, if we accept a hypothesis that should be rejected, we have made a Type II error.

Our level of significance, traditionally $\alpha = .05$ or $.01$, is the level of Type I error we select before the sample is taken. If we arbitrarily choose $\alpha = .01$ in our experiment, this means that we will reject a true hypothesis (Type I error) only 1% of the time.



For the accept-reject graph given above, the two-tailed test means that for a z value greater than 2.57 or less than -2.57, we reject our null hypothesis, H_o, (usually $\mu_1 = \mu_2$). Here μ_1 represents the first population mean while μ_2 represents the second population mean. We rarely, if ever, have all the data from an entire population, so we use the sample data to determine $\bar{x}_1$, the first sample mean, and $\bar{x}_2$, the second sample mean. From this, we make inferences about the population means μ_1 and μ_2.

For any z value, $-2.57 < z < 2.57$, we accept our null hypothesis. The probabilities .005, .495, .495, and .005 represent the likelihood of our computed z value falling in each of the four regions A, B, C and D.

EXERCISES 1.5

1. For setting up an experiment testing the effectiveness of a new AIDS drug, why does the researcher avoid a level of Type I error = .00001?

2. If we develop a confidence interval for mean waiting time of post office customers, which is $\overline{x} \pm 2.57 \frac{2}{\sqrt{n}}$, what is the chance that the population mean is outside this interval? How does this relate to Type I error?

3. Suppose we decided to test whether the slot machines in a gambling casino were fair. You observe that the casino advertises that someone wins every twelve trials. In an evening you record that thirty times out of one thousand tries someone has won on the slot machine. What level of Type I error would you select for your study? Would you be satisfied if $\alpha = .3$? Would .00001 be appropriate for α?

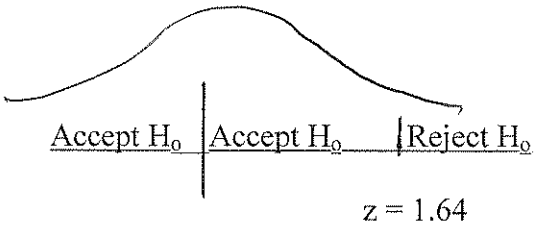
1.6 Populations - One Sample Difference of Means

One-Tailed/Two-Tailed Tests

In comparing two results, if we are testing whether one result is higher or lower than another, we use a one-tailed test. We select a z or t value that puts all the Type I error in the critical region to the extreme left or right of the distribution.

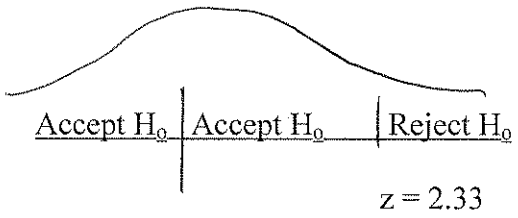
For example, if we are testing whether students' reading scores have dropped over a ten year period, we use a one-tailed test. The graphs are below:

If $\alpha = .05$



Rej H_o if $z > 1.64$ Critical z

If $\alpha = .01$

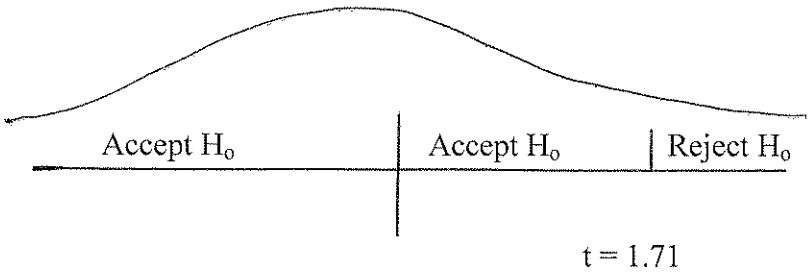


Rej H_o if $z > 2.33$ Critical z

Small Sample t

Since the t value changes with n, let n = 26. Look up $t_{.05,(n-1)df}$ or $t_{.05,25df} = 1.71$.

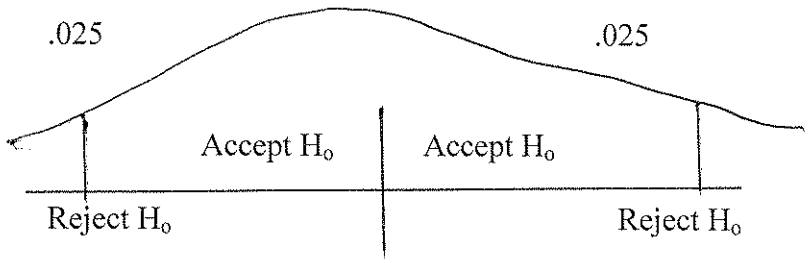
$\alpha = .05$



Reject H_o if $t > 1.71$

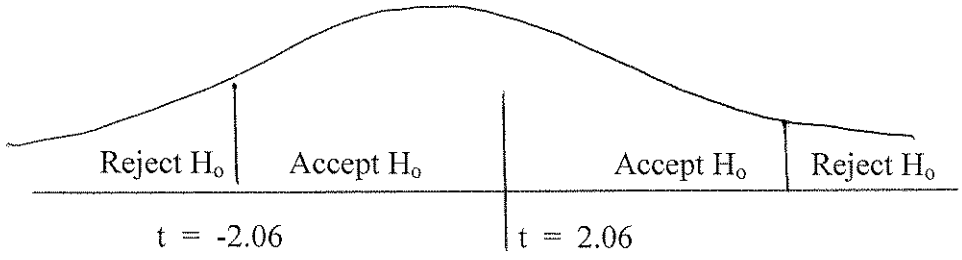
Two-Tailed Tests

If we compare whether two results are different, then we have two possibilities, a significant increase or a significant decrease, we must split the critical region as follows:



Accept H_o if $-1.96 < z < 1.96$

For $\alpha = .05$, $n = 26$, a two-tailed t test would have a graph as follows:



Accept H_o if $-2.06 < t < 2.06$

Reject H_o if $t < -2.06$ or $z > 2.06$

Large Sample ($n \geq 30$)

If you are fortunate enough to know the mean of the entire population and wish to compare your sample mean to the population mean, use the formula:

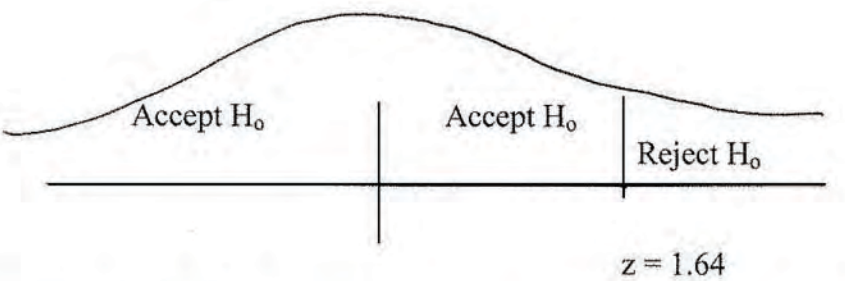
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

[μ refers to the population mean, a rarely known value.]

If necessary, use the sample standard deviation s to estimate σ.

Example 1: Women in 1990 have a life expectancy of 75.5 with a historical standard deviation of 16.2. Has their life expectancy significantly increased over the 73.2 figure reported in 1980? [n = 1000 for the 1990 sample].

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{75.5 - 73.2}{16.2/\sqrt{1000}} = 4.49$$



Since you have asked whether life expectancy has increased, use a one-tailed test where critical $z = 1.64$ for $\alpha = .05$.

We conclude that women have a significant increase in life expectancy over the five years 1985-1990.

Small Sample ($n < 30$)

If our population is known and we wish to repeat the previous procedure with a sample less than 30, we use the formula

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

and utilize the appropriate value from a t table with (n - 1) degrees of freedom.

Example 2: A simulation bank advertises an average wait in line for a teller of two minutes. An efficiency expert visits the simulation bank and notes an average wait of 2.4 minutes with $s = .6$. Can we conclude that this wait is different from the simulation bank's claim? [$n = 25$, $\alpha = .05$]

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad t = \frac{2.4 - 2}{.6\sqrt{25}} = 3.3$$

The key word "different" calls for a two-tailed test. Critical $t_{.025,24df} = 2.06$. Reject H_o because 2.4 minutes is significantly greater than the simulation bank's claim.

EXERCISES 1.6

1. Redo example 1, if we change the question to read: Has there been a significant change in women's life expectancy?

2. Work out example 2, if the average wait is found to be 1.8 minutes.

3. For the following sample of gas station prices, test whether unleaded gasoline has significantly increased from \$1.39 per gallon. [Let $\alpha = .05$]

1.41	1.79
1.36	1.96
1.58	1.56
1.29	1.41
1.39	1.82

4. A mathematical model of the stock market forecast the following percentage increases in blue chip stocks. Is the forecast the same as the historical average of 5.1%? Let $\alpha = .05$

4.6	3.9	4.2	6.4	6.3	1.8
3.9	8.2	3.8	8.2	2.9	6.4
5.3	6.4	5.6	5.6	4.7	2.6
6.2	5.8	6.2	6.4	6.2	7.1
5.6	3.9	4.8	8.2	8.7	2.3

1.7 Two Sample Difference of Means

Normally (no pun intended), you don't have knowledge of an entire population. As a result, we frequently test the mean of one sample against the mean of another. For example, suppose we are testing the effectiveness of a bank teller training program.

We could obtain a sample of waiting times from a bank without the procedure and a separate sample of waiting times from a bank with the new procedure. Depending upon the size of the samples, use either the large sample z test or small sample t test. We choose our critical value and use either a one or two tailed test depending on whether we are testing > or < (one tailed) or different (two tailed).

1) Large sample difference of means

$n_1 \geq 30$ $\qquad$ $n_2 \geq 30$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

2) Small sample difference of means

$n_1 < 30$ $\qquad$ or $\qquad$ $n_2 < 30$

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

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

$$df = n_1 + n_2 - 2$$

Example 1: Suppose average waiting time at our bank with the new teller training program is 2.6 minutes compared to 2.9 minutes in a bank with no such procedure. If $n_1 = n_2 = 100$ and $s = .8$, does the new program significantly lower waiting time? Let μ_1 = mean time with no training. Let μ_2 = mean time with training.

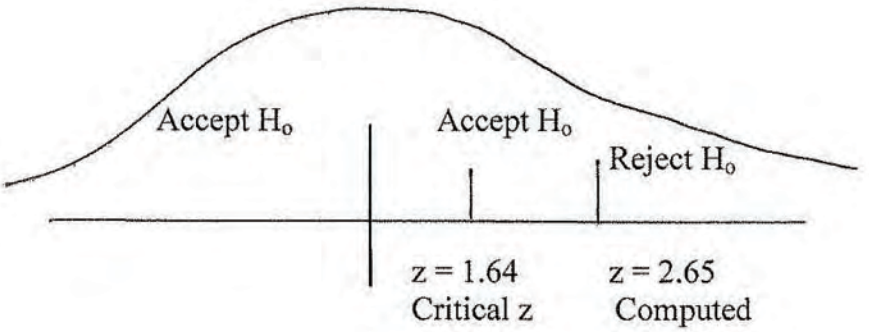
H_o: $\mu_1 = \mu_2$

H_a: $\mu_2 < \mu_1$ or $\mu_1 > \mu_2$

Use the following formula with $\overline{x_1}$ = sample mean waiting time with no training, x_2 = sample mean waiting time with the new procedure.

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{2.9 - 2.6}{\sqrt{\frac{(.8)^2}{100} + \frac{(.8)^2}{100}}} = 2.65$$

Let $\alpha = .05$



Conclusion - Reject H_o

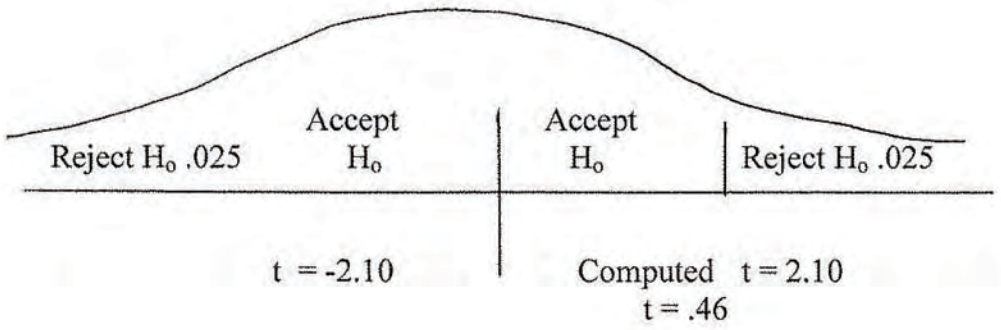
The new program has evidenced significantly lower waiting time. Since $\alpha = .05$, you can be 95% confident that the new procedure significantly reduces waiting time.

Example 2: Suppose a 10 day model of the stock market yields a 6.2% annual increase in prices with $s_1 = .015$ and we are interested in comparing our model with the actual 10 day figures which are 5.9% with $s_2 = .014$. Is the model significantly different from the stock market results?

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.062 - .059}{\sqrt{.0002105\ (1/10 + 1/10)}}$$

$$t = .46$$

Let $\alpha = .05$



The computed t is in the accept H_o region. We conclude that there is no significant difference between the model and the stock market results.

EXERCISES 1.7

1. Redo example 1, if the new teller training program resulted in 2.8 minutes average waiting time with all other statistics remaining the same.

2. Redo example 2, if the actual stock market average was 6.9%, all other statistics remaining the same.

3. For a new AIDS drug, the following longevity figures were recorded:

	Mean Longevity	s^2	n
New (population 1)	9.2	2.5	100
Placebo (population 2)	7.1	2.7	100

Can we conclude that those patients who took the new drug significantly increased their longevity?

1.8 Sample Size

In order to test hypotheses, we need to select a random, independent, and significantly large sample. We would not go to the University of California at Berkeley or Oshkosh Community College as our only place to sample college student opinion. We also want samples that are independent. The researcher would not ask for names of a respondent's relatives or friends to add to the sample since this would likely lead to related or dependent responses. For further analysis of this deep and difficult statistical problem, please refer to my article, "Dependent Random Variables and Hypothesis Testing," International Journal of Mathematical and Computer Modeling, Fall 1990.

Selecting appropriate sample size is critical (again, no pun intended) to carrying out a sound statistical study. If we wanted results that were close to certainty, we would survey the entire population. However, we rarely, if ever, have the financial or person power resources to survey the entire population. For example, by the time we surveyed the American population about which presidential candidate they favor, the President's four year term would have expired. Also, the financial costs would be staggering.

However, there is an elegant relationship between our previous discussion of confidence interval and sample size. To illustrate, consider the formula for a confidence interval that we have previously discussed:

Confidence Interval for Population Proportion

$$\bar{p} \pm z \sqrt{\frac{p(1-p)}{n}}$$

Let the maximum value for the error (at our arbitrarily selected level of confidence) be e.

$$e = z \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}}$$

Solve for n

$$e^2 = z^2 \cdot \frac{\bar{p}(1 - \bar{p})}{n}$$

$$n = \frac{z^2\ \bar{p}(1 - \bar{p})}{e^2}$$

Therefore, any n greater than or equal to $\frac{z^2\ \bar{p}(1 - \bar{p})}{e^2}$ will give us accuracy to

e units of measure with $(1 - \alpha)$ level of confidence.

Example 1: How many phone calls would the New York Times have to make in order to be 99% confident that the very closely contested Senate race winner would be correctly predicted? Let maximum error equal .01. This is arbitrary and is left to the researcher and interested parties to decide. It may be very expensive to reduce the error level in real world situations.

$$\text{Use } n = \frac{z^2\ \bar{p}(1 - \bar{p})}{e^2}$$

$$z_{(1-\alpha)/2} = z_{.495} = 2.57$$

Let $p = .5$, $e = .01$

Therefore, $n = \frac{(2.57)^2\,(.5)\,(1 - .5)}{(.01)^2} = 16{,}512.25$, so that $n \geq 16{,}513$ will give us 99% confidence in predicting the winner.

The formula can easily be adjusted to either of the remaining confidence interval formulas. For example, our formula for estimating the mean of the population based upon sample result $\bar{x}$ was:

1) $\bar{x} \pm z\sigma/\sqrt{n}$ [$n \geq 30$] or

2) $\bar{x} \pm t\sigma/\sqrt{n}$ [$n < 30$]

Simply let $e = \frac{z \cdot \sigma}{\sqrt{n}}$

To determine the minimum sample size for an arbitrary level of confidence and predetermined error level, solve for n:

$$e^2 = \frac{z^2 \sigma^2}{n}$$

$$n = \frac{z^2 \sigma^2}{e^2}$$

Example 2: How many college liberal arts students would have to be sampled in order to establish the mean GPA of such students nationwide with 95% confidence, given maximum allowable error of .1? [A sample of 300 students yields an estimate s = .3]

$$n = \frac{z^2 \sigma^2}{e^2}$$

Since $\alpha = .05$, $\alpha/2 = .025$, and $z_{.025} = 1.96$, we have

$n = \frac{(1.96)^2 (.31)^2}{(.1)^2} = 34.57$, so that for $n \geq 35$ we can have 95% confidence in our results.

EXERCISES 1.8

1. Redo example 1, if we decide to permit 95% confidence. Would this be a wise idea for the New York Times?
2. Redo example 2, if we set $\alpha = .01$.
3. Solve for n if we use the formula for small sample population mean estimation.
4. Suppose for example 2, we determine that $n = 25$ but don't know σ. Estimate σ with 95% confidence.
5. How large a sample of AIDS patients would be required in order to estimate mean longevity at 99% confidence, if a sample of 25 yielded $s = 2.5$ and our maximum error were set at 1 year?

1.9 Two Sample Difference of Proportions

Suppose we are testing whether an innovative type of teaching is more effective than traditional instruction. Further, let us define success as attaining a score of 60% or better, with failure defined as a score less than 60%. Such a design naturally leads itself to testing whether $P_1 > P_2$, that is, whether the proportion of students passing with the innovative style is greater than the proportion passing with the traditional style.

The z test is used with: $$z = \frac{P_1 - P_2}{\sigma_{P1-P2}}$$

$$\sigma_{P1-P2} = \sqrt{p(1 - p)\,(1/n_1 + 1/n_2)}$$

Since we are not sure that $P_1 > P_2$, to obtain p, compute the proportion passing overall:

$$P = \frac{x_1 + x_2}{n_1 + n_2}$$

x_1 = number passing with style 1

x_2 = number passing with style 2

Example 1: Consider that 67 of 108 college freshmen pass Calculus I; if they have a scheduled problem session with a teaching assistant, the proportion is 70 of 102. Is the practice of scheduling the problem session effective in increasing passing percentage?

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}}$$

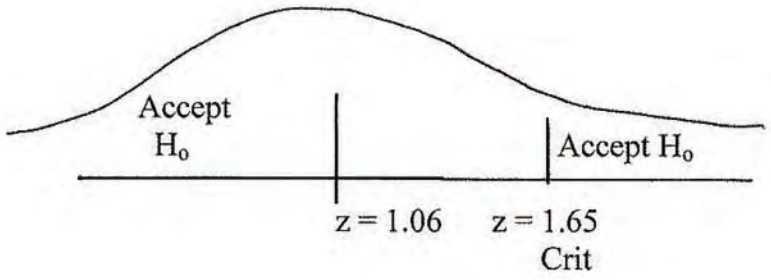
$$P_1 = \frac{70}{102} = .69$$

$$P_2 = \frac{67}{108} = .62$$

$$P = \frac{70 + 67}{102 + 108} = .65$$

$$z = \frac{.69 - .62}{\sqrt{.65(.35)\,(1/102 + 1/108)}} = \frac{.07}{.066}$$

$$z = 1.06$$



There is no significant difference between the two results.

This doesn't mean we should discontinue the lab hours. The value z = 1.06 corresponds to a probability of .3554.

Crudely, we could place 85% confidence in the value of our lab hours. However, a coin flip carries with it 50% confidence, so 85% confidence is much better than a coin flip but much worse than the 95% or 99% confidence that is the minimum standard for research.

A review of the one sample difference of proportions is included in Appendix B.

EXERCISES 1.9

1. Redo example 1, if $P_1 = 105/210$, $P_2 = 106/183$.

2. The FDA has examined the results of an experimental drug. Of 100 patients given the new drug, 53 fully recovered. Of 100 patients who are given a placebo, 49 fully recovered. Should the FDA conclude that the drug was effective in patient recovery?

3. An experiment with college individualized instruction resulted in 60 of 110 students receiving a grade of C or better. For the students who enrolled in conventional classroom instruction, 57 of 95 passed with a grade of C or better. Is individualized instruction significantly more effective than conventional instruction?

5. In a longitudinal study of college majors, it was found that 1150 of 2000 remained (10 years after graduation) in careers closely related to their undergraduate major. Is this result different from an earlier finding that 900 of 1520 students remained in closely related careers?

1.10 Correlation Analysis

In correlation analysis we study the degree of linear association between two variables.

For example, a marketing executive would like to show that there is a high positive correlation between n, the number of newspaper ads for a product, and y, gross sales.

Consider the table below:

n	3	5	0	7	11	13
y	21	29	18	29	32	25

Correlation coefficients r range: $-1 \leq r \leq 1$. The value -1 is the strongest possible negative correlation, i.e., the relation between how many pieces of cake you have at a party and your resultant self esteem. The value 0 shows no association between the two variables, i.e., the relation between your salary and your shoe size. Actually, there may be a modest relationship between shoe size and salary since men have larger sizes than women. However, with more women taking advanced mathematics courses like modeling, this situation will certainly change soon. Statistics frequently has unexpected results. To compute r, use the standard formula:

$$r = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n \Sigma x_i^2 - (\Sigma x_i)^2}\ \sqrt{n \Sigma y_i^2 - (\Sigma y_i)^2}}$$

Please use a pocket calculator with a correlation key (for example a Tandy PC-7) or a software package. The computation of correlation can be cumbersome with paper and pencil, particularly with large n.

For the earlier marketing example, $r = .66$. This is a high positive relationship which should encourage further investment.

To test whether a linear correlation coefficient is significantly less than, greater than, or equal to zero, we employ a t test.

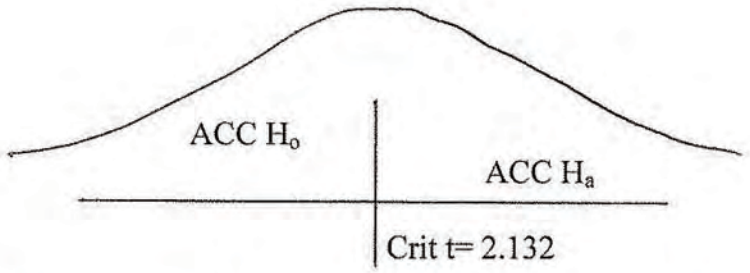
$$t = r\sqrt{\frac{n-2}{1-r^2}} \text{ with } (n-2)\text{df.}$$

Example 1: Test whether .66 is a significant positive correlation related to the earlier marketing data.

$$t = .66\sqrt{\frac{6-2}{1-(.66)^2}} = 1.76$$

Let H_o: $r = 0$

H_a: $r > 0$ Conclusion: Accept H_o
$r = 0$



To obtain critical t use a t table where you look up (n - 2) degrees of freedom, which in this example is 4df. Since we are testing whether the correlation is significant positive, we use a one-tailed test and look up $t_{.05,4df} = 2.132$. For this example, the correlation of .66 would not be a significant positive correlation. This is unusual and due to the small sample size. In the vast majority of cases, this level of positive correlation is statistically significant and is almost as large as one could reasonably expect to discover. For example, the correlation between your

high school SAT score and college grades is much lower than .66 but is statistically significant because the number of students in the correlation study exceeds one million.

EXERCISES 1.10

1. For the pseudo-random number 517724, compute the correlation between a_n and a_{n+1}, the nth and (n+1)st digit.

   Hint: Let $x_1 = 5 \quad y_1 = 1$

   $x_2 = 1 \quad y_2 = 7$

   etc.

2. Use a numerical package to calculate π to 10 digits. Compute the correlation between a_n and a_{n+1}. Is it significantly different from 0? $\pi = 3.14159$. Let $a_1 = 1$, $a_2 = 4$, $a_3 = 1$, $a_4 = 5$, $a_5 = 9$, etc.

3. Graph the sample data and try to determine by the distribution of the points whether there is a linear relationship (straight line) between a and b. Compute r by software package and by hand.

a	b
3	8
3	6
4	3
7	2
8	2
9	6
10	8

4. a) For what value of n would a correlation coefficient of .66 be significantly greater than zero? ($\alpha = .05$)

b) Is there any way other than trial and error (let n = 6,7,8...) to solve the inequality above for n?

5. Research Question

Use Numerical Analysis to compute $\sqrt{2}$ beyond calculator accuracy. Test the digits for autocorrelation between successive digits. Is there a correlation (R=0) between successive digits? Test using alpha = 5%.

H_o: $R = 0$

H_a: $R \neq 0$

Even if you find a zero correlation between the first billion digits, there is no assurance that the next billion digits will have the same result. As a mathematical researcher in randomness once commented: "Capturing randomness is life trying to capture the fog."

1.11 Elementary Linear Regression

One of the most important topics in statistics is ***regression analysis***, in which y (the dependent variable) is estimated by an equation of the form $y = a_o + a_1x$ (***elementary linear regression***), $y = a_o + a_1x_1 + a_2x_2 + \ldots a_nx_n$ (***multiple regression***), or myriad other forms such as $y = ae^{bx}$, $y = a + bx + cx^2$, etc. In this section we will focus on the simplest regression form, $y = a_o + a_1x$ (***elementary regression***).

Sometimes the equation is trivial. For example, in a supermarket, one head of broccoli costs \$.99, 2 cost 2(.99), and we could write y(cost) = .99(x), where x stands for the number of heads of broccoli.

In the real world we seldom obtain a perfect fit between our regression curve and the data. For this reason, we use a technique called ***least square*** to define the best fitting curve.

Consider the data (real world) points below. Each error, ℓ_i, is the vertical deviation from the point to the line.

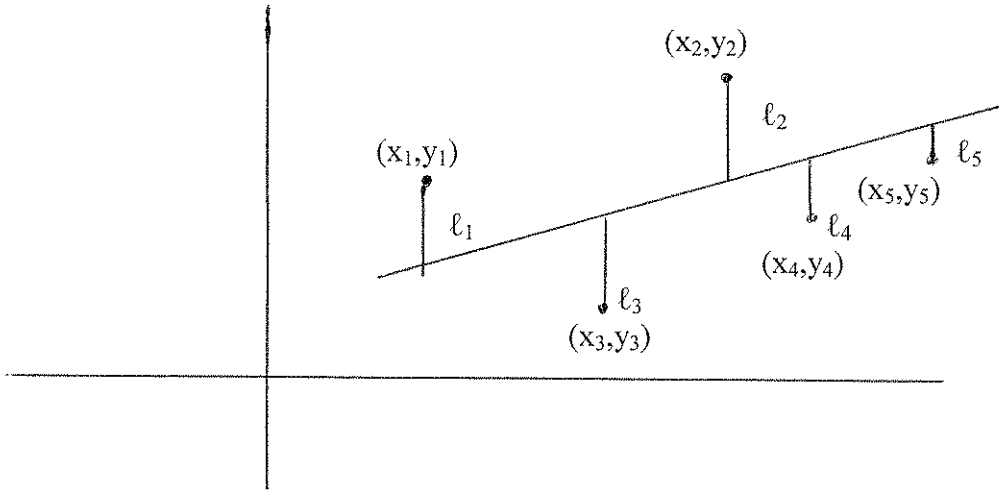


Figure 1.

We want $\sum \ell_i^2$ to be a minimum. This leads to a ***least-square line***. To find the best fitting regression line, we use Calculus III and partial differentiation. We start with n points (x_1,y_1), (x_2,y_2), ... (x_n,y_n) and want the best fit $y = a_o + a_1x$.

Consider our errors from Figure I. Since we assume $y = a_o + a_1x$ is our best fitting regression line, we define this by $\sum \ell_i^2$ to be a minimum.

We can write $\ell_1 = a_o + a_1x_1 - y_1$. We then write $\sum \ell_i^2 = \sum (a_o + a_1x - y)^2$. This is a function of a_o and a_1. To yield minimum error, take the partial derivatives as follows:

$$\frac{\delta (\Sigma \ell_i^2)}{\delta a_o} = \Sigma\, 2\,(a_o + a_1x - y)$$

$$\frac{\delta (\Sigma \ell_i^2)}{\delta a_1} = \Sigma\, 2x\,(a_o + a_1x - y)$$

Set both partials equal to 0. This gives us $\Sigma\, y = a_o n + a_1 \Sigma\, x$

$$\Sigma\, xy = a_o \Sigma\, x + a_1\ \Sigma\, x_2$$

(These are called the ***normal equations*** for the least-squares line.)

Solve for a_o and a_1:

$$a_o = \frac{(\Sigma y)\,(\Sigma x^2) - (\Sigma x)\,(\Sigma xy)}{n \Sigma x^2 - (\Sigma x)^2}$$

$$a_1 = \frac{n \Sigma xy - (\Sigma x)\,(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$

As an alternative procedure, we could solve for a_1 and use the equation $y = a_o + a_1 x$ to solve for a_o, where $\overline{x} = \frac{\Sigma x_i}{n}$, $\overline{y} = \frac{\Sigma y_i}{n}$.

To illustrate, consider the data points below relating x, price of a product, to y, sales.

x	y (thousands)
100	50
150	40
200	38
300	35

To find the least squares regression line of y on x, use the second set of equations.

$$y = a_o + a_1 x$$

$$a_1 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{4(29.100) - (750)163}{4(162{,}500) - 750^2}$$

$$a_1 = \frac{-5850}{87500} = -.07$$

Then use the simpler equation $\overline{y} = a_o + a_1 \overline{x}$ to solve for a_o.

$$\overline{y} = 40.75$$

$$\overline{x} = 187.5 \text{ and } a_1 = -.07$$

Substituting these values into the equation

$$40.75 = a_o + (-.07)(187.5)$$

$$a_o = 53.88$$

$$\therefore y = a_o + a_1 x, \qquad y = 53.88 - .07x$$

The preceding example only illustrates linear regression. We need more than four points and also should graph the points to see whether the least squares function is linear. It could be exponential, quadratic, etc. After obtaining the least squares line, it is sound to compute the

coefficient of determination, $r^2 = 1 - \frac{\Sigma (y - \hat{y})^2}{(y - \overline{y})^2}$, where $\hat{y}$ is the value of y using the least squares line. If the regression line explains no variation, then $r^2 = 0$ and $r = 0$. If $r^2 = 1$, then we have the rare perfect linear correlation and perfect linear regression. Let us return to our example in the chart below. We will compute the coefficient of determination by the formula given above.

x	y	$\hat{y}$	
100	50	46.88	$\hat{y} = 53.88 - .07x$
150	40	43.38	$\hat{y} = 40.75$
200	38	39.88	
300	35	32.88	

$$r^2 = 1 - \frac{\Sigma (y - \hat{y})^2}{\Sigma (y - \hat{y})^2}$$

$$r^2 = 1 - \frac{(50 - 46.88)^2 + (40 - 43.38)^2 + (38 - 39.88)^2 + (35 - 32.88)^2}{(50 - 40.75)^2 + (40 - 40.75)^2 + (38 - 40.74)^2 + (35 - 40.75)^2}$$

$$r^2 = 1 - \frac{29.1876}{126.75} = .77$$

This computation and example captures the relationship between correlation and linear regression. The coefficient of variation, r^2, equals .77. This value, r^2, represents the percentage of variation in the observed y values that is explained by the least-squares regression line. The value r explains how well the regression line fits the sample points. If $r^2 = 1$, then $r = \pm 1$, which is perfect linear regression. If $r = 0$, then the total variation in y is not explained at all by a least-squares line.

For our example $r^2 = .77$. Therefore, 77% of the variation of y is captured by our least-squares regression line. This is a high level of total variation explained by the line. Of course, 23% is unexplained variation. Whenever performing real world regression analyses, this author recommends a follow-up data analysis with new points to check whether the first equation that one obtained was sound. The least-squares technique simply reduced error, $\Sigma \ell_i^2$, to a minimum. It can never tell whether we have uncovered a valid relationship or simply minimized error.

EXERCISES 1.11

1. Write the least-squares regression line for the following data points. Let x = temperature, y = # of milk shakes sold.

x	y
82	130
86	200
92	330
96	410
92	550

2. Determiner r^2, the coefficient of determination, for the above data.

3. Write the least-squares regression line for the following data points. Let x = sales, y = profits.

x	y (thousand)
160	230
200	310
320	430
400	510
500	620

4. Find r^2 for the above data.

5. Derive, using standard calculus techniques, the equation

$$a_1 = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$

Hint: Start with $f = \ell_1^2 + \ell_2^2 + \ldots + \ell_n^2$

Use partial differentiation, setting the partials equal to zero

and solving.

Note: $\ell_1^2 = (a_o + a_1x_1 - y_1)^2$

6. Show that the least-squares line passes through the point (x,y). This justifies our using the second set of equations in our elementary linear regression model.

1.12 Quadratic and Exponential Regression

Quadration Regression

We could extend the previous concepts of elementary linear regression to parabolic, exponential, etc. To illustrate, consider ***quadratic regression*** if the data points resemble a parabola. If we consider our regression equation to be of the form:

(1) $y = a + bx + cx^2$

we can solve for a, b and c by the following equations, called ***normal equations***:

(2) $\Sigma y = na + b\,\Sigma x + c\,\Sigma x^2$

(3) $\Sigma xy = a\,\Sigma x + b\,\Sigma x^2 + c\,\Sigma x^3$

(4) $\Sigma x^2y = a\,\Sigma x^2 + b\,\Sigma x^3 + c\,\Sigma x^4$

Equations 2, 3, and 4 are derived by multiplying equation 1 by 1, x, and x^2 respectively. Naturally, we could extend this reasoning to $y = a + bx + cx^2 + \ldots + dx^n$, obtaining n equations (hopefully all independent equations) in n unknowns. For example, $x + y = 2$ and $2x + 2y = 4$ are really the same equation with the same straight line for a graph and are called ***dependent***. For a more rigorous treatment of this concept, the reader should take a course in linear algebra, where the concept of linear independence deals with this issue in its general form.

For most types of regression, one uses computer packages. As the regression analysis increases in complexity, the computations usually are so extensive that the statistical software packages become the only viable approach.

Exponential Regression

If we are given the equation of modeling either continuous growth or continuous decay, the problem could be translated into the differential

equation $\frac{dy}{dx} = f(x,y)$. For example, consider that we are given the question

of computing the amount of money we accumulate in our savings account with a rate of interest equal to 5% continuously compounded. Suppose we start with \$100 in our account and are asked for the amount that we have after four years.

This problem can be approached as follows:

Let P = total amount of money in savings at time t.

(1) $\frac{dP}{dt} = .05P$

We can now solve for P using elementary calculus. This method, called separation of variables, is the early part of the content of differential equations.

Divide both sides of equation (1) by P and multiply both sides by (dt). We obtain:

(2) $\frac{dP}{P} = .05dt$

Next integrate both sides of equation (2). We derive:

(3) $\ell n\, P = .05t + c$

We know at time $t = 0$, that $P = 100$. Therefore, substitute to find c.

$$\ell n\, 100 = c$$

Now by exponentiating both sides of equation (3) using base e, we obtain:

$$\ell\, \ell n^{P} = P = e^{.05t + \ell n 100}$$

This can be simplified to the conventional form that we find in textbooks that cover continuous compounding of interest, which is $P = 100\, e^{.05t}$. This means that total savings accumulated (with our notation P) equals the original principal multiplied by e^{it}, where i is the

interest rate and t is the time. You might see the formula for continuous compounding $A = P e^{it}$.

To complete our problem, let $t = 4$, $i = .05$.

Total amount $= 100\, e^{.05(4)} = 100\, e^{.2} = 122.14$

A similar model would be used to predict population growth, effective timing of drug doses, and myriad continuous phenomena. For population growth consider the following model:

Let P = population of the United States at time t

We can start with $\frac{dP}{dt} = cP$, where c is a positive number

We should have some initial condition, $P(t_o) = P_o$. For example, we know that the United States population was 150,697,000 in 1950. Let $t_o = 1950$. Let $P_o = 150{,}697{,}000$.

We can derive the Malthusian population growth equation:

$$P(t) = P_o\, e^{c(t-t_o)}$$

If we know that the 1970 population of the United States was 203,211,926, we can find c as follows:

$$\frac{203{,}211{,}926}{150{,}697{,}000} = e^{c(20)}, \text{ so that } c = .015\ *$$

We then can use this value of c to predict future U.S. populations. We leave the computation of the U.S. population in the year 2100 as one of the exercises for this section.

---

*Giordano, F. and Weir, M. *A First Course in Mathematical Modeling* (Monterey: Brooks/Cole Pub., 1985), pp. 306-307.

Though you will likely spend a semester studying differential equations, we present at the end of this chapter two research activities from differential equations by prominent mathematicians. These research problems will be presented in a manner that does not require a preliminary course in differential equations and were designed to stimulate your interest in this fertile mathematical specialty.

EXERCISES 1.12

1. Find the least squares parabola for the following data points:

x	y
2	5
3	10
6	39
10	112

2. Find the least squares parabola for the following data points:

x	y
1.5	3
2.6	6.2
3.4	12.1
5.1	27.2

3. How would one extend least squares parabolic regression to cubic [$y = a + bx + cx^2 + dx^3$] ? Generalize to nth degree regression.

4. Predict the United States population in the year 2100 from the exponential equation that we derived for population growth. Compute the estimate of U.S. population for the years 2200, 2300, and 3000. Is the present level of population growth possible?

1.13 Multiple Linear Regression

One of the most useful topics in statistics is multiple linear regression. In multiple linear regression, the dependent variable is estimated as a linear combination of n independent variables, $x_1, x_2,...x_n$. For example, consider y to be the Dow Jones stock average. Let x_1 = the prime rate, x_2 = the current Federal deficit, etc. If Marie's multiple regression equation has better and more refined predictor variables, $x_1, x_2,...x_n$, she makes a fortune.

As a second illustration, consider the prediction of college GPA (y) by a test score such as the SAT (x_1) and other measures such as high school rank (x_2). If we can write a valid, good fitting equation such as $y = a_1x_1 + a_2x_2 + a_3x_3$, we can direct counselors to intervene early with students whose predicted GPAs are in the trouble zone. Consequently, we can reduce the drop out rate.

In practice, multiple linear regression is particularly cumbersome and is nearly always done by computer. We present an illustrative example together with an analysis of the mathematics that underlies the process.

The form of a multiple linear regression equation is:

$$y = a_o + a_1x_1 + a_2x_2 + ... a_nx_n + e$$

where $x_1, x_2, ...x_n$ are the independent predictor variables, y is the dependent variable, and e is the error associated with the model. If n = 2, we have a plane, provided x_1 and x_2 are not the same variable.

We can solve for the regression coefficients for only the simplest examples. Even then we need the power of the computer to determine whether all independent variables should be

kept. Sometimes one is so closely related to another (i.e., correlation coefficient of .9), that the computer can delete one of the variables.

To illustrate the process without considering relatedness of the independent variables, consider the problem of predicting GPA (y) on the basis of SAT stanine rank (1-9)(x_1) and rank in class, where 1 = bottom 10 percentile, 2 = 10-20 percentile, ..., 10 = 90-99+ percentile (x_2). We need a large sample, but for the ease of computation, consider a sample of ten students. Their statistics after freshman year are given below:

Student	GPA (y)	SAT (x_1)	Rank (x_2)
1	2.6	4	4
2	3.2	7	7
3	4.0	10	10
4	1.6	7	8
5	2.7	6	6
6	3.3	8	7
7	2.8	6	8
8	1.8	4	4
9	3.7	9	9
10	2.5	5	8

The objective is to find a_o, a_1, and a_2 where $y = a_o + a_1x_1 + a_2x_2$.

y = GPA (at the end of freshman year), x_1 = SAT (stanine score), and x_2 = rank in class (1-10).

To find the least squares regression plane, use the normal equations: [We solve for a_o, a_1, and a_2]

$$\Sigma y = na_o + a_1\Sigma x_1 + a_2\Sigma x_2$$

$$\Sigma x_1y = a_o\Sigma x_1 + a_1\Sigma(x_1)^2 + a_2\Sigma x_1x_2$$

$$\Sigma x_2y = a_o\Sigma x_2 + a_2\Sigma x_1x_2 + a_2\Sigma(x_2)^2$$

These equations are derived by starting with $y = a_o + a_1x_1 + a_2x_2$ and multiplying by 1, x, and y and then summing.

To complete our example,

$\Sigma y = 28.2$ $\qquad$ $\Sigma x_1y = 196.4$

$n = 10$ $\qquad$ $\Sigma (x_1)^2 = 472$

$\Sigma x_1 = 66$ $\qquad$ $\Sigma x_1x_2 = 498$

$\Sigma x_2 = 71$ $\qquad$ $\Sigma x_2y = 207.8$

$\Sigma y^2 = 84.8$

Substitute into the normal equations:

$$28.2 = 10\,a_o + 66\,a_1 + 71\,a_2$$

$$196.4 = 66\,a_o + 472\,a_1 + 498\,a_2$$

$$207.8 = 71\,a_o + 498\,a_1 + 572\,a_2$$

Solve using determinants:

$$a_0 = \frac{\begin{vmatrix} 28.2 & 66 & 71 \\ 196.4 & 472 & 498 \\ 207.8 & 498 & 472 \end{vmatrix}}{\begin{vmatrix} 10 & 66 & 71 \\ 66 & 472 & 498 \\ 71 & 498 & 472 \end{vmatrix}}$$

$$a_0 = \frac{-18968.8}{-20328} = .93$$

Substitute and find that $a_1 = .27$ and $a_2 = .02$. The final multiple regression equation relating GPA (y) and SAT (x_1) and class rank (x_2) is:

$$y = .93 + .27\,x_1 + .02\,x_2$$

After completing one such example, you will probably agree that multiple regression is better left to the computer.

Several assumptions are associated with multiple linear regression including:

1) There exists a mathematical relationship between the dependent variables and independent variables.

2) The errors are normally distributed and independent.

It is recommended in practice that the researcher perform two independent data samples and compute the multiple regression equation independently to verify the first equation. After all, all we are performing is a process that minimizes error - not necessarily establishing a relationship between the variables.

Interpreting Computer Output

In practice the computer will test whether each coefficient a_o, a_1, ...a_n is zero - which means that the independent variable should be deleted from the multiple linear regression equation.

There are many statistical packages to ease the burden of lengthy computations. One of the most popular systems is Minitab which is easy to use and powerful for most analyses that the statistics student would need. Consider this invented example in the form of Minitab output:

MTB > REGRESS Y IN C1 USING 3 PREDICTORS IN C2-C4

The regression equation is:

C1 = 2.53 + .196 C2 + 2.46 C3 - 0.162 C4

Predictor	Coefficient	St. Dev.	t-ratio	P
Constant	2.5281	1.1310	2.31	.050
C2	.1956	.0721	3.36	.010
C3	2.4610	.8713	2.90	.020
C4	-.1619	.1685	.889	.400

The p value of C4 is .40. This is related to the two-tailed t ratio, which tested whether the regression weight Ci = 0. Since the p value for C4 is .40 and is greater than .05 (assuming type I error α = .05), the computer would reject C4 as a valid independent variable. The researcher should then repeat the process - if the computer package hasn't already done this - with the data associated with C1, C2, and C3.

Another problem in multiple regression analysis is multi-collinearity, in which certain independent variables x_i are highly correlated or some x_i can be represented as a linear

combination of other independent variables. To eliminate this problem, a sound statistical package - such as SPSS - will use a stepwise selection process. Independent variables are put into the equation one at a time and the package is designed to generally delete variables that are highly correlated with other variables in the equation. For a much fuller discussion of multiple linear regression, consult *Introduction to Business Statistics*, A. Kvanli, C. Guynes, and R. Pavur (St. Paul: West Publishing, 1989), pp. 561-641.

EXERCISES 1.13

1. Company profits y are assumed to be related to advertising (x_1) and efficiency of quality control (x_2).

   a. Use normal equations to write the multiple linear regression equation $y = a_o + a_1x_1 + a_2x_2$, relating profits to the given two independent variables, x_1 and x_2 listed below.

   b. Use a software package to find the multiple linear regression equation. Interpret the output associated with the computer analysis.

y (millions)	x_1 (millions)	x_2 (1-10)
23	2	9
11	3	7
36	7	9
10	8	4
36	25	8
5	3	7
56	21	10
32	11	9
19	7	5
28	11	4

1.14 Modeling Research Problem One

Multiple Regression Models to Measure Our Long-Term Influence

It may surprise you that as a gifted sophomore, junior, or senior you can become a player in the exciting world of creating new mathematics. But in modeling, you may be able to contribute ideas that assist in solving many problems that face researchers today. This is because modeling is a relatively new discipline, emerging as an outgrowth of discrete mathematics, computer science, and calculus-based probability and statistics - a late twentieth century convergence of knowledge. Mathematical modeling can be used in deterministic processes, such as the estimation of area under a curve. If we don't rely upon chance and random variables, we are using a deterministic process. In elementary calculus you learned that if $y = x^2$, then $dy/dx = 2x$. This is a deterministic process; you were not estimating the derivative. However, we will shortly show how the estimation of a deterministic process like integration can be made using probabilities and modeling simulations. If you wished to contribute new knowledge to Real Analysis, it would take several years of study to begin to even ask research questions appropriately. That is because mathematicians have been applying their collective energies to this field for several hundred years.

The research directions in this book are deep, difficult problems that may never yield final solutions. However, if we advance from square one to square two in understanding, that is progress.

The several questions that will be presented as material develops in the book come from extremely varied disciplines which include number theory, psychometrics (statistics and

psychology), and mathematical statistics. If you are excited by one or more of the problems, this may signal that a doctorate in the specialized field would be appropriate for you.

In the words of my graduate school professor, Anthony Ventriglia (Manhattan College), "Mathematics is a participant sport, not a spectator sport. Become involved."

RESEARCH AREA 1 (PSYCHOMETRICS)

SIGFLUENCE [Significant, long-term, positive influence]

How can we measure the key personal influences that shape our lives? I was reluctant to include a psychology/statistics research question in a mathematical modeling text. However, each of my Numerical Analysis students encouraged me to include this material, and they represent the gifted students that this book is geared toward.

This simple question is thought by many people to be too complex to even attempt. In fact, a colleague of mine from the Canadian Society for the History and Philosophy of Mathematics, who was trained at the University of California, Berkeley in mathematical logic, told me that he thought this concept was inscrutable. I disagree and have spent nearly ten years trying to advance our knowledge of this neglected experience. Before discussing these ideas further, please complete the Sigfluence Survey, which is included at the end of this chapter. Also score your survey using the key at the end of the survey. Instructions for scoring are given below.

Scoring the Survey

Now that you have taken the survey, you have to score it and obtain three scores for yourself. The scoring key is included immediately after the survey. You will have to add the scores to the items that you have circled. Each item is scored from 1 to 11 (there are 11 x's). If

an item has an R, this means that the number of the extreme left x is reversed and becomes an 11; the extreme right x (for an R item) becomes a 1. Add up the numbers corresponding to the individual items and obtain three scores:*

1) Actual Sigfluence

2) Potential for Sigfluence

3) Need for Sigfluence

Sigfluence is my new word that means significant, long-term, positive, interpersonal influence. The importance of understanding sigfluence better should be fairly obvious. After all, many people believe that we are eventually going to destroy ourselves with nuclear weapons or environmental insensitivity. Also, many young, idealistic people enter teaching or social work to make a difference in peoples' lives. Some do and are satisfied with their professional life. Some don't and experience burn-out and malaise.

My theory is that sigfluence is a basic need of every person. This is different from the way psychology and education view this today.

According to current (and in my opinion naïve) thinking, sigfluence is an interest that helping professionals satisfy in their "altruistic" career. Some people like sports; helping professionals like helping. Since we don't have an adequate measure of actual sigfluence at present, we rely on the common perception that teachers and social workers help a great many people. In time we may find that our common perceptions are wrong.

---

*If you have no children, give yourself a score of neutral (6) for the two items related to your children.

Now let's consider several activities that are reasonable for fresh, creative minds like yours to pursue. First, we need to develop a more effective measure of Actual Sigfluence than my survey gives. The Actual Sigfluence score that you obtained is the result of several years of computer analyses, correlation matrices, consultations with experts in test development, and extensive revisions. But it needs improvement and refinement.

One possible way of developing an improved Actual Sigfluence score is by interviewing people and asking them about the nature of their Actual Sigfluence. You should consult a text on qualitative research before you interview people. Otherwise you will likely bias the results. Qualitative research, according to researchers at Clark University's International Conference on Theoretical Psychology (June 1991), is an extremely important means of advancing our knowledge of phenomena often considered in strictly numerical terms. Even quantitative experts like former Federal Reserve Chairman expressed interest in qualitative research to determine why Americans had such a deep pessimism about the economy. The National Science Foundation gave me little hope for a grant if I submitted a proposal to study sigfluence from a qualitative and quantitative perspective. Fortunately I changed gears and was awarded a three year grant to innovate mathematical modeling activities for gifted students like you. And by the time that you are ready for National Science Foundation grants, perhaps qualitative research will be accepted as the important focus that it deserves.

Let us turn to a more conventional multiple regression research approach. A multiple regression study that one of my graduate students recently completed took data from 292

completed surveys and related seven survey variables to the status of the present occupation of the person (completing the survey).

Status is a difficult occupational variable to assign a value, and my student used the MSE 12 index of job status from the 1963 National Opinion Research Corporation. A follow-up study with a more current status measure would be profitable for future research.

The dependent variable Y was the status of the respondent. The salient variables related to Y were determined by SPSS. SPSS is a comprehensive tool to analyze data. An extraordinary range of procedures are available ranging from elementary to advanced statistics. SPSS is available at most colleges and universities; it is also available for purchase by individuals but is fairly expensive.

The computations are so involved in most multiple regression analyses that nearly all multiple regression studies have to be done with appropriate software. Our results identified the seven significant independent variables from the survey as:

(1) x_1 - status of the key influencer

(2) x_2 - age when influenced

(3) x_3 - age today

(4) x_4 - the degree of influence you perceive affecting your life (0-10)

(5) x_5 - perception of home environment (1=permissive, 10=authoritarian)

(6) x_6 - the length of time to determine the positive impact

(7) x_7 - the length of time of contact with the influential agent.

The SPSS generated multiple regression equation was:

$$Y_1 = 44.39 + .209x_1 - .213x_2 - .087x_3 + .233x_4 - .161x_5 - .146x_6 + .06x_7$$

This is a very preliminary result. We need better measures of status and better measures of the independent variables. But this is only a start and should only be considered as such.

A promising multiple regression study would be to use the Actual Sigfluence score as a dependent variable and to use SPSS or a similar powerful computer package to identify salient survey variables. These variables would yield a multiple regression equation to quantify Actual Sigfluence based on two or more survey variables. Of course, improved measures of Actual Sigfluence may translate into improved independent variables in any multiple regression analysis.

A preliminary multiple regression analysis was completed by one of my graduate school students - Gloria Rousseau of Iona College (in 1989). She used SPSS to write a multiple regression equation from the survey data with 61 original cases - 39 females (code = 1), 22 males (code = 0). I have included the multiple regression output together with Gloria's coding sheets. You should obtain a more recent socioeconomic measure of occupational status than Gloria used if you wish to experiment with multiple regression analysis of Actual Sigfluence.

One immediate problem with Gloria's study was the small sample size (For $k = 2$ independent variables, $n - k - 1 = 21$. Therefore, $n = 24$ cases had complete data and were used).

The resulting multiple regression equation:

$Y(\text{Actual Sigfluence}) = .328x_1 + 8.241x_2 + 54.742$ was so counter to my previous research and preliminary understanding that I believed that we were at an impasse. For example, my initial SPSS correlation years ago between Actual Sigfluence and occupational status of the respondents was -.01 for a sample of 292. Additionally, in an exhaustive study, Jencks found a

significant relation between father's occupational status (not mother's) and the subsequent achievement of the sons.* Also, Gloria Rousseau's separate multiple regression equations for male, female and pooled data were so different that I questioned whether the statistical or mathematical route was the best way to explore sigfluence.

I believed, and still do, that Rousseau's result simply was an equation that allows us to reject the null hypothesis in multiple regression - namely that the beta weights are all zero. A common misperception in regression is that there is a necessary relationship between the dependent variables and the independent variables. If you review the ways that computer packages determine the values of $b_o, b_1, \ldots b_n$ in a multiple regression equation, you will find that they minimize the value $SSE = (Y - \overline{Y})^2$. Minimizing this error (between our predicted Y and observed Y values) is quite different from demonstrating that there is a valid relationship between the dependent and independent variables. A follow-up study is always helpful to establish whether your equation truly holds up. A follow-up study with a larger sample is called for to determine if we can replicate Rousseau's results.

This is not my current direction. At Clark University's International Conference on Theoretical Psychology (June 1992), I described these results and much material from my two books on Sigfluence, whose references follow. Two researchers recommended that I read Dr. Amedeo Giorgi's work on qualitative research, and this has been my major focus for the past year. I have given two references at the end of this chapter for those readers interested in

---

*Jencks, C. et. al., *Inequality* (New York: Basic Books, 1972), p. 194

qualitative research. But for those of you who would like to perform more conventional statistical analyses, another modeling avenue involves ***goodness of fit testing***.

Consider gathering data regarding the age when people encounter their key influencer. You would have to interview a large sample of people (perhaps $n > 100$) and perform a goodness of fit test of the distribution of present age of the respondents.

Ideally, the age of the respondents should be representative of the United States population. To check this, a goodness of fit test is required. Then we should analyze the distribution of ages X corresponding to the onset of the key influence - when the sigfluence began in the perception of the person. For example, if your ninth grade algebra teacher was your key influencer (judged at age 30) and you were 13 years, 6 months when you first entered her class, your X score for the Sigfluence Distribution would be 13.5. We have tried two goodness of fit tests with our original set of 292 valid, completed surveys. Both Poisson and normal distributions were used to fit our Sigfluence Distribution and neither was a good fit. To be fair, we didn't test whether our 292 respondents were representative of the American population.

But be careful. If we keep trying goodness of fit tests, one will eventually fit as a result of chance. And each goodness of fit test carries a conventional level of Type I error of 5%. So modeling and statistics have to be employed wisely.

Also perform (if time permits) follow-up studies with new data to determine whether your goodness of fit test holds up. Positive results are frequently ephemeral, particularly when you are trying to predict one of the most elusive mysteries - human behavior.

To assist your research into "sigfluence," the following references are suggested:

PUBLICATIONS by John Loase

1. *Statistics Made Easy.* The Graduate Group, 2009.
2. *The Sigfluence Generation.* AEG Publishing, 2009. Second place Benjamin Franklin National Book Award, non-fiction, 2011.
3. "How to Excel at Mathe Transformation," Mathematical Association of America, *Focus*, April 2009.
4. "A Variation on the Hardy-Ramanujan Partition Function," *College Mathematics Journal*, Sept. 2005.
5. *Theory and Measurement of Sigfluence.* University Press of America, 2002.
6. *Our Neglect, Denial and Fear.* Nova Science, 2000.
7. *Sigfluence.* BlueBird Publishing, 1997.
8. *Sigfluence: The Key to 'It's a Wonderful Life.'* University Press of America, 1996.
9. *Sigfluence: Long-Term, Positive Influence.* University Press of America, 1994.
10. "A Quantitative and Qualitative Analysis of Sigfluence," Mathematical Modeling. Pergamon Press, 1995, (invited article by editor Dr. X. Avula).
11. "Interdisciplinary Statistics," *Inquiry*, Fall 1991.
12. "Can a Polynomial Function with Whole Number Exponents, Integer Coefficients and Domain Natural Numbers Map to the Infinite Decimal Expansion of Any Irrational (mod k), k a Whole Number?", proposed and solved by Loase and several others, *College Mathematics Journal*, Fall 1990.
13. "Dependent Random Variables and the Central Limit Theorem," *International Journal of Mathematical Modeling*, Fall 1989.

14. *Enduring Positive Influence*. New York: Peter Lang University Studies Series, 1988.

15. "Certain Mathematical, Statistical, Logical and Philosophical Issues at the Foundation of Mathematical Modeling," *Mathematical Modeling in Science and Technology*, Pergamon Press, 1987.

16. "Toward a Joint Disciplinary Validation of Sigfluence," *Mathematical Modeling in Science and Technology*, Pergamon Press, 1987.

17. "Pure Thought in Its Relation to the Computer," invited article that led to editorship, *Thought*, Fall 1986.

18. "Sigfluence," *The Counselor*, Fall 1985.

19. "Sigfluence - One of Two Neologisms Emerging From Harvard's 1984 International Conference on Thinking," *New York Times*, Aug. 27, 1984.

20. "Does $\sqrt{2}$ Truly Exist?" *American Mathematical Society Abstract* (led to an invitation for Loase to deliver the Bertrand Russell Lecture at the Canadian Learned Societies' Annual Conference), Fall 1986.

21. "Statistical Dependence and the Central Limit Theorem," *American Mathematical Society* Abstract, Fall 1986.

Qualitative Research References

1. Giorgi, Amedeo ed. *Phenomenology and Psychological Research.* Pittsburgh: Duquesne University Press, 1985.

2. Greenbaum, Thomas. *The Handbook for Focus Group Research.* New York: Lexington Bokos, 1993.

SIGFLUENCE SURVEY PART I DR. JOHN LOASE

1. Check: Male_____, Female_____

2. How old are you today? _____

3. How many brothers and sisters did you grow up with? _____

4. What order are you in with respect to your brothers and sisters? _____

5. Check your educational background. No high school diploma _____, High school diploma _____, Two years college _____, Bachelor's degree _____, Master's degree _____, Doctorate _____.

6. What is your present occupation? _____

7. Check your present annual salary. Below $10,000 _____, $10,000-19,999 _____, $20,000-29,999 _____, $30,000-39,999 _____, $40,000-49,999 _____, $50,000+ _____.

8. Check your father's educational background. (See #5)_____

9. What is/was your father's occupation? _____

10. Check your mother's educational background. (See #5)_____

11. What is/was your mother's occupation? _____

12. What was your family's economic status when you first entered high school? Poverty _____, Lower middle class _____, Middle class _____, Upper middle class _____, Wealthy _____.

13. Who, outside of your family, had the most significant, long-term positive influence upon your life? (Can be left black; if left blank, go to question #19).

_____

14. What is/was the occupation of the influencer? _____

15. To what extent did the influencer impact your life? (Write a sentence or two)

_____
_____
_____

16. Under what circumstance(s) did the influencer impact your life? (Write a sentence or two)

_____
_____
_____

17. Why do you feel that this person's influence was significant upon your life? (Write a sentence of two)

_____
_____
_____

18. How long was it after the close contact before you knew that it was significant?

_____

19. a) In regard to your personal influence, write a sentence or two highlighting your major achievements.

_____
_____
_____

b) In regard to your personal influence, write a sentence or two highlighting your major disappointments.

_____
_____
_____

PART II For each of the following sentences, circle the response that would be most nearly true for you. The responses always extend from one extreme to its opposite. Please use the neutral rating as little as possible, since it means no judgment in either direction.

1. I usually have:

X X X X X X X X X X X

Negative impact on the people I meet | Neutral | Positive impact on the people I meet

2. Life is filled with a lot of possibilities for positive influence toward people.

X X X X X X X X X X X

Strongly agree Neutral Strongly disagree

3. My present or recent job has:

X X X X X X X X X X X

Little opportunity for positive influence towards people

Neutral

Has a lot of opportunity for positive influence towards people

4. My friends would say, if asked, that I have a positive influence on their lives.

X X X X X X X X X X X

Strongly agree Neutral Strongly disagree

5. Having positive personal influence is important to me.

X X X X X X X X X X X

Strongly disagree Neutral Strongly agree

6. My life has been satisfying.

X X X X X X X X X X X

Strongly disagree Neutral Strongly agree

7. In my life I have:

X X X X X X X X X X X

Helped a great many people — Helped some — Helped no one

8. In my present or recent job I have achieved:

X X X X X X X X X X X

Considerable positive influence — Some positive influence — No positive influence

9. In my present or recent job I have achieved:

X X X X X X X X X X X

No negative influence — Some negative influence — Considerable negative influence

10. My children are:

X X X X X X X X X X X

A source of considerable pain — A source of some pain — A source of no pain

11. My children are:

X X X X X X X X X X X

A source of considerable pleasure — A source of some pleasure — A source of no pleasure

12. In terms of helping others, I am capable of:

X X X X X X X X X X X

Considerable positive influence — Some positive influence — No positive influence

13. In terms of helping others, I am capable of:

X X X X X X X X X X X

Considerable negative influence — Some negative influence — No negative influence

14. My intimate relationships have been characterized by:

X X X X X X X X X X X

Considerable reciprocal harm — Some reciprocal harm — No reciprocal harm

15. My intimate relationships have been characterized by:

X X X X X X X X X X X

Considerable reciprocal benefit — Some reciprocal benefit — No reciprocal benefit

16. I have been told frequently by people that I have helped them:

X X X X X X X X X X X

Strongly disagree — Neutral — Strongly agree

17. The people who come into contact with me feel that they benefit from our interaction.

X X X X X X X X X X X

Strongly agree — Neutral — Strongly disagree

18. Life is a sequence of people influencing people.

X X X X X X X X X X X

Strongly disagree Neutral Strongly agree

19. "The whole world of loneliness, poverty, and pain make a mockery of what human life should be." (Bertrand Russell)

X X X X X X X X X X X

Strongly expresses my feeling Neutral Is just the opposite of my feeling

20. People who help the poor, like Mother Teresa:

X X X X X X X X X X X

I would like to use as models Neutral I would not use as models

21. The meaning in my life comes from the positive influence that I have contributed toward others.

X X X X X X X X X X X

Strongly disagree Neutral Strongly agree

22. Dr. Albert Sabin, who developed the oral vaccine that wiped out polio, is a person I would like to meet.

X X X X X X X X X X X

Strongly agree Neutral Strongly disagree

23. I would like to be in a position to increase the effectiveness of aid to starving people.

X X X X X X X X X X X

Strongly agree Neutral Strongly disagree

SIGFLUENCE SURVEY REVISED SCORING KEY
(January 1992)

Please compute your scores for these three sigfluence related constructs. The scale ranges from 1 to 11. R indicates to reverse the score, i.e. (1=11), (2=10), (3=9), (4=8), (5=7), (6=6).

A. Actual Sigfluence – To arrive at your total score, add your responses to items 4(R), 7(R), 8(R), 9(R), 10, 11(R), 14, 15(R), 16, 17(R).

B. Potential for Sigfluence – To determine your score, add your responses to items 1, 2(R), 3, 12(R), 13 and 18.

C. Awareness of Personal Need for Sigfluence – To compute this score, add your responses to items 5, 19(R), 20(R), 21, and 22(R).

Results of the Sigfluence Survey
(Gloria Rousseau - Iona College)

**** MULTIPLE REGRESSION****

Equation Number 1 Dependent Variable.. ACTUAL Actual Sigfluence

Variable(s) Removed on Step Number
20.. EDUCFATH Father's Educational Background

Multiple R	.54899
R Square	.30139
Adjusted R Square	.23486
Standard Error	13.50026

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	1651.22861	825.61430
Residual	21	3827.39639	182.25697

F = 4.52995 Signif F = .0231

----------Variables in the Equation----------

Variable	B	SE B	Beta	T	Sig T
STATUINF	.32819	.15406	.38859	2.130	.0451
EDUCMOTH	8.24107	3.92484	.38300	2.100	.0480
(Constant)	54.74189	11.69051		4.683	.0001

Variables not in the Equation

Variable	Beta In	Partial	Min Toler	T	Sig T
AGETHEN	.15377	.18390	.99909	.837	.4127
STATUPER	-.05737	-.06283	.83775	-.282	.7812
AGENOW	.20361	.23791	.95377	1.095	.2864
STATUFAT	-.23779	-.25645	.81253	-1.187	.2493
EDUCFATH	-.27721	-.27583	.69163	-1.283	.2141
STATUMOT	-.07048	-.07085	.70606	-.318	.7540
SALARPER	.12189	.13883	.90633	.627	.5378
STATUFAM	-.09973	-.09835	.68939	-.442	.6633
AGEPOSIT	1.8475E-03	.00217	.96470	.010	.9923

End Block Number 2 POUT = .100 Limits reached

November 13, 1989

Gloria Jasmine Rousseau
Modeling and Simulation

Coding the Sigfluence Survey

Record Columns	Answer from Survey	
1-3	ACTUAL	
4-5	POTENTIAL	
6-7	NEED	
8-9	Question #2.	Occupation of the Influencer
10-11		Status of the Influencer
12	Question #5.	Length of the Influence (1-9)
13	Question #6.	Length After (1-9)
14-15	Question #7.	Age Then
16	Question #8.	Tell the Influencer (1 Yes 0 No)
17-18	Question #9.	Respondent's Occupation
19-20		Status Job of Respondents
21-22	Question #10.	Age Now
23-24	Question #11.	Father's Occupation
25-26		Status Job of the Father
27	Question #12	(1-6)
28-29	Question #13	Mother's Occupation
30-31		Status Job of the Mother
32	Question #14	(1-6)
33	Question #15	Salary of the Respondent
34	Question #16	(1-6)
35	Question #17	Sex (0 Male 1 Female)
36	Question #22	(1-5)
37	Question #23	(1-9)
38	Question #24.	

JOB / STATUS TABLE

Accountant	29/70
Actor (Entertainer)	46/56
Architect	03/78
Artist	27/56
Small Businessman	13/56
Business Executive	13/56
Finance Manager	13/56
Chauffer/Doorman	90/19
Child Care	33/13
Clergy	01/62
Clerk	02/40
Coach	18/79
Computer Programmer	15/64
Counselor	26/76
Customer	38/69
Doctor	22/78
Engineer	32/75
Farmer	17/22
Hairdresser	21/27
Housewife/Housekeeper	14/35
Law Enforce/Police Dept.	16/37
Lawyer	23/90
Manager	10/45
Mechanical Technician	04/25
Electrical Technician	04/25
Miner	91/18
Musician	40/41
Nurse	25/47
LPN	51/21
Politician	28/60
Prisoner	41/13
Professor	12/79
Psychologist	06/81
Real Estate Broker	19/55
Receptionist	05/48
Reporter	20/70
Retiree	34/-
Sales	45/47
Secretary	05/48
Serviceman/Former Member	07/28
Social Worker	35/64
Student	31/-
Teacher Secondary	36/73
Teacher Elementary	36/69
Tradesman	11/21
Waiter	22/22

Variables in the Data File

Variable	Description	Columns
Actual	'Actual Sigfluence'	1-3
Potencia	'Potential for Sigfluence'	4-5
Need	'Need for Sigfluence'q	6-7
OccupInf	'Influencer's Occupation'	8-9
StatuInf	'Influencer's Job Status'	10-11
Contact	'Length of the Contact'	12
Contafte	'Length to Recognize Sigfluence Afterwards'	13
AgeThen	'Age When Influenced'	14-15
TellInfl	'Tell Influencer of Influence'	16
OccuPers	'Individual's Occupation'	17-18
StatuPer	'Individual's Job Status'	19-20
AgeNow	'Current Age'	21-22
OccupFat	'Father's Occupation'	23-24
StatuFat	'Father's Job Status'	25-26
EducFat	'Father's Educational Background'	27
OccupMot	'Mother's Occupation'	28-29
StatuMot	'Mother's Job Status'	30-31
EducMoth	'Mother's Educational Background'	32
SalarPer	'Individual's Salary'	33
EducPers	'Individual's Educational Background'	34
Sex	'Individual's Sex'	35
StatuFam	'Family's Economic Status'	36
BroSisNu	'Number of Brothers and Sisters'	37
AgePosit	'Age Position With Respect to Brothers/Sisters'	38

SPSS ACTIVITY ONE

CORRELATION, REGRESSION, AND HYPOTHESIS TESTING WITH A LARGE DATA SET, MEASURING COLLEGE STUDENTS' BELIEFS TOWARD MONEY AND MEANING

Please review and take the Marketing and Sigfluence Survey in Appendix A. The Survey was taken by 542 undergraduates at Concordia College - NY and Iona College. There were 104 responses or variables derived from the students' responses. For example, gender was scored 1 = female; 0 = male. Item 5 yielded seven variables - English spoken at home = 1; not spoken at home = 0.

We reduced the set of variables from 104 to 50 after looking at the correlation matrix of $(104)^2 = 10{,}816$ correlations, looking for key clusters of variables that were statistically significant, and relying on the experience of Dr. Teresa Piliouras (data mining) and Dr. Loase (statistics - sigfluence measures). We then mapped each data set to the interval (0, 1) which enabled us to graph pairs of variables to discover geometric relationships.

Please now install SPSS on your computer and follow these steps:

1) Go to SPSS Data Editor

2) Go to File

3) Go to Open

4) Go to Data

5) Go to SPSS 50 var final - the data disk supplied with this book.

6) Press Open

7) You now have the 542 row by 50 variable data set that you can research.

Please analyze elementary statistics of several data sets to begin. For example, go to NORMACT and then go to

1) Analyze

2) Descriptive Statistics

Go to the Crosstabs index. Crosstabs will form two-way tables and can reveal relationships, such as gender or income differences in the selected variables. You could explore which variables demonstrate higher levels of satisfaction with life (Item 6). Our main finding in 2007 was that of the thousands of hypotheses we explored from this data set, the sigfluence variables (Part II of the survey) had greater explanatory power than the variables from Part I, the marketing dimension.

There are millions of hypotheses to be tested and future articles and books could be written from statistical analysis of this extensive data set. Again, there are $50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 = 254{,}251{,}200$ sets of 5 variables to explore 50 P 5, if order of analysis is significant. Next, let us move from Crosstabs to correlational analysis

Correlation

Let us start by correlating health with education.

1) Go to Analyze

2) Go to Correlate

3) Go to Bivariate

4) Click Health

5) Click ▶ to Variables

6) Click Education

7) Click ▶ to Variables

8) Click Pearson (correlation)

9) Click One Tailed

This is because you hypothesize health correlates positively with education. If you simply wanted your hypothesis to be that the variables are significantly positively or negatively correlated, you should click two-tailed. Click Flag significant correlation.

10) Click OK

The correlation R = .07, was not statistically significant. This is different from the result from Harvard's longitudinal study over several decades that discovered EDUCATION IN YEARS as a significant positive correlate to long, healthy life. The discrepancy is due to the restricted age range, college students 18-25, who are generally in good/excellent health. Also, education is limited to a narrow range, because the 542 respondents were all undergraduate college students.

It took twenty years of research to identify this cluster of important correlations. Now let us share the finding. We want to correlate Item 6 (My Life Has Been Satisfying), Actual Sigfluence, Need for Sigfluence, and Potential for Sigfluence; 4 variables with themselves. This will yield 16 correlations but only 4 C 2 = 6 correlations that are non-trivial.

Go to the SPSS file and perform the following steps:

1) Press Analyze

2) Correlate

3) Bivariate

4) Press Quest 6, NORMACT, NORMPOT, NORMNEED

5) Press ▶ The 4 variables should appear in variables.

6) Press OK

Write down the correlations that have two stars (**). These correlations are statistically significant at the .01 alpha level.

Question 6 has a .435 correlation with Actual Sigfluence, a .476 correlation with Potential for Sigfluence, and .377 correlation with Need for Sigfluence. Potential, Actual and Need are also significantly correlated. This suggests that satisfaction with life is associated with sigfluence. In a remarkable result, Potential for Sigfluence is the strongest correlation with Satisfaction with Life. This suggests that society has to focus on giving more opportunities for people to believe they have the potential to effect lasting positive influence.

Another important direction of our research was the influence of gender. Let us explore the difference between men and women on these 4 variables.

To explore whether women have higher scores in Actual Sigfluence, follow these steps:

1) Click NORMACT

2) Click ▶ NORMACT should be in Test Variable(s)

3) Click GENDER

4) Click ▶ into Grouping Variable. GENDER should be in Grouping Variable

5) Click Define Groups

6) Group 1 - Type 1 (Code for Female)

7) Group 2 - Type 0 (Code for Male)

8) Click Continue

9) Click OK

10) Observe the t value of t = 4.463 (df = 540) and significance level of .000. This confirms that women report significantly higher levels of Actual Sigfluence than men. This research result should lead to societal strategies to increase men's perceptions and actions related to sigfluence.

Now return and test the other 3 variables. You will find that women score significantly higher in each comparison. We need to help men increase the Potential, Need and Actual Sigfluence in coming years. Please read *The Sigfluence Generation: Our Young People's Potential to Transform America* for a fuller explanation of the discoveries you are now sharing. *The Sigfluence Generation* is free on my website sigfluence.com and won a second place award in the recent national Benjamin Franklin Book Contest for non-fiction.

Please explore the data set. Test your own hypotheses with basic two sample hypothesis tests and correlation. If you are ambitious, print a copy of the 50 x 50 variable correlation matrix. This explores the correlation of each of the 50 variables with themselves. Naturally, half are redundant because the correlation between x and y equals the correlation between y and x. Also there are 50 perfect correlations that add little - the correlation of x with x = 1.

At the end of Chapter Three, the next SPSS activity is written for intermediate level of statistical analysis and features multiple regression, analysis of variance and goodness of fit exercises with the enclosed data set.