

APPENDIX A
MARKETING AND SIGFLUENCE SURVEY

Part I – General Information

1. Gender:
 - ☐ Male
 - ☐ Female
2. What is the zip code where you currently live? _____
3. What is your year of birth? _____
4. What is your ethnic background?
 - ☐ Caucasian
 - ☐ African-American
 - ☐ Asian/Pacific Islander
 - ☐ American Indian/Alaskan
 - ☐ Spanish/Hispanic/Latino
 - ☐ Decline to answer
 - ☐ Other. Please specify _____
5. What language(s) do you speak at home?
 - ☐ English
 - ☐ Spanish
 - ☐ French
 - ☐ German
 - ☐ Russian
 - ☐ Chinese
 - ☐ Other. Please specify _____
6. What is the condition of your general health?
 - ☐ Excellent
 - ☐ Good
 - ☐ Fair
 - ☐ Poor

7. What was your family's general economic status while you were economically dependent on them?
- ☐ Affluent
 - ☐ Upper Middle Class
 - ☐ Middle Class
 - ☐ Lower Middle Class
 - ☐ Poverty-stricken
8. What is the highest level of education that you have completed?
- ☐ Some high school
 - ☐ High school graduate or equivalent
 - ☐ Some college
 - ☐ Associate's degree
 - ☐ Bachelor's degree
 - ☐ Master's degree
 - ☐ Doctorate
9. In school, would you consider yourself to be an:
- ☐ An "A" average student
 - ☐ A "B" average student
 - ☐ A "C" average student
 - ☐ Below "C" student
10. What is your present occupation?
-
11. Current job status:
- ☐ Working full time
 - ☐ Working part time
 - ☐ Full time student
 - ☐ Not employed
12. What was/is the strongest influence in your choice of a career?
- ☐ Parents
 - ☐ Money
 - ☐ Status
 - ☐ Intellectual stimulation
 - ☐ Other. Please specify _____
13. What is your total annual household income?
- ☐ Less than \$25,000
 - ☐ \$25,000 - \$50,000
 - ☐ \$50,000 - \$75,000
 - ☐ \$75,000 - \$100,000
 - ☐ \$100,000 +

14. What kind(s) of bank account(s) do you currently have? Check all that apply.
- ☐ Personal checking account
 - ☐ Personal savings account
 - ☐ Business checking account
 - ☐ Business savings account
 - ☐ None of the above
 - ☐ Other. Please specify _____
15. Given the current market, what would you feel the most comfortable investing in?
- ☐ Property
 - ☐ Money market fund
 - ☐ An aggressive growth mutual fund
 - ☐ A conservative growth mutual fund
 - ☐ Certificate of deposit (CD)
 - ☐ U.S. savings bond
 - ☐ Other. Please specify _____
16. Where do you make your purchases? Check all that apply.
- ☐ At the store
 - ☐ Online
 - ☐ From a wholesaler (Costco Wholesale, BJ's Club, etc.)
 - ☐ Other. Please specify _____
17. Do you currently own a credit card, and if so, what kind of credit cards to you have? Check all that apply.
- ☐ Visa
 - ☐ Mastercard
 - ☐ American Express
 - ☐ Diner's Club
 - ☐ Debit card
 - ☐ Store card (Macy's, etc.)
 - ☐ None
18. How many cars does your household presently have?
- ☐ 0
 - ☐ 1
 - ☐ 2
 - ☐ 3
 - ☐ 4+

19. Other than the basic necessities such as rent/mortgage, car payments/insurance, and tuition, what do you spend the most amount of money on? Number and select the top three.

- ☐ Entertainment
- ☐ Electronic devices
- ☐ Clothing and shoes
- ☐ Cosmetics/hair supplies
- ☐ Fitness club fees
- ☐ Car accessories
- ☐ Other. Please specify _____

20. How would you prefer to spend your spare time? Number and select the top three.

- ☐ Shopping
- ☐ Reading a book/magazine
- ☐ Working out
- ☐ Hiking
- ☐ Going out to a bar/night club
- ☐ Going to church/church activities
- ☐ Sleeping
- ☐ Hanging out with family and friends
- ☐ Other. Please specify _____

21. Who would you consider to have had the most significant, long-term, positive influence on your life?

- ☐ Parents
- ☐ Siblings
- ☐ Grandparents
- ☐ Teacher
- ☐ Friend
- ☐ Other. Please specify _____

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:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
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.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly agree Neutral Strongly disagree
3. My present or recent job has:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
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.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly agree Neutral Strongly disagree
5. Having positive personal influence is important to me.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly disagree Neutral Strongly agree
6. My life has been satisfying.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly disagree Neutral Strongly agree
7. In my life I have:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
Helped a great many people Helped some Helped no one

8. In my present or recent job, I have achieved:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Considerable positive influence Some positive influence No positive influence
9. In my present or recent job, I have achieved:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 No negative influence Some negative influence Considerable negative influence
10. My children are:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 A source of considerable pain A source of some pain A source of pain
11. My children are:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 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:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Considerable positive influence Some positive influence No positive influence
13. In terms of helping others, I am capable of:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Considerable negative influence Some negative influence No negative influence
14. My intimate relationships have been characterized by:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Considerable reciprocal harm Some reciprocal harm No reciprocal harm
15. My intimate relationships have been characterized by:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Considerable reciprocal benefit Some reciprocal benefit No reciprocal benefit

16. I have been told frequently by people that I have helped them.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Strongly disagree Neutral Strongly agree
17. The people who come into contact with me feel that they benefit from our interaction.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Strongly agree Neutral Strongly disagree
18. Life is a sequence of people influencing people.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Strongly disagree Neutral Strongly agree
19. "The whole world of loneliness, poverty, and pain makes a mockery of what human life should be." (Bertrand Russell)
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Strongly express my feeling Neutral Is just the opposite of my feeling
20. People who help the poor, like Mother Teresa:
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 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.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 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 model.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 Strongly agree Neutral Strongly disagree
23. I would like to be in a position to increase the effectiveness of aid to starving people.
☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
 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).

If you have no children, use the neutral rating of 6 for items 10 and 11.

Now obtain your percentile score from the table of norms that immediately follows:

ACTUAL SIGFLUENCE NORMS
(n = 282)

<u>Data Interval</u>	<u>Cumulative Proportion</u>
33 – 39	0.004
40 – 46	0.011
47 – 53	0.021
54 – 60	0.057
61 – 67	0.191
68 – 74	0.323
75 – 81	0.482
82 – 88	0.660
89 – 95	0.833
96 – 102	0.929
103 – 109	0.972
110 – 116	0.996
117 – 123	1.000

To use this table, locate your score on the left. Use the cumulative proportion on the right corresponding to your score. For example, if you scored 86 in Actual, the cumulative proportion is 66%. Your percentile score is approximately 66%, a little less since a score of 88 (the end point of the interval) corresponds to the 66th percentile. A score of 66 percentile means that 34% of the sample scored higher than you and 66% scored at your level or below.

POTENTIAL FOR SIGFLUENCE NORMS
(n = 282)

<u>Data Interval</u>	<u>Cumulative Proportion</u>
29 – 31	0.011
32 – 34	0.014
35 – 37	0.032
38 – 40	0.082
41 – 43	0.167
44 – 46	0.294
47 – 49	0.394
50 – 52	0.564
53 – 55	0.691
56 – 58	0.830
59 – 61	0.954
62 – 64	0.996
65 – 67	1.000

NEED FOR SIGFLUENCE NORMS
(n = 282)

<u>Data Interval</u>	<u>Cumulative Proportion</u>
10 – 13	0.004
14 – 17	0.014
18 – 21	0.021
22 – 25	0.039
26 – 29	0.117
30 – 33	0.277
34 – 37	0.475
38 – 41	0.663
42 – 45	0.805
46 – 49	0.926
50 – 53	0.986
54 – 57	1.000

RECOMMENDED FOLLOW UP ACTIVITIES

1. The Sigfluence Generation: Our Young People's Potential to Transform America is free for download at my website, sigfluence.com. The book was awarded a Silver Medal for non-fiction in the 2012 Benjamin Franklin Book Contest. The book relied on over 20 years of Statistics and Mathematical Modeling to discover the positive transformational potential of our young people.

If you have not as yet obtained the 50 variable SPSS data file to use for your own research, please email me at John.Loase@concordia-ny.edu requesting the file.

For a refresher in Statistics, please turn the page. I included half of my recently published book, *Statistics Made Easy* (Graduate Group, 2010). You may wish to purchase the text from Graduate Group if you want a more extensive real-world introduction to each topic.

Now that you have taken the Marketing and Sigfluence Survey, you may have questions about the validity and reliability of your three sigfluence scores - Actual, Potential and Need. An entire book, *Theory and Measurement of Sigfluence* (University Press of America, 2002) by Loase, is devoted to the statistical foundation of the Sigfluence Survey.

Other works that may be of interest in both books and film include:

Our Neglect, Denial and Fear - Nova Science (Kroshka Books) - 2000

Sigfluence: The Key to Our Earthly Immortality - 1997 - Bluebird.

Sigfluence: The Key to It's a Wonderful Life - University Press of America - 1996.

Sigfluence: Long-Term, Positive Influence - University Press of America - 1994

Sigfluence: Enduring, Positive Influence - Peter Lang University Studies Series - 1988.

2. Film - *Wild Strawberries* - Bergman
Ikiru - Kurosawa
3. Psychology - *Man's Search for Meaning* - Frankl
Man and His Symbols - Jung
Memories, Dreams and Reflections - Jung
4. Literature - *Death in Venice* - Mann

5. Language - Language - Sapir, E. Harcourt Brace - 1921
6. Similar Finding - God on the Quad - Naomi Riley, St. Martins, 2005

SHARE YOUR REACTIONS. John Loase

Dr. John Loase/Professor of Mathematics
Concordia College
171 White Plains Rd.
Bronxville, NY 10708
email: John.Loase@concordia-ny.edu

APPENDIX B

A TI-83 BASED PRIMER ON BASIC STATISTICS

CHAPTER ONE - ELEMENTARY PROBABILITY

Probability is a concept that you already have familiarity with. The probability of a head in a coin flip is $\frac{1}{2}$. The probability of someone running a mile under one minute is near 0. The probability that having a good financial plan early will benefit you is nearly 100%.

Properties of Probability

1. The probability of an event lies between 0 and 1. This is written $0 \leq P(E) \leq 1$.

The \leq sign means that the probability could equal 0 or 1.

2. The sum of probabilities of all possible outcomes of an event equals 100% or 1.
3. The probability of an event not happening is $1 - P(E)$.

Sixty percent of baby boomers plan to retire early. If you asked a baby boomer if he/she plans to retire early, what do you think the chances are that they do not plan to retire early?

Well, this is a simple application of probability. $1 - 60\% = 100\% - 60\% = 40\%$.

Surveys show that two-thirds of baby boomers felt they could not invest for the long-term. What is the probability that your boomer acquaintance is investing for the long-term? Of course, the probability is $1 - \frac{2}{3} = \frac{1}{3}$. Things are pretty dismal unless you start a systematic, early retirement plan.

The median annual salary for Money Magazine readers is \$84,000 per year. This means 50% make more; 50% make less. The median is the middle number if salaries (or any set of

data) is arranged in order (high to low or low to high). What is the probability that a reader picked at random makes less than \$84,000? (50%).

What is the probability that a reader picked at random makes more than \$84,000? $(1 - \frac{1}{2}) = 50\%$.

Definition

The way we frequently obtain probability is through frequencies. We count the number of times an event occurs and divide by the total number of events.

$$P(E) = \frac{\text{number of times E occurs}}{\text{number of trials}}$$

Usually, we are estimating more complicated events than coin flips. It is hard to get a handle on the true probabilities. As a result, we count.

For example, to obtain our median salary of Money readers of \$84,000 per year, we count the number of readers making more than \$84,000, say one million, and divide by total readership, say 2 million.

$$\frac{1 \text{ million}}{2 \text{ million}} = \frac{1}{2} = 50\%$$

Racial disparities are expected to decline in the future with respect to pension. In 1990, 51% of whites received a pension versus 32% of African Americans. By 2030 the gap should close to 83% of whites - 79% of African Americans.

In 2030 what is the probability that a random African American will not have a pension?
 $(1 - 79\%) = 100\% - 79\% = 21\%$

Consider the risks of inflation on a baby boomer planning to retire at age 50. Look at the table below.

PLAN TO BEAT THE AVERAGES

How long should you expect your assets to last once you have retired? Choosing the average life expectancy means you have a 50 percent chance of outliving your money. The table below is a safer benchmark. Your probability of outliving these figures was only 20 percent in the most recently published government life tables. To be really safe, add a couple more years.

At age.....	Plan to live another..... Female	Male
45 yrs.	46 yrs.	43 yrs.
50	41	38
55	36	32
60	31	27
65	27	22

Source: National Center for Health Statistics Data for 1991

If you are a 50 year old female baby boomer, your chances of living 41 years or more is roughly 20%. If you are a man, your chances of living more than 38 years is roughly 20%. It is a good idea to use tables that give you a high probability (80%) of having enough money in retirement.

Look at the table below for an analysis of what the effect is of 3%, 5% and 10% inflation on \$1 over 10, 20, 30 and 40 years.

YEARS	3%	5%	10%
10	1.34	1.63	2.59
20	1.81	2.65	6.73
30	2.43	4.32	17.45
40	3.26	7.04	45.26

If your 50 year old baby boomer retires, he/she has to expect to live (with approximately 80% probability) forty more years. The retirement income that he/she starts with has to be divided by 3.26 to determine the real income he/she will have at age 90 if inflation increases by

3% per year. For example, if you retire with a comfortable \$70,000 retirement income at age 50, by age 90 you will have a real income of $\$70,000 \div 3.26 = \$21,472$.

As you can see, most 90 year olds will be struggling with \$21,472 in retirement income (in buying power after 40 years). If you plan young in life, you reap great benefits later in life, as we learn from mathematics of finance and specifically the power of compound interest and annuities. This material is covered in detail in Statistics Made Easy (Graduate Group, 2010).

Homework Probability (Answers to chapter homework are at the end of Chapter 7.)

1. What is the probability of living less than 46 years if you are a female retiring at age 45?
2. What is the probability of living an additional 27 years or more if you are a male retiring at age 60? What is the probability of living less than 27 years for this same person?
3. Using the data from the New York Times article "Vast Advance is Reported in Preventing Heart Illness," August 6, 1999:
 - a) Find the probability of dying of cancer in 1997.
 - b) Find the probability of dying of cancer in 1979.
 - c) Find the probability of dying from heart disease in 1950.
 - d) Find the probability of dying from heart disease in 1996.
4. In July, 1999 the unemployment rate was 4.3%. What percent of the American population of workers was employed.

EXPERIENCE 1

Interview your parents about their plans for retirement. Help them plan for their financial needs using the charts from this chapter. Also discuss with them the knowledge you have gained

from reading The Millionaire Next Door. What have you learned? If it is not convenient to interview your parents, make up a fictitious account.

CHAPTER TWO - STANDARD DEVIATION/VARIANCE

Suppose you studied the want ads for accountants for a week and observed five jobs with starting salaries (in thousands) of \$40, \$35, \$40, \$50 and \$35. The range between the high salary of \$50 and \$35 is $50 - 35 = \$15\text{K}$. Also each salary is fairly close to the mean. To calculate the mean, add the five numbers and divide by 5.

$$\frac{40 + 35 + 40 + 50 + 35}{5} = \frac{200}{5} = \$40\text{K}$$

As you can see, the salaries are fairly close to \$40K.

Now suppose you observed these five starting salaries for accountants: \$100, \$20, \$15, \$30, \$35. The mean would still be the same \$40K. But there would be a larger average deviation from the mean. The concept of deviation from the mean is the essence of variance or standard deviation.

The variance (σ^2) of a population (the entire data set of a statistic of interest) is defined as:

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

n = size of the sample
 X = individual score
 \bar{X} = population mean

We rarely know all the values of the population. In fact, much of statistics is based on estimating the mean of the population based on a small sample ($n < 30$) or large sample ($n \geq 30$).

The sample variance, written as S^2 , is the sum of the squares of deviations from the mean, divided by $n - 1$.

$$S^2 = \frac{\sum (X - \bar{X})^2}{(n - 1)}$$

The sample standard deviation, S , is the square root of the variance.

Let us go back to our first example, where the five numbers for salary are: \$40, \$35, \$40, \$50, \$35. $\bar{X} = 40$.

You can use a table to simplify your calculations:

X	$X - \bar{X}$	$(X - \bar{X})^2$
40	40-40	$0^2 = 0$
35	35-40	$(-5)^2 = 25$
40	40-40	$0^2 = 0$
50	50-40	$10^2 = 100$
35	35-40	$(-5)^2 = 25$
$\Sigma X = 200$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 150$

This always equals 0

$$\text{The sample variance} = \frac{150}{5-1} = \frac{150}{4} = 37.5$$

To obtain the sample standard deviation, simply take $\sqrt{37.5} = 6.1$.

Usually, the larger standard deviation means greater variability within the sample. To compensate for samples with large numbers, like major league baseball players' salaries, you could use a statistic called the coefficient of variation:

$$\text{Coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}} \times 100\%$$

To illustrate, use the standard deviation we have just calculated of 6.1, mean = 40. The coefficient of variation for this sample is:

$$\frac{6.1}{40} \times 100\% = 15\%$$

This 15% means fairly low variation.

There is a short-cut formula that eliminates the calculation of the $(X - \bar{X})^2$. We will use this formula, which is:

$$\text{SHORT CUT FORMULA} \quad S^2 = \frac{n \Sigma X^2 - (\Sigma X)^2}{n(n-1)}$$

You will get slightly different answers for the variance depending upon whether you use the population or sample variance formula. If you take three semesters of calculus and a calculus based probability and statistics course, you will find out why. You will also understand the proofs for the formulas in the primer and become facile in an ever widening field of critical importance to our nation - mathematical modeling.

Rounding errors are frequently an issue in statistics. Drs. Jay Devore and Roxy Peck recommend using four or five digits of decimal accuracy beyond the decimal accuracy of the data values themselves (Introductory Statistics, 1994, p. 83).

Let us calculate our variance using the shortcut formula of the salary data:

$$n = 5 \text{ (salaries)}$$

$$\Sigma X^2 = 40^2 + 35^2 + 40^2 + 50^2 + 35^2 = 8150$$

$$\Sigma X = 40 + 35 + 40 + 50 + 35 = 200$$

$$(\Sigma X)^2 = 40,000$$

$$\text{Note: } \Sigma X^2 \text{ does not equal } (\Sigma X)^2$$

$$\text{Variance (sample)} = S^2 = \frac{5(8150) - 40,000}{5(4)} = 37.5$$

$$\text{Standard deviation (sample)} = S = \sqrt{37.5} = 6.1$$

Let us try another example of variance. Consider the cost of a three bedroom house in the city you plan to work. Your largest expense is usually your house mortgage and taxes. This is vital statistics. You need to understand your finances and how to estimate your house expenses. Consider downsizing and moving to a highly rated city like Saranac Lake, New York. Houses are relatively inexpensive in Saranac Lake. On any given day you might have a wide

variety of houses for sale in the \$60,000 - \$350,000 range. Most sell between \$60,000 and \$150,000 - quite a significant difference from high priced areas such as Westchester.

It is essential that you subscribe to the local newspaper of the town in which you plan to move. You can look at crime, job openings, and our immediate concern - house prices. The Adirondack Daily Enterprise on August 6, 1999 advertised homes for sale. Their prices (for the immediate Saranac Lake area) were \$345,000, \$215,000, \$145,000 and \$55,000. We can leave out the thousands and consider this housing sample of four as \$345, \$215, \$145, and \$55 K.

Let us use our TI-83 calculators to compute the variance and standard deviation of this sample of four house prices. Here are the steps to enter our data:

- 1) Press 2nd.
- 2) (
- 3) 345, 215, 145, 55
- 4) Press 2nd
- 5))
- 6) Press STO
- 7) Press 2nd
- 8) Press 1
- 9) Press Enter

Your data is stored in List 1. Now press

- 1) 2nd List
- 2) Use the ▲key to go to highlight MATH
- 3) Use the ▼key to go to Std Dev
- 4) Press Enter

5) Press 2nd 1

6) Press Enter

The answer is 122.3. The standard deviation is 122 if rounded to the nearest whole number (thousand).

Now use a similar process to obtain the mean.

The answer is 190 K. You have just seen how easy it is to analyze house prices. The mean house price was \$190K with standard deviation of \$122. The coefficient of variation was:

$$\frac{\text{S. Dev.}}{\text{Mean}} \times 100\% = \frac{122}{190} \times 100\% = 64\%$$

This is a very high coefficient of variation. The house prices are very widely variable. This was due to a small sample ($n = 4$) and unusually expensive house prices. Since I had sampled dozens of houses last year for an earlier book, I know that this one day sample of four does not represent the true house price situation in Saranac Lake. You should analyze your prospective area by taking a two month sample of house prices and preferably the actual selling price - not the price advertised in the newspaper or quoted by real estate agents.

Homework

1. Find the standard deviation, variance, and coefficient of variation for this sample of teacher salaries: \$30, \$40, \$36, \$28, \$42 K. Use both the definition formula and shortcut formula to compute variance.
2. Use your TI-83 calculator to confirm your answer to #1.
3. Your six stocks went up (in %) 0, 7, 8, 6, 10, and 28 last year.
 - a) Find the mean % gain for your six stocks

- b) Find the mean, variance, standard deviation, and coefficient of variation.

Use two different formulas to compute the variance.

4. Physician salaries were sampled and found to be \$110, \$150, and \$135 K. Find the mean, variance, standard deviation and coefficient of variation three different ways.

EXPERIENCE 2

Obtain a sample of at least twenty house prices and at least 20 professional salaries in your city of prospective relocation. Use a month or two months of samples from the local newspaper to obtain your data. Compute the mean, variance, standard deviation, and coefficient of variation for both samples. (Keep the data separate. Do not mix house prices and salaries.)

CHAPTER THREE - SAMPLING METHODS/CONFIDENCE INTERVALS

Most statistical analyses want to make an estimate of some population characteristic - for example the average salary you might make as a starting teacher in Mariana, Florida or Saranac Lake, New York. It is inefficient to ask every starting teacher from the previous year about their salary. It would be hard to find every one, very costly, and a tortuous process. The better way is to take a sample of say 30 or 60 or 100 starting salaries in the area. You can use that data to come up with an interval that with 95% confidence captures the true population mean (99% confidence if your are satisfied with a larger interval - say \$24,000 - \$46,000 as contrasted to a 95% confidence interval that might look like \$27,000 - \$43,000).

In order to ensure that your sample is well constructed, it must have certain characteristics, including:

1. Randomness

If we went to Greenwich (a wealthy town) to look at average house prices for the Northeast, we would have little evidence to draw valid conclusions. Greenwich is not representative of the Northeast. Its houses are among the highest priced in the nation.

You cannot make sound inferences as to the average house prices in the Northeast, unless you have a sample that is representative of the Northeast. You need a random sample.

A random sample means that each house in the Northeast has an equal chance of being picked. You could somehow put each index card with a house price into a very big hat, shuffle the cards well, and pick say 1000 cards. The key is to try to make your sample as random as humanly possible. You can never achieve perfect randomness, but you can easily eliminate glaring biases. If a sample is not random, it is biased.

If your population can be matched with numbers, you could use a random number generator from a standard calculator or computer to help you select your sample. Unfortunately, computer generated random numbers repeat every K times, so they are not random. But K can be a very large number if you use a well constructed random number generator.

2. Stratified Sampling

Sometimes you can break up your population into a set of n subpopulations, that we call strata. For example, you could break the Northeast into say 200 counties and take a random sample of each county house price.

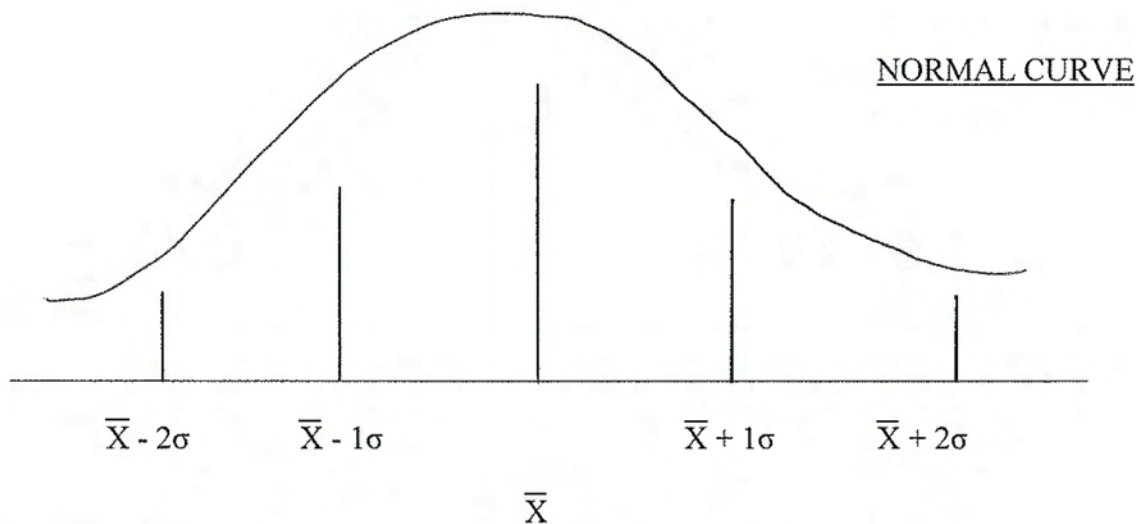
Stratified sampling usually ensures that you have thought about the subgroups that make up your population and included each as part of your sample. For example, if there are three different subgroups of teachers you wish to study, for example private school, public school and college, you should include each subpopulation in your sample. Otherwise you might limit your sample to public school teachers, enter private school teaching, and find that you had overestimated your starting salary. After all, public school teachers generally make significantly more than private school teachers. Public defenders make much less than corporate lawyers. You shortchange yourself if you do not try to achieve a random sample with the underlying strata proportionately represented.

Confidence Intervals

Before we compute a confidence interval, there are two powerful concepts that should be introduced:

1. The normal curve - this curve is a basic and remarkable foundation for statistics.

The curve below could be approximated by the graph $y = k \cdot 3^{-x^2}$. This is a poor approximation, but at least you get some idea of what the algebra is behind the normal curve below.



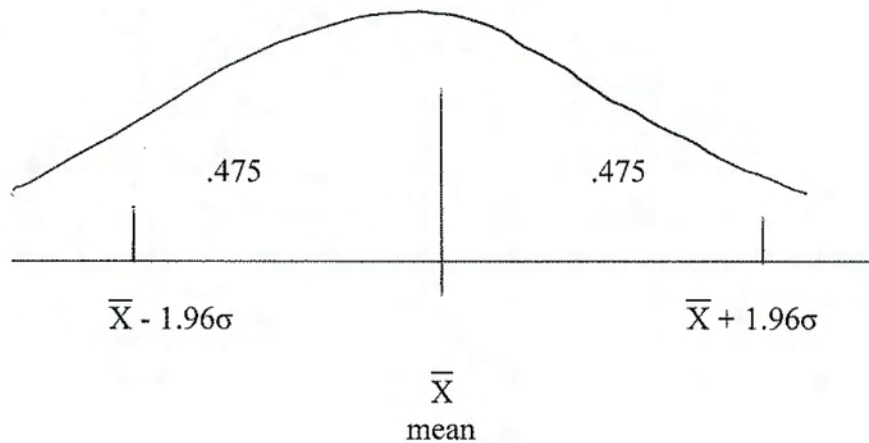
The probability of a value lying between the mean \bar{X} and positive infinity is 50%. The corresponding probability of a value lying between the mean \bar{X} and $-\infty$ is 50%.

The area under the normal curve is $100\% = 1$. This doesn't help us very much. What helps us is the fact that we can look up the probability of a value lying between any two values if they are expressed in standard form, Z measures where:

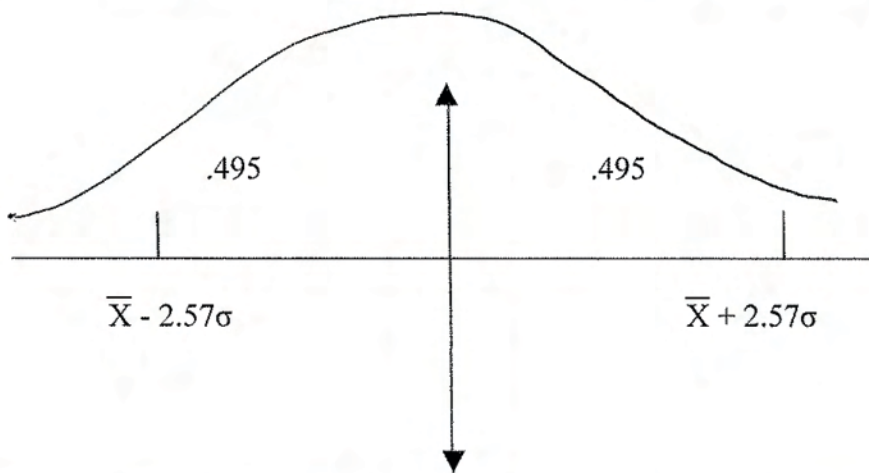
$$Z = \frac{X - \mu}{\sigma} \quad X = \text{value}; \mu = \text{population mean}; \sigma = \text{standard deviation}$$

The normal curve can be teamed with the central limit theorem to help us achieve a great many powerful statistical ideas. The central limit theorem tells us that when sample size is

sufficiently large, say $n \geq 30$, the samples of \bar{X} (mean) can be approximated with probabilities from the normal curve. We will just use two normal curve probabilities below, the probability of being between the mean - 1.96 (standard deviations) and the mean +1.96 (standard deviations). The probability of a value falling in this interval is 95%. See the graph below (.475 + .475 = 95%)



The probability of a value lying between $\bar{X} + 2.57\sigma$ and $\bar{X} - 2.57\sigma$ is 99%. See the graph below.



Even if the population from which you are sampling is not a normal distribution, you can always use this approximation. The normal curve is an amazingly generalizable and powerful invention, one of the most useful of the twentieth century.

We now can apply the central limit theorem and use the normal curve for our next result. If $n \geq 30$, the sampling distribution of \bar{X} (mean) can be described by a normal curve with mean $\mu_x = \mu$ and standard deviation σ/\sqrt{n} .

This gives us our two powerful formulas for large samples ($n \geq 30$).

95% Confidence Interval for Population Mean

$$\bar{X} \pm 1.96 (\sigma/\sqrt{n})$$

And 99% Confidence Interval for Population Mean

$$\bar{X} \pm 2.57 (\sigma/\sqrt{n})$$

Let's use these two results to calculate 95% and 99% confidence intervals for your home price if you move to Saranac Lake. Assume that you have inspected the Adirondack Daily Enterprise for several months and written down 50 house prices. You need to find the mean, say \$95,000. You next need to find the standard deviation, say \$25,000.

We can now calculate the 95% and 99% confidence intervals for the average home price in Saranac Lake. To calculate the 95% confidence interval,

$$95\% \text{ C.I.} = \bar{X} \pm 1.96 (\sigma/\sqrt{n}) = \$95,000 \pm 1.96 (25,000/\sqrt{50})$$

$$95\% \text{ C.I.} = 95,000 \pm 1.96 (25,000/7.07107)$$

$$95\% \text{ C.I.} = 95,000 \pm 6929.644311$$

$$95\% \text{ C.I.} = (95,000 - 6929.64\dots, 95,000 + 6929.64\dots)$$

Now round to the nearest thousand at the end.

$$95\% \text{ C.I.} = (88,000, 102,000)$$

You can assume that your home in Saranac Lake will cost somewhere between \$88K and \$102K (with 95% confidence). The reason that you rounded to the nearest thousand at the end was that all your house prices were in thousands. If you ended up with a confidence interval that

had house prices to the nearest penny, you would give people the illusion that you could estimate house prices much more accurately than you can. After all, most house prices rise or fall by tens of thousands or more of dollars as a result of price changes or negotiations. Always think of your audience and try to use statistics that make sense and reflect reality.

Now let us use the same data to compute a 99% confidence interval for home prices. This has to be a wider interval because you have more probability of being correct.

$$99\% \text{ C.I.} = \bar{X} \pm 2.57 (\sigma/\sqrt{n})$$

$$99\% \text{ C.I.} = 95,000 \pm 2.57 (25,000/7.07107)$$

$$99\% \text{ C.I.} = 95,000 \pm 9086.319326$$

Round to thousands

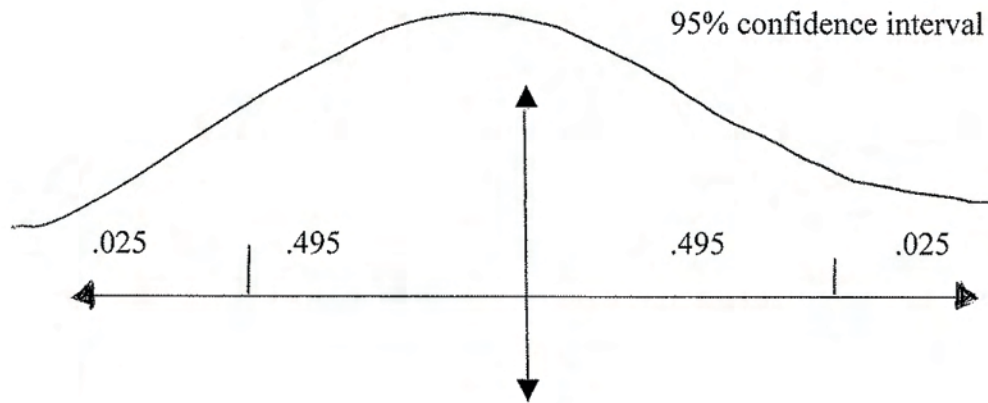
$$99\% \text{ C.I.} = 95,000 \pm 9,000$$

$$99\% \text{ C.I.} = (86,000, 104,000)$$

You now have 99% confidence that your house will cost between \$86K and \$104K. If you wish more than 99% confidence, it usually is not practical. For example, you could be 100% confident that your house will cost somewhere between \$0 and \$1 billion. But it is difficult to plan with this type of confidence interval.

2. Small Sample Confidence Intervals

Frequently, we have to make analyses with smaller samples. If your sample is smaller than 30, you need to change the Z value of 1.96 or 2.57 to the corresponding value from a t-distribution. To determine the t value, use a t table and look up the value by determining df and p. The df (degrees of freedom) is (n-1), where n is sample size. Your p value is .025 (for 95% confidence) or .005 (for 99% confidence). Look at the graph below.



You now see the reason for relying on a p-value of .025 from the t-table to obtain a confidence interval (95%). There is 2.5% probability of landing in either the extreme left tail (house price below \$88K) or in the extreme right tail (house price above \$102K).

Let us calculate a 95% confidence interval for your starting salary if you have a sample of ten accountants' salaries from Saranac Lake and they are (in thousands):

22, 24, 18, 26, 28, 20, 26, 25, 29, 30

Calculate \bar{X} , σ using the data editor of the TI-83.

$$\bar{X} = 24.8$$

$$\sigma = 3.9$$

For 95% confidence, use $n = 10$, $df = (n - 1) = 10 - 1 = 9$

$$t_{.025, 9df} = 2.26$$

$$95\% \text{ C.I.} = 24.8 \pm 2.26 (3.9 / \sqrt{10})$$

$$95\% \text{ C.I.} = 24.8 \pm 8.814/3.162 = 24.8 \pm 2.787$$

$$95\% \text{ C.I.} = (22K, 28K)$$

3. Central Limit Theorem

Fun exercise: Let us use the random number program on your TI-83 to glimpse the genius of the central limit theorem. We are going to see the concept of the binomial density function. For a binomial random variable, we need something like a coin flip with two possible outcomes, heads or tails. The probability of a head does not change from coin flip to coin flip. We want to estimate the number of heads if we flip a coin say 100 times. The mean of the binomial random variable $\mu = np$, where $n = 100$, $p = \frac{1}{2}$

$$\mu = 100(1/2) = 50$$

The standard deviation of the random variable $\sigma = \sqrt{np(1-p)}$

For example, with $n = 100$, $p = \frac{1}{2}$ $\sigma = \sqrt{100(1/2)(1-1/2)} = \sqrt{25} = 5$

We can use our confidence interval knowledge to calculate the 95% confidence interval for the expected number of heads in 100 coin flips as:

$$\bar{X} \pm 1.96 [\sigma] = 50 \pm 1.96(5) = (40, 60)$$

With 95% confidence, the mean of heads will be between 40 and 60.

- a) Calculate the 99% confidence interval for # of heads.
- b) Use the Random Number key to simulate a coin flip as follows:
 - 1) Press MATH
 - 2) Go to PRB [Probability]
 - 3) Enter
 - 4) Now press enter 30 times

Write down a count of random numbers that are greater than .5. These are heads; random numbers less than .5 could be considered as tails. In the unlikely event of obtaining a perfect .5, discard this outcome. I obtained 13/30 my first time. Repeat this exercise say ten times.

Suppose you now have a sample size of 300, $n = 300$, $p = \frac{1}{2}$. Suppose you obtained 140 heads, 160 tails. Your confidence interval estimate could be determined as follows:

$$\mu = np = 300(1/2) = 150$$

$$\sigma = \sqrt{np(1-p)} = 6$$

$$95\% \text{ C.I.} = 150 \pm 1.96(6) = (138, 162)$$

You may conclude that there was a 95% probability of getting between 138 and 162 heads. If you obtained 140, you were within chance levels. If you obtained 200, your calculator was probably broken. This is way outside chance. This is one of the great many exciting applications of the central limit theorem. We obtained our value of 1.96 from the normal curve, the centerpiece of Statistics.

Please take several courses in Statistics to better understand the way statistics can be applied to everyday decisions and improve your ability to plan.

Sample say 300 random numbers from 0-1 in a similar way to our example. Count the number of heads. Is your result within that predicted by a 95% confidence interval? Try a 99% confidence interval.

Homework

- 1, Calculate a 99% confidence interval for the data related to accountants' salaries. Would you expect the interval to be wider or narrower than the 95% confidence interval?
2. Use your TI-83 to enter your data for accountants' salaries. Obtain the mean, \bar{X} , and standard deviation, σ , as shown.

Next use the following to calculate the 99% confidence interval using your TI-83.

a) STAT

b) ◀

- c) Go to 8: T interval
- d) Enter
- e) Press Data
- f) List: L1
- g) Freq: 1
- h) C Level: 95
- i) Calculate
- j) Enter

The 95% confidence interval is now given: (22.023, 27.577) The TI-83 also gives $\bar{X} = 24.8$, $S_x = 3.881580434$, $n = 10$. You can check your computations using the TI-83.

3. Find a 95%, 99% confidence interval for your home price in Scarsdale if a recent New York Times advertised price list was as follows: (in thousands)

495, 650, 1300, 369, 599, 750, 375, 2400, 650, 475

4. Find a 95%, 99% confidence interval for your starting salary as a nurse if you sampled 36 starting salaries from Mariana and obtained an average of \$50K with standard deviation of 10K.

EXPERIENCE 3

Let us analyze your house price and salary estimated to determine whether you can afford your house. We will follow an example but ask you to subscribe to a local newspaper and obtain accurate housing and salary figures to ensure realistic estimates. Imagine living in Saranac Lake.

a) Your 95% confidence interval for house prices was calculated as (88K, 102K). To play it safe use an estimate of 102K. This way you can be 97.5% confident that your house will be cheaper.

b) Next use your salary estimate as accountant. Your 95% confidence interval was (22K, 27K). Take the 22K salary. You are 97.5% confident of making more.

c) Use your TI-83 Mathematics of Finance program to calculate your mortgage on your \$102K house. You would usually need 20K as a down payment. Your mortgage will be for approximately \$82K. Use TVM Solver as follows:

1. 2nd Finance [TVM Solver]
2. Enter
3. $n = 360$ (30 years, 12 months/year) You usually take a 30 year loan.
4. $I\% = \text{say } 8\%$, the rate has been around $8\% \pm 1/2\%$ for sometime
5. $PV = -82,000$. This means you start with debt (mortgage) of \$82K
6. $PMT =$ This is your monthly payment that SOLVER is used to determine.

You can leave the 0 there.

7. $FV = 0$ When you pay off the loan in 30 years, your future debt is 0.
8. $P/Y = 12$ (payments per year)
9. $C/Y = 12$ (the bank compounds your interest 12 times per year)
10. BEGIN
11. Go to PMT
12. Press ALPHA Enter

Your monthly payment is calculated at \$597.70. You must also add in taxes. Taxes in Saranac Lake are around \$200/month on such a house. So your monthly payment = mortgage + taxes = \$800 approximately.

Your monthly salary is approximately $\$22,000/12 = \$1,833$.

Banks like to loan people mortgage money if the ratio of the monthly payment/salary is 25% or less. Consider your ratio: $800/1833 = 44\%$.

No bank would grant you a mortgage with a 44% ratio. You will need to rent, buy a much cheaper house, or have another salary to afford a house in Saranac Lake, one of the more affordable areas to live in New York State. Imagine how impossible it is to start in an expensive area like Scarsdale. Mortgages carry a lot of stress, so try to live within your means.

Your experiential project is to:

- 1) Obtain 95% confidence estimates of your future housing and salary using the local newspaper.

- 2) Use TVM SOLVER to determine whether you can afford your future house.

Don't forget closing costs and insurance on your future home. We were way over 25% before we considered these staples of home ownership.

CHAPTER FOUR - TESTS OF SAMPLE MEAN

This is a course of Vital Statistics. It covers life and death matters, which impact your life. We are ready now to test hypotheses using our knowledge. A hypothesis is an educated guess. When I first began researching a book for baby boomers two years ago, I had a hypothesis that boomers would live longer if they retired in the Adirondack village of Saranac Lake as if they moved to Mariana, Florida. These were the two most highly rated small towns in New York State and Florida by Norman Crampton, *The 100 Best Small Towns in America* (Macmillan, 1995).

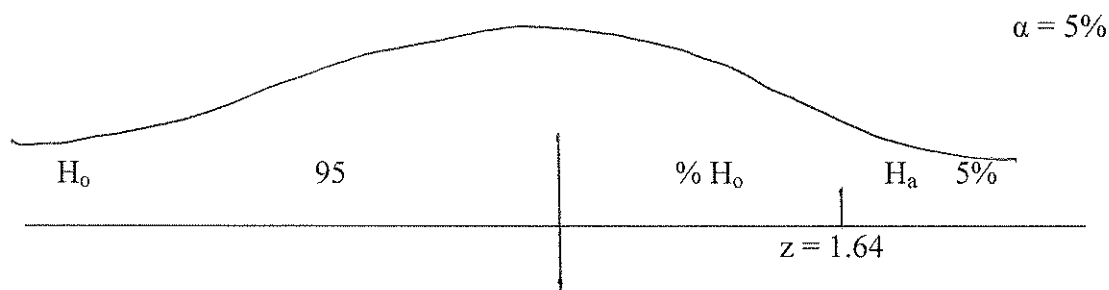
I sent away for the death records from the Adirondack Medical Center and Jackson County Health Department to determine whether Saranac Lake residents or Mariana residents lived above the national average of 75 years. The average for Mariana was 74.26 and the average for Saranac Lake was 75.27. We can use a simple formula to determine whether either was greater than (from a statistically significant perspective) the national average of 75. We need a brief introduction to the essential concept of significance. Please remember to take a complete course in statistics, since we have reduced statistics to bare essentials. There is no way for the reader to fully appreciate the enormous scope and power of statistics from these eight highlights.

Statistical Significance

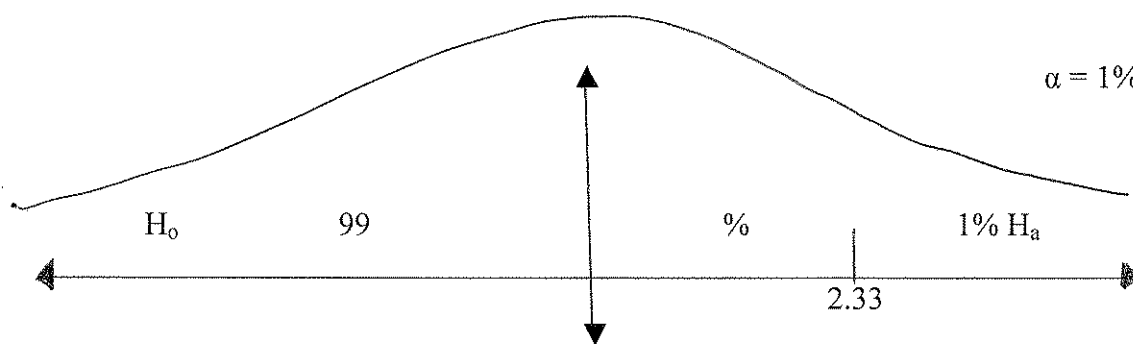
We are usually trying to establish that one value of a test statistic is significantly greater than another. For example, we know that Saranac Lake average life span of 75.27 based on a sample of say 30 is greater than 75. The question is whether it is statistically significant. We first need to set a level of Type I error. This level, usually 1% or 5%, means that even if we

conclude that our test statistic is statistically significantly greater, there is a 1% or 5% chance that the result was due to sampling. The sample simply was a biased one, not a reflection of the underlying superiority of one statistic. We are going to greatly simplify statistics by confining our analyses to two possibilities, 1% type error or 5% Type I error and the hypothesis that one statistic is greater than another.

If sample size is large ($n \geq 30$), we only need two pictures below:



Picture I



Picture II

We also assume that σ , the population standard deviation, is known. We rarely know σ . Therefore, we suggest that small sample t test for all hypotheses tests of means using the sample standard deviation S in place of σ .

H_0 is always the hypothesis that the two statistics are equal (from a statistical perspective). The two numbers are hardly ever exactly equal, but H_0 means that they are too

close to conclude that one is statistically significantly greater than the other. Two levels of statistical significance are common, 1% and 5%. These error levels, specifically Type I error levels, mean that 1% or 5% of the time sampling error will account for one sample to end up statistically significantly greater than the other. Type II error means that you have concluded H_0 (means are equal) when they are not (H_a is true).

We can now use the formula for one sample mean hypothesis tests below:

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

n = sample size
 \bar{X} = sample mean
 μ = population mean
 σ = standard deviation

Now let us test whether Saranac Lake's average life span 75.27 is statistically significantly greater than the national average of 75.

H_0 : (null hypothesis) Saranac Lake's life span is equal to that of the United States or $X = \mu$.

H_a : (alternative hypothesis) Saranac Lake's life span is greater than that of the United States or $X > \mu$.

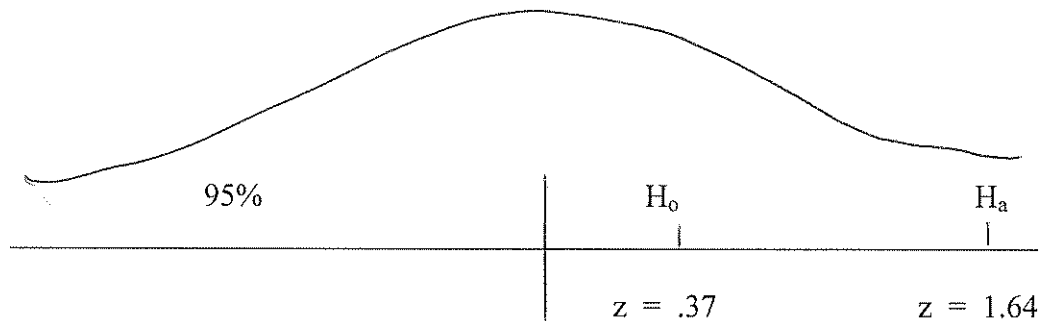
We arbitrarily select $\alpha = 5\%$, our confidence level is 95% and we use Picture I. If our z (computed value) is greater than 1.64, we conclude H_a ; any number less than 1.64, we conclude H_0 .

Calculate below: Let $\sigma = 4$, $n = 30$

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{75.27 - 75}{4/\sqrt{30}} = \frac{.27}{(4/5.477)}$$

$$z = \frac{.27}{.7303} = .37$$

Look at where $z = .37$ falls in Picture I, well to the left of 1.64.



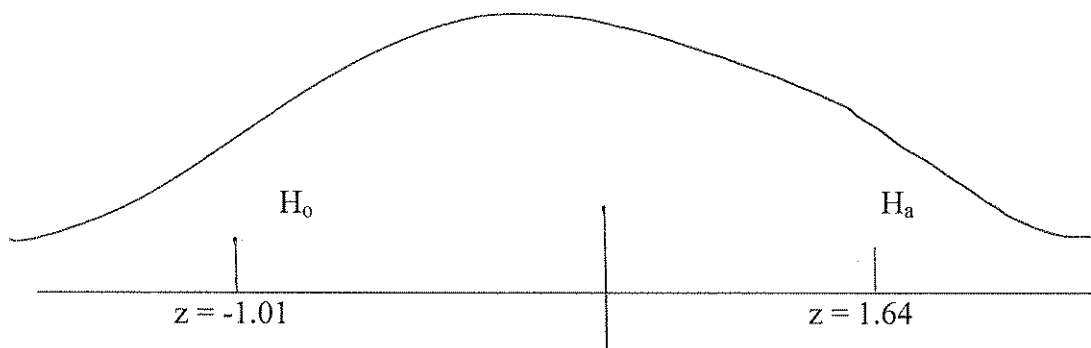
We conclude that the average life span in Saranac Lake is the same as the average American life span. If we were to require $\alpha = 1\%$, we would reach the same conclusion using Picture II.

The average life span of the sample of Mariana residents was 74.26. Let us use $\alpha = 5\%$ to determine whether Mariana was higher than the national average. (Of course, we will conclude H_0 since we are going to obtain a negative z value.)

$$Z = \frac{74.26 - 75}{4/\sqrt{30}}$$

$\bar{X} = 74.26$ (Mariana mean); $\mu = 75$ population mean
 $\sigma = 4$ population standard deviation; $n = 30$ sample size

$$Z = \frac{.74}{.7303} = -1.01$$



Our conclusion is H_0 , the two means are equal (from a statistical perspective).

It would be natural to ask whether the national average is higher than the Mariana average. Of course, you now are testing a second hypothesis and must add another 5% Type I error to your accumulated error. The result is summarized below:

$$Z = \frac{75 - 74.26}{4/\sqrt{30}} = 1.01$$

Since $1.01 < 1.64$, we conclude H_0 : The two means are equal. To be precise, the national average of 75 may have some rounding error, making our study somewhat inconclusive. However, you have just been introduced to one of the most powerful concepts of the twentieth century - statistical significance.

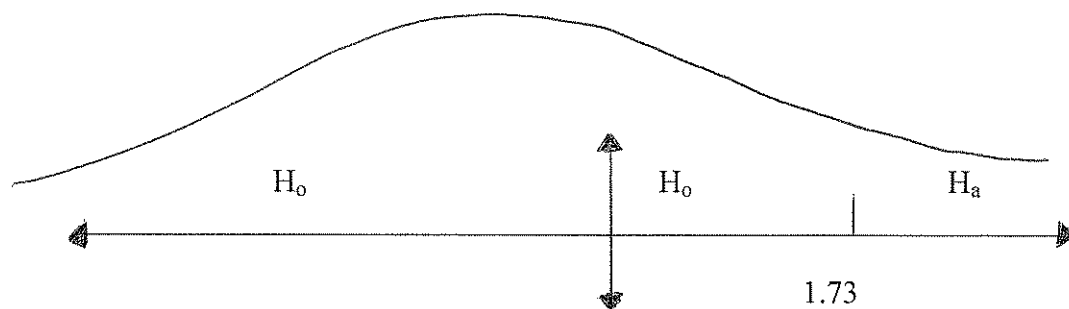
Small Sample Analysis

We often have samples that are small (less than 30). In order to perform the hypothesis test with one sample mean compared to a population mean, we simply have to adjust the critical value ($z = 1.65$ or 2.33). We can use the same formula to obtain a computed value, but we have to use the t-table with $(n-1)$ degrees of freedom to change our critical value. For example, let us change our original problem in only one element, sample size. Suppose we have an average life span in Saranac Lake of 75.27 but the sample size was 20. We compute the critical t value as follows:

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{75.27 - 75}{(4/\sqrt{20})} = \frac{.27}{.8944}$$

$$t = .30$$

Now, if you want 95% confidence, $\alpha = \text{Type I error} = 5\%$, you look up $\alpha = .05$, $df = n-1 = 20-1 = 19$. Use the picture below. $t_{.05, 19df} = 1.73$

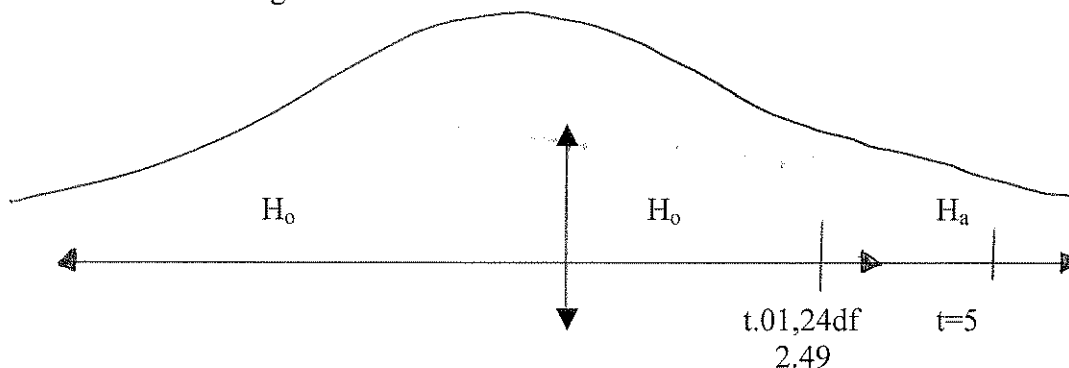


Since $.3 < 1.73$, we accept H_0 . There is no difference (statistically significantly) between Saranac Lake and the United States in longevity.

Let us analyze whether accountants make more than teachers, based on a sample of 25 accountants and the knowledge that teachers average salary (mean) = \$35K. The sample result for accountants was a mean of \$45K, $\sigma = 10K$. If we do not know the population average, we use the sample standard deviation. Usually we do not know population statistics since populations are typically very large. In general, since we rarely know σ , this is currently the preferred test. Consider the calculation for our problem below:

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{45 - 35}{10/\sqrt{25}} = \frac{10}{2} = 5$$

If we select $\alpha = 1\%$, $t_{.01, 24df} = 2.49$, consider H_0 : the means are equal. H_a : accountants' mean is higher.



Since $5 > 2.49$, we conclude H_a : accountants make statistically significantly higher salaries than teachers ($\alpha = .01$).

TI-83 Test for One Sample Mean

Our TI-83 has two separate menus, one for the z test (13-10) and one for the t-test (Guidebook 13-11). We have left out analyses of $\mu \neq \mu_0$ (means not equal). This is because the real world is usually testing whether one average is higher than another. We are typically testing whether one group lives longer, makes more money, is more satisfied.....than another. You can either use Data if you have sample data entered in L_1, L_2 , etc. Otherwise you can enter the data, \bar{X} , n , σ by the Stats option.

The calculated results will give you the same z or t that we calculated by hand. The major result is the p value. This leads to the probability that the results were statistically significant. For example, if $p = .3$, this is the Type I error associated with concluding that one mean was higher than the other. Since 5% is the maximum conventional α level, a p value of .3 means that the result was not statistically significant.

If you obtained a p value less than .05, the results are statistically significant with $\alpha = .05$. If you obtained a p value less than .01, your results are statistically significant at the .01 level. You may conclude that one mean is significantly greater than the other.

Let us use the TI-83 calculator to test whether a sample of 10 teachers' salaries in New York State were higher than the national average of \$35K. Our New York State sample was 28, 50, 60, 38, 85, 18, 62, 82, 90, 42.

1. Enter Sample in List 1
2. Go to STAT
3. Press \blacktriangleleft
4. Go to t-test (#2)
5. Enter

6. Input DATA
7. $\mu_o = 35$
8. List: L_1
9. Freq = 1 (each element in the sample is used once)
10. $\mu > \mu_o$ (Enter)

This tests whether the sample mean is greater than 35K

- 11, Calculate

The result is: $t = 2.62$

$$p = .013$$

$$\bar{X} = 55.5$$

$$S_x = 24.74$$

$$n = 10$$

This result means that New York State teachers' salaries are higher than the national average ($\alpha = .013886$).

So if you tested this hypothesis with $\alpha = .05$, you can conclude H_a , the sample mean was higher. If you set $\alpha = .01$, you have to conclude H_o , the two means are equal. This is because $.013886 > .01$. You are less than 99% confident that the sample of teacher salaries was significantly greater than the national average.

Homework

Complete each problem with both the TI-83 and by calculating the computed t or z value from the given data.

1. Engineers are said to average \$50K per year. A sample of 36 engineering salaries revealed a mean of \$55K with standard deviation of 10K. Is the sample result greater than the population average? Let $\alpha = .05$, $\alpha = .01$.
2. The salaries of eight doctors in Saranac Lake were 50K, 36, 72, 64, 56, 60, 70, 80. Are they below the national average of 80K? Use the sample standard deviation for your population standard deviation. $\alpha = .05$, $\alpha = .01$
3. Saranac Lake house prices were found as 80K, 110, 65, 90, 64, 86, 120, 72 in July. Are they below the national average of \$100K, $\sigma = 10$? $\alpha = .05$, $\alpha = .01$
4. A sample of 100 health records of Alaska residents revealed a mean life span of 80, $\sigma = 10$. Is this result above the national average of 75? $\alpha = .05$, $\alpha = .01$

EXPERIENCE 4

Survey 30 or more students about any measure, their GPA at college, how many hours they study or work per week. Make a guess as to the mean before you gather data. Test whether the true mean is larger (or smaller) than your guess. Use $\alpha = .05$ and $.01$

CHAPTER FIVE - TESTS OF TWO SAMPLE MEANS

This section will give you the tools to compare salaries between two different occupations. You could test whether house prices are higher in one town as compared to another. The more you plan and use statistics, the less likely you are to encounter financial difficulties during your life.

I researched two cities that were considered to be very desirable places to live - Marianna, Florida and Saranac Lake, New York. These two choices were low in crime, low in median house price, and recommended retirement places for baby boomers.

I took a sample of house prices from each city. Saranac Lake had 30 houses to choose from last year. The median price was \$63,000. This means that half the houses cost more; half the houses cost less.

If you decided to move to Marianna, you had 45 houses from which to pick last year. The median price was \$92,000.

If you decided to move to Scarsdale, the median house price last year was \$645,000. You have to make a great deal of money to support a \$645,000 house price, which does not include taxes and house maintenance.

In order to compare samples of houses, the median is a good place to start, but it does not allow you to determine whether one sample mean is statistically significantly higher than another. You need the mean and standard deviation for each sample. Then you substitute in the formula below if you are comparing two large samples (each sample has 30 or more elements).

Two Large Sample Difference of Mean Formula:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

\bar{X}_1 = sample mean for group 1

\bar{X}_2 = sample mean for group 2

σ_1^2 = sample variance for group 1

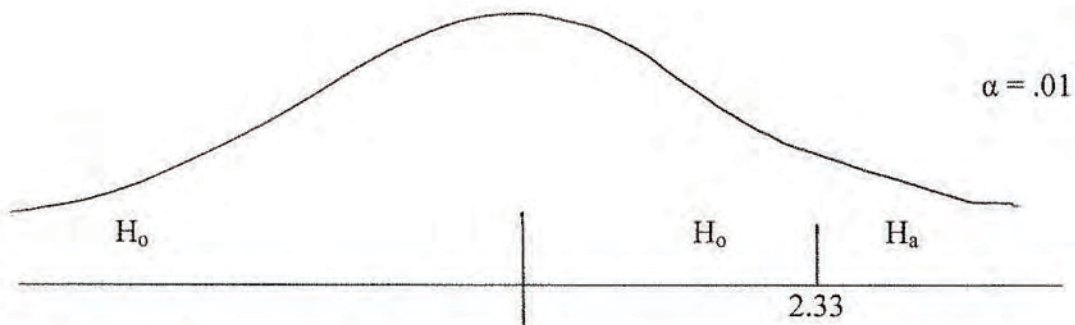
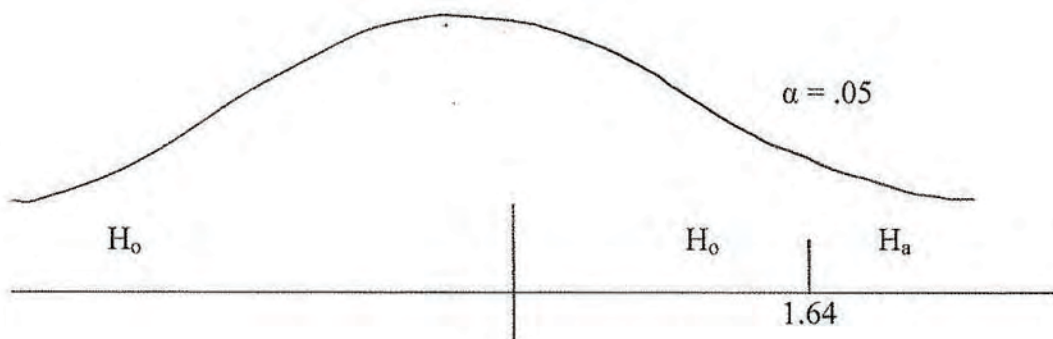
σ_2^2 = sample variance for group 2

Then you test whether

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Using two pictures



To illustrate, suppose you are comparing whether attorneys make more than doctors.

Suppose you find 40 attorneys' salaries which have a mean of 100K, $\sigma = 12$. We are being very

flexible in using σ in place of S , the sample standard deviation. Let our sample of 36 doctors have mean salary 98K, $\sigma = 5 = 10K$. Substitute S for σ , the population standard deviation σ we rarely know. All data is in thousands.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{100 - 98}{\sqrt{12^2/40 + 10^2/36}}$$

$$Z = \frac{2}{\sqrt{6.38}} = \frac{2}{2.5} = .80$$

If we let $\alpha = .05$, $.80 < 1.64$. See Picture I. Our conclusion is:

Accept H_0 : $\mu_1 = \mu_2$

OR there is no statistically significant Z value from which you could conclude that lawyers earn more than doctors. If $\alpha = .01$, $.80 < 2.33$. Your conclusion is the same.

Often you do not have one or both large samples. For example, biotechnology trials frequently use samples smaller than 30. Some hospitals, who are researching new drugs, have a limited number of patients. They must then take half of the sample for a control group, which receives a placebo (or no drug treatment).

We can easily adjust statistics to the small sample case by using the formula below:

Two Small Sample Difference of Means ($n_1 < 30$ or $n_2 < 30$)

Follow these steps:

1. Compute the pooled variance

First calculate σ_1^2 , σ_2^2 . Then find $S_p^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}$

2. Compute the two sample means, \bar{X}_1 and \bar{X}_2

3. Compute $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$

Now adjust your previous large sample critical value by looking up your t value ($\alpha = .05$ or $\alpha = .01$) with $(n_1 + n_2 - 2)$ degrees of freedom. We are limiting our discussion to the most common comparison: $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 > \mu_2$.

Please follow this example. Suppose cancer patients starting chemotherapy with drug A live 12 years longer on average, $\sigma_1 = 4$, sample size of 10. The placebo resulted in an average of 8 additional years, $\sigma_2 = 3$, sample size of 10. Can we conclude that drug A is effective in prolonging life?

Let $\alpha = .05$, which is what the FDA requires before a new drug can be marketed nationally. Substitute in the small sample t formula:

$$1. \quad \sigma_1^2 = 16, \quad \sigma_2^2 = 9, \quad n_1 = 10, \quad n_2 = 10$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Substitute σ for S .

$$S_p^2 = \frac{(10-1)16 + (10-1)9}{10 + 10 - 2}$$

$$S_p^2 = \frac{9 \cdot 16 + 9 \cdot 9}{18} = 12.5$$

You should always obtain a value of S_p^2 somewhere between (or equal) to the values of S_1^2 and S_2^2 . The pooled variance S_p^2 is an average of the two variances and it enables us to use the t test, which adjusts for the small sample size.

$$2. \quad X_1 = 12, \quad X_2 = 8. \text{ These are the values of the two sample means.}$$

$$3. \quad \text{Now calculate the computed t value.}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 (1/n_1 + 1/n_2)}}$$

$$t = \frac{12 - 8}{\sqrt{12.5(1/10 + 1/10)}}$$

$$t = \frac{4}{1.5811} = 2.5298 = 2.53$$

Note, to calculate $\sqrt{12.5(1/10 + 1/10)}$, there are several ways to compute. One way is as follows:

1. .1 + .1
2. Enter
3. x 12.5 (multiply by 12.5)
4. Press ^ This is the exponent key.
5. .5 (the square root means the $\frac{1}{2}$ power or 5/10).
6. Enter

The answer is 1.58113883. Now look up your critical t in the t table.

$$\text{degrees of freedom} = n_1 + n_2$$

Set $\alpha = .05$, which is what the FDA requires for the approval of a new drug for cancer treatment (or virtually any new experimental trial).

$$df = 10 + 10 - 2$$

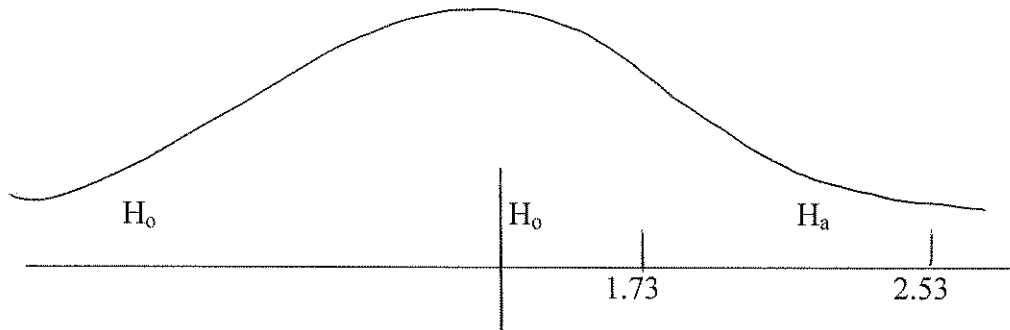
$$t_{.05,18} df = 1.73$$

Use Picture IV below with:

$H_0: \mu_1 = \mu_2$ (the means are equal from a statistical perspective)

$H_a: \mu_1 > \mu_2$ (people live significantly longer with the treatment utilizing drug A)

Picture IV



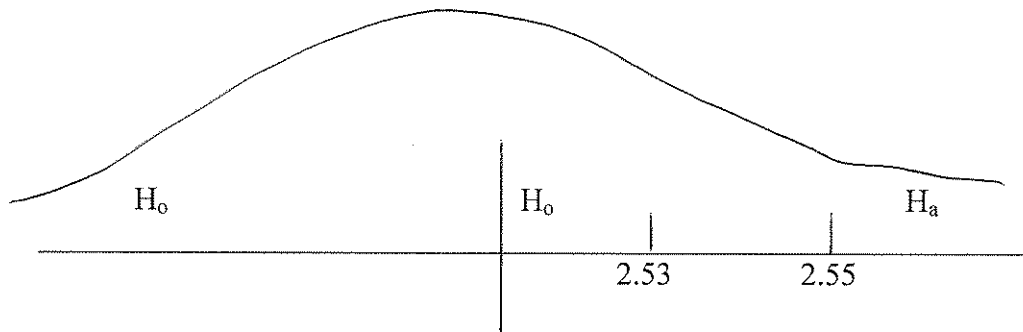
We conclude H_a : Since $2.53 > 1.73$, we accept H_a , the mean of Group I is significantly higher than Group II.

If we decided on an alpha value, $\alpha = .01$, this is a harder level of statistical significance to achieve. It is usually considered too stringent, since many effective treatments may fall short. After all, if you achieved a confidence level of 98% or $\alpha = .02$ for the efficacy of some medical advance, you would have to consider the method a failure at an $\alpha = .01$ level.

Now if we used $\alpha = .01$, look up the critical t value.

$$t_{.01,18} \text{ df} = 2.55$$

As a result, we would have to conclude that $\mu_1 = \mu_2$ or the drug did not work if we used $t_{.01,18} \text{ df} = 2.55$ for our test. Accept H_0 : $\mu_1 = \mu_2$ The two means are equal.



TI-83 Section

The TI-83 is programmed to enable the student to either enter the data in List 1 (Sample 1) or List 2 (Sample 2). Then you can go to the STAT menu and use wither #2, the 2-sample Z test or #3, the 2-sample t test.

Let us re-do our previous example using the TI-83 according to the following steps. We summarize the data below:

Cancer patient data: $\bar{X}_1 = 12$

$$\bar{X}_2 = 8$$

$$n_1 = 10$$

$$n_2 = 10$$

$$\sigma_1 = 4$$

$$\sigma_2 = 3$$

$$\alpha = .05$$

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Now please follow the steps to program the TI-83 to complete the analysis:

1. Press STAT
2. Press \blacktriangleleft
3. Go to #4, 2 samp t test
4. Enter
5. Input stats. Enter (This is because you have the values $\bar{X}_1, \bar{X}_2, \sigma_1^2, \sigma_2^2$. If

you have the data in List 1 and List 2, you would press Enter on Data.

Put in the following data:

6. $X_1 = 12$
7. $S_{x1} = 4$
8. $n_1 = 10$
9. $X_2 = 8$
10. $S_{x2} = 3$
11. $n_2 = 10$
12. We want to test $\mu_1 > \mu_2$, so go to $> \mu_2$. Enter
13. Pooled go to Yes. Enter
14. Calculate. Enter

Our t value confirms our work. $t = 2.5298$ $p = .0104$. As we know, we narrowly missed statistically significant results with $\alpha = .01$ ($p = .01$) but easily concluded that $\mu_1 > \mu_2$ at $\alpha = .05$.

Homework

1. Redo the comparison of attorney and physician salaries using the TI-83, $\alpha = .05$.
2. If Philadelphia's average house price is \$210K and Atlanta's is \$185K, where the sample size for Philadelphia was 100, $\sigma_1 = 50$, the sample size for Atlanta was 60, $\sigma_2 = 40$, can we conclude that Philadelphia's mean house price is significantly greater than Atlanta's, $\alpha = .05$? Use the computed Z value with the formula and check your work with the TI-83.

3. The following is a list of salaries for teachers in public schools (List 1) and a list of salaries for private schools (List 2). Use both calculated t and the TI-83 to analyze whether public school teachers make significantly more than private school teachers. Use $\alpha = .05$ and $\alpha = .01$.

Public ($n_1 = 12$) 65, 55, 36, 80, 42, 33, 36, 38, 35, 40, 45, 50

Private ($n_2 = 11$) 26, 60, 35, 30, 53, 28, 30, 28, 28, 29, 29

4. Test whether people who live in the Adirondacks live longer than people who live in Mariana County based on an average life span in Saranac Lake of 75.27 with standard deviation 4, based on a sample of 20. In Mariana, the average life span was 74.26, with standard deviation of 4, based on a sample of 30. Use both the small sample t formula and the TI-83 to test your hypothesis with $\alpha = .05$ and $\alpha = .01$.

EXPERIENCE 5

Test any hypothesis with a sample from two populations. For example, do women have a higher GPA than men? Was your hypothesis confirmed or rejected by your study? What have you learned?

CHAPTER SIX - DIFFERENCE OF PROPORTIONS

Many times, there are only two possibilities in a category - success or failure in sports, defective or properly working in a mechanical process, divorced or continued married, or employed/unemployed. We can call P the sample proportion of success and define P as:

$$P = \frac{\text{number of successes in the sample}}{\text{total sample size}}$$

For example, if ten transmissions are defective from auto dealer A, the proportion of proper functioning transmissions is:

$$P = \frac{990}{1000} = .99$$

If you want to get several job offers, it may be necessary to send out hundreds, if not thousands, of resumes and cover letters. Sometimes the number of resumes it takes to have one job interview may be in the range of 100 to 1, sometimes 1000 to 1.

The ratio of successes P is defined as 1/100 (1%) or 1/1000 (.1%). You could use knowledge of job demand to your advantage in selecting your college major or your career change. For example, if 29 of 30 graduates of Columbia Law School obtained legal positions with starting salaries about \$80,000, your admission to law school would be a likely ticket to professional success. Of course, you should inspect what percentage of students who are admitted to Columbia Law School drop out. If 200 students enter and only 30 graduate, the 29/30 success ratio is not so impressive as a placement statistic.

It is natural to compare two ratios to decide whether one is significantly higher than another. For example, employers have to be sensitive to issues related to job discrimination. In Introductory Statistics (J. Devore and R. Peck, West, 1994, NY: p. 344-345.), the authors cite an article that discussed the court case Swain v. Alabama (1965) in which there was alleged

discrimination against blacks in grand jury selection. The census data indicated that blacks constituted 25% of the grand jury pool and a random sample of 1050 people called for jury service yielded 177 blacks. We can use the following formula and our earlier two large sample Z pictures to test the hypothesis H_0 vs. H_a where:

$$H_0: \bar{P} = P$$

\bar{P} = the sample proportion

P = the population proportion

$$H_a: P > \bar{P}$$

The population proportion is significantly greater than the sample proportion.

One Sample Formula

There is one sample that you are comparing to a population statistic. The one sample z test for a population proportion is:

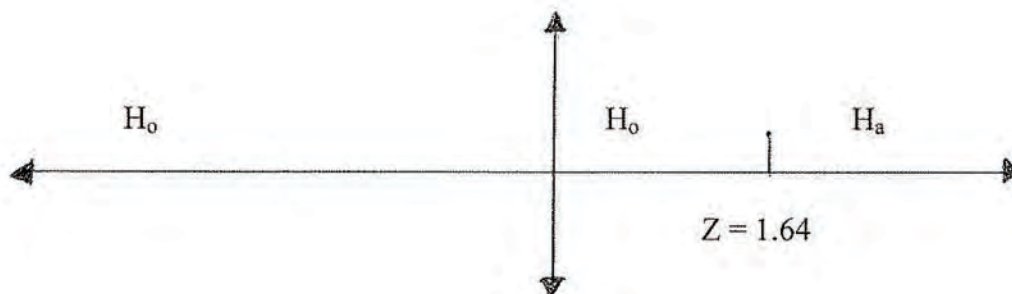
$$Z = \frac{P - \bar{P}}{\sqrt{P(1-P)(1/n)}}$$

Use the population proportion for the denominator

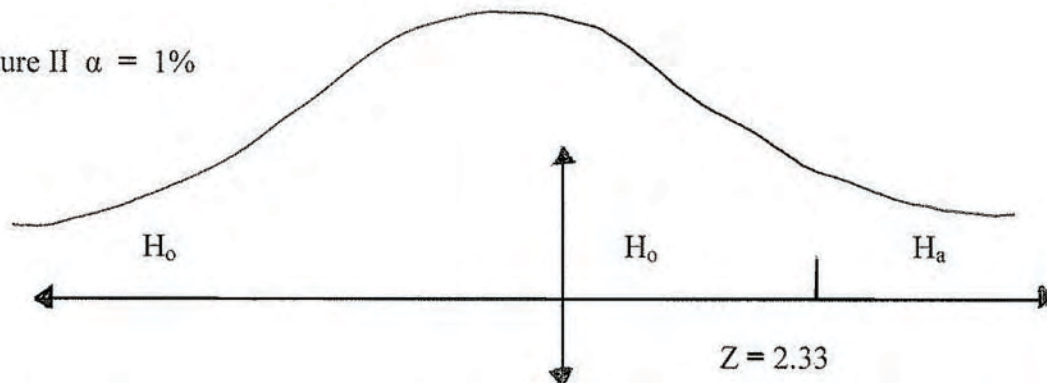
where n = sample size

You can use either Picture I or Picture II below to test your hypothesis:

Picture I $\alpha = 5\%$



Picture II $\alpha = 1\%$



Now test whether our result of 177 blacks from a sample of 1050 is significantly less than the population proportion of 25%. To simplify the algebra, use the higher proportion as the first value. This avoids negative numbers and a shifting of Picture I and II to the negative side of the x-axis.

$$Z = \frac{P - \bar{P}}{\sqrt{P(1-P)(1/n)}}$$

$$P = 25\%$$

$$\bar{P} = 177/1050 = .1686$$

$$n = 1050$$

$$Z = \frac{.25 - .1686}{\sqrt{.25(1-.25)(1/1050)}}$$

$$Z = \frac{.0814}{.0134} = 6.07$$


This value is much higher than $Z = 2.33$ (Picture II). We can conclude that $H_a: P > \bar{P}$ with much higher confidence than 99%. Lawyers sometimes use 99% as the probability corresponding to "beyond a shadow of a doubt," a key element of criminal prosecution. We may conclude that there was discrimination against blacks at any conventional level of Type I error.

The court only looked at the numerator and observed a difference of only 8%. They concluded (incorrectly) that the difference was not large enough to establish a prima facie case (a case without further examination).

TI-83

Let us complete the statistical analysis (using the TI-83) of the discrimination case data.

Use the following steps:

1. Press STAT
2. 
3. Go to #5: 1 Prop Z Test
4. Enter
5. P_o : .25
6. X: 177
7. n: 1050
8. Prop < P_o (population proportion)

We are testing whether our population is greater than our sample proportion. This is simply reversing the inequality sign. For example, $10 > 5$ is the same expression as $5 < 10$.

9. Enter

10. Calculate

$Z = -6.0935563$

$P = .999999$

The Z value is negative, because the calculator uses the value $(\bar{P} - P)$ in the numerator where we used $(P - \bar{P})$ to avoid negative numbers. The slight discrepancy between our Z of 7.02 and the calculator result is due to the calculator using 9 significant digits after the decimal. The P value is the main issue to keep in mind. The result (accept H_a) is statistically significant at any conventional level of α . You can have .999999 etc. confidence in your result. Discrimination is well demonstrated by our objective statistical analysis.

Two Sample Difference of Proportions

The stock market has made a great number of millionaires in the 1990s. Stocks like EMC and Dell have gone up way over 1000% in the past ten years.

One guide to assist you in picking the Dells and EMCs of the 2000s is to consult analysts' ratings. Analysts rate stock with 1 = strong buy, 2 = buy, 3 = hold, and 4 = sell. The best rating a stock could have is a mean of 1; the worst is a 4.

Analysts tend to be optimists so you might look at the fraction of analysts who rate a stock as a strong buy. On August 30, 1999, a visit to www.morningstar.com revealed the following analyst ratings.

$\frac{\text{Strong Buys}}{\text{Total \# of Analysts}}$	= P
--	-----

Intel	P = 15/36
-------	-----------

AOL	P = 25/40
-----	-----------

$$\text{Dell} \qquad P = 13/32$$

It is natural to compare the strength of analyst ratings of two stocks, such as America On Line compared to Dell. The Z test for the difference of proportions is the test of choice. The test has several requirements, before you can use it. Consider the following:

$$n_1 = \text{sample size for sample 1}$$

$$n_2 = \text{sample size for sample 2}$$

$$P_1 = \frac{\text{\# of successes in sample 1}}{n_1} = \frac{x_1}{n_1}$$

$$P_2 = \frac{\text{\# of successes in sample 2}}{n_2} = \frac{x_2}{n_2}$$

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

Requirements:

1. $n_1 \geq 30$
2. $n_2 \geq 30$
3. $n_1 P_1 \geq 5$
4. $n_2 P_2 \geq 5$
5. $n_2 (1 - P_2) \geq 5$

Let us test our 5 requirements in the test of whether the rating of American On Line (AOL) is higher than the rating of Dell.

$$P_1 = \frac{x_1}{n_1} = \frac{25}{40}$$

$$P_2 = \frac{x_2}{n_2} = \frac{13}{32}$$

To test the five requirements:

1. $n_1 = 40 > 30$
2. $n_2 = 32 > 30$
3. $n_1 P_1 = 40 \cdot 25/40 = 25 > 5$
4. $n_2 P_2 = 32 \cdot 13/32 = 13 > 5$
5. $n_2 (1 - P_2) = 32(1 - 13/32) = 32(19/32) = 19 > 5$

The five requirements are satisfied, so we can use the following test:

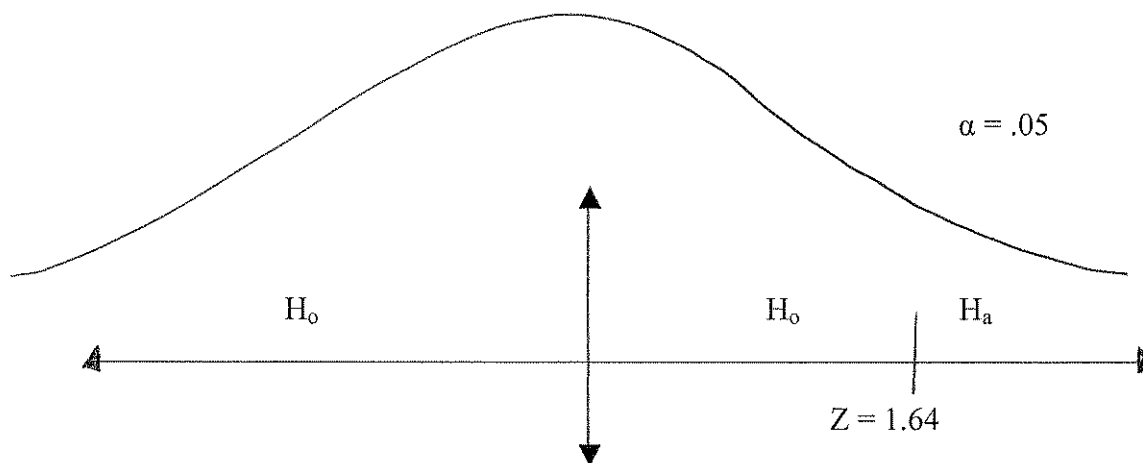
$$H_0: P_1 = P_2$$

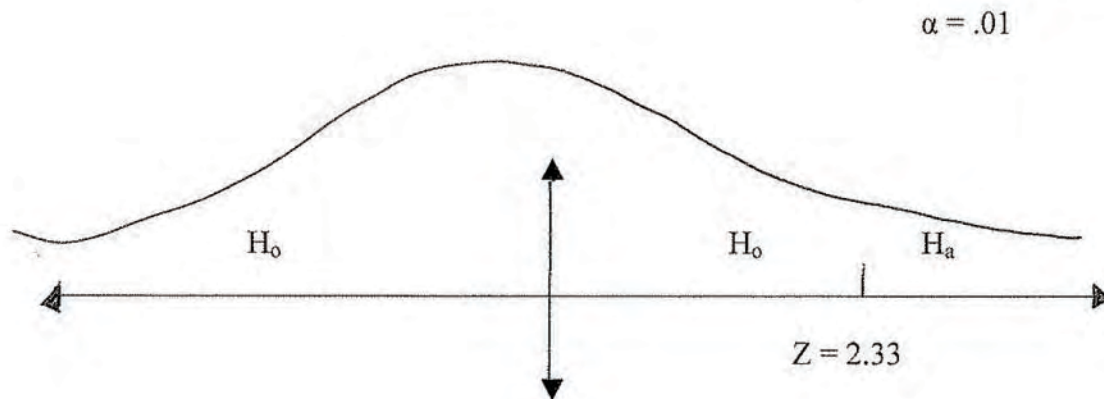
The two proportions are equal.

$$H_a: P_1 > P_2$$

P_1 is significantly greater than P_2 .

Set $\alpha = .05$ or $.01$ and use Picture I or II below.





The formula for a 2 sample difference of proportion is:

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2}}}$$

Substitute: $P_1 = 25/40$, $P_2 = 13/32$, $n_1 = 40$, $n_2 = 32$

$$P = \frac{25 + 13}{40 + 32} = \frac{38}{72} \text{ is called the pooled proportion}$$

$$Z = \frac{(25/40) - (13/32)}{\sqrt{\frac{38/72(1-38/72)}{40} + \frac{38/72(1-38/72)}{32}}}$$

$$Z = \frac{.625 - .406}{\sqrt{.0062 + .0078}} = \frac{.219}{.118}$$

$$Z = 1.86$$

Use Picture I and Picture II

If you select $\alpha = .05$, your conclusion is accept H_a : $1.86 > 1.64$

$\therefore P_1 > P_2$, AOL has a significantly higher rating than Dell.

If you select $\alpha = .01$, your conclusion is accept H_0 : $1.86 < 2.33$

$\therefore P_1 = P_2$, AOL has an equal rating as Dell.

This remarkable result, two contradictory findings based on your selected α level, illustrates how essential an understanding of statistics is to your financial and personal life. Most studies rely on an α level of 5%, so as to not reject significant results that usually fall short of the highly stringent $\alpha = .01$ level.

TI-83

Let us complete the same problem using the TI-83's statistical program. Use the following steps:

1. STAT
2. \blacktriangleleft
3. Go to #6 - 2 Prop Z Test
4. Enter
5. X_1 : 25
6. n_1 : 40
7. X_2 : 13
8. n_2 : 32
9. P_1 : $>$ P_2 Enter
10. Calculate Enter

The result is $Z = 1.85$; the P value is .0323. This confirms our earlier result. We concluded that $P_1 > P_2$ ($\alpha = .05$) and $P_1 = P_2$ ($\alpha = .01$). The precise level of alpha is .03. If you set alpha lower than .03, you fail to accept H_0 . If you set alpha higher than .03 ($\alpha = .05$), you can conclude $P_1 > P_2$.

Homework

1. Do more than 50% of analysts rate AOL as a strong buy? $\alpha = .05$, $\alpha = .01$. Use both the formula and the TI-83.
2. You are told that 10% of your resume letters will result in an interview. If you send out 1000 letters and receive 60 requests for an interview, is this significantly below the advertised claim of 10%? Let $\alpha = .05$. Use both the formula and the TI-83.
3. Intel has a strong buy rating from 15 analysts of 36 who have rated the stock (August 31, 1999). Is this significantly lower than the 25/40 rating AOL a strong buy? Use $\alpha = .05$, $.01$ and both formula and TI-83.
4. Presidential candidate A was given a favorable rating by 100 of 250 respondents. Candidate B was given a favorable rating by 90 of 300 respondents. Does Candidate A demonstrate significantly higher support at $\alpha = .05$? Use both the formula and TI-83.

EXPERIENCE 6

Ask two samples of thirty or more any question with a yes/no or true/false response. Compare the two resulting fractions. Was the % of the group response higher as you expected? Use $\alpha = .05$ and $.01$. What have you learned?

CHAPTER SEVEN - LINEAR CORRELATION/REGRESSION

The stock market has made a lot of people very wealthy in the 1990s. One of the key factors that has increased the value of stocks is the increase in the earnings of stocks in the 90s. For example, if IBM made an average of \$1 per share for 1990 and earned an average of \$5 per share in 2000, the stock may increase in price by a factor of 5.

Of course, there are many factors that relate to stock price besides earnings. The purpose of this section is to develop a method for determining the most accurate line, $y = mx + b$, that relates variables such as earnings (x) to y (stock price). Also, the Siamese twin of the line is the related topic, linear correlation, which measures the strength of the linear relationship between x and y .

Sometimes, there is a perfect linear relationship between two variables. For example, if you go to a store that charges 7% sales tax, the equation $y(\text{tax}) = .07x$ (purchase) is a deterministic relation between x and y . If you buy \$100 worth of goods, your tax is determined as follows:

$$y = .07x = .07(100) = \$7$$

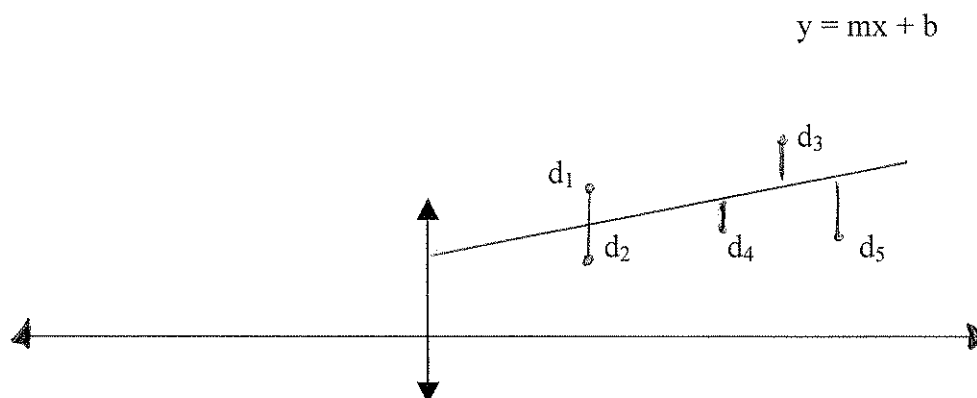
There is no error. You have a perfect prediction.

In real life, the prediction of one variable (y) based upon an independent variable (x) is rarely perfect. In fact, stock prediction could be improved by a complex regression method called multiple regression, where stock price $y = a_0 + a_1 x_1 + a_2 x_2 + \dots a_n x_n$. Some models use 5 predictors of stock price, while others use dozens. This topic is too complicated for our brief introduction to key topics from Statistics.

Before you decide upon a method to link two variables, it is wise to plot the points (x, y). The points are called the scatter plot and may show whether a linear relationship ($y = mx + b$)

exists between x and y or perhaps a quadratic relationship ($y = ax^2 + bx + c$) or an exponential relationship ($y = e^x$) or a variant.

We are going to use the principle of least squares. See the picture below.



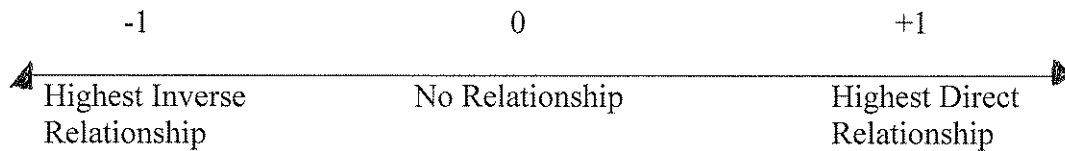
The d 's represent the vertical distances between five points and the best fitting line $y = mx + b$, which minimizes the sum of the squares of the differences in y between the actual y value and the y value of the regression line. Algebraically we are finding the equation $y = mx + b$ to obtain a minimum for $\sum d^2$.

Let us use recent data from Barrons (August 30, 1999), which gives the current price y of selected stocks and the analysts' expected earnings (per share) for the year 1999. The data for 5 stocks is presented below:

	<u>x (earnings)</u>	<u>y (price)</u>
America On Line	.60	99 1/8
EMC	1.06	60 9/16
General Electric	3.21	116 9/16
General Motors	4.55	65
IBM	3.91	124

Let us calculate the coefficients $y = mx + b$ and the correlation coefficient r for the data.

First calculate the correlation coefficient, r .



The correlation coefficient, r , ranges between -1 and +1. The closer the r value is to -1, the higher the inverse relation between x and y . For example, if you are analyzing major league pitchers' salaries, there is a strong inverse relation between your earned run average (average number of runs you allow per game) and your salary. The lower the number of runs you allow, the higher your salary will be.

The relation between your level of education (x) and your future salary (y) is nearly a perfect 1. The more education you pursue (masters, medical degree, etc.), the more money you will likely make. No relation is a perfect +1 or -1, except for trivial deterministic relations.

The formula for calculating the correlation coefficient, r , is below:

$$r = \frac{n \sum (xy) - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Complete the table for our given data, which we rounded to the nearest whole number to facilitate calculation.

	x	y	xy	x^2	y^2
AOL	1	99	99	1	9801
EMC	1	61	61	1	3721
GE	3	117	351	9	13,689
GM	5	65	325	25	4,225
IBM	4	124	496	16	15,376

Calculate the following:

$$n = 5 \text{ (\# of pairs)}$$

$$\Sigma x = 14$$

$$\Sigma y = 466$$

$$\Sigma xy = 1,332$$

$$\Sigma x^2 = 52$$

$$\Sigma y^2 = 46,812$$

Substitute in the formula for r:

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

$$r = \frac{5(1332) - (14)(466)}{\sqrt{5(52) - 14^2} \sqrt{5(46,812) - 466^2}}$$

$$r = \frac{136}{\sqrt{64} \sqrt{16,904}} = \frac{136}{1040} = .13$$

This is a low positive correlation. You have to look for other factors besides earnings to predict stock price with a good level of accuracy. The correlation of .13 demonstrates the bubble in the stock market in 1999. The price to earnings ratio is historically around 12:1. The P/E for EMC was $\frac{60.9}{1.06} = 5.71$. This high P/E predicted the subsequent stock market crash that came

within a year. EMC was to lose over 90% of its value. Today the P/E ratio for most stocks approaches the historical average of 12. We have learned about bubbles and the importance of earnings in support of stock prices.

You can use the calculations from your correlation analysis to determine the least-squares line $y = a + bx$. This line is a predictor of y based upon values of x.

The slope of the least squares line is

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Use the values from your correlation analysis to substitute:

$$n \sum xy - (\sum x)(\sum y) = 136$$

$$n \sum x^2 - (\sum x)^2 = 64$$

$$b = \frac{136}{64} = 2.1$$

To find the y intercept of $y = a + bx$, use the formula:

$a = \bar{y} - b\bar{x}$, where \bar{y} , \bar{x} are the means of y and x respectively.

$$\bar{y} = \frac{\sum y}{n} = \frac{466}{5} = 93.2$$

$$\bar{x} = \frac{\sum x}{n} = \frac{14}{5} = 2.8$$

$$a = \bar{y} - b\bar{x} = 93.2 - 2.1(2.8)$$

$$a = 87.3$$

The least squares line $y = a + bx = 87.3 + 2.1x$

You can use the line, $y = 87.3 + 2.1x$ to predict y based on various values of x. For example, if $x = 3$, we can predict $y = 87.3 + 2.1(3) = 93.6$

It is important to evaluate how well the least squares line predicts y. A standard method for evaluating the effectiveness of a least square line is the coefficient of variation. If the points in a scatter plot are close to the least squares line, the line is a good fit.

Again, please take a complete Statistics course. This topic will be covered in depth in a standard elementary Statistics course.

First compute the residual sum of squares, SSRESID. This gives you the sum of the squares of the differences between your given y values and your predicted y values, given the line $y = a + bx$.

$$SSRESID = \sum y^2 - a \sum y - b \sum xy$$

Next compute $\sum (y - \bar{y})^2 = SSTO$

This gives you the squared differences between your y values and the mean.

You can use the short cut formula: $SSTO = \sum y^2 - \frac{(\sum y)^2}{n}$

Let us proceed with our calculations:

$$SSRESID = \sum y^2 - a \sum y - b \sum xy$$

$$SSRESID = 46,812 - 87.3(466) - 2.1(1332)$$

$$SSRESID = 3333$$

$$SSTO = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SSTO = 46,812 - \frac{466^2}{5}$$

$$SSTO = 3381$$

We now compute the COEFFICIENT OF DETERMINATION

$$R^2 = 1 - \frac{SSRESID}{SSTO} = 1 - \frac{3333}{3381} = .01$$

R^2 ranges between 0 and 1.

$R^2 = .01$ for our least squares line

The closer the value of R^2 is to 1, the better prediction you can make of y based upon x.

Since our R^2 value is so small, we conclude that we cannot very effectively predict stock price on

the basis of earnings. Our current stock market is based on high growth stocks with currently low earnings. Our result confirms that earnings is not a key factor in determining a stock's price.

We can use the TI-83 to confirm our work. Let us work out the same problem with the TI-83.

1. Put the values of x in List 1: {1,1,3,5,4} → L1
2. Put the values of y in List 2: {99,61,117,65,124} → L2
3. STAT
4. CALC
5. Go to #4 Lin Reg (ax + b)
6. Enter
7. Enter

The result is: $y = ax + b$

$$a = 2.125$$

$$b = 87.25$$

This confirms our earlier result. To calculate the correlation, R, follow these steps:

1. 2nd catalog
2. Go to Diagnostic On
3. Enter
4. Enter
5. Go to STAT
6. CALC
7. Go to #4 Lin Reg (ax + b)
8. Enter

9. Enter

Your result is: $R^2 = .017$

$R = .13$ which confirms our conclusions that R^2 is a low value or the least squares line is not very accurate.

Homework

From Barrons (August 30, 1999)

	This Year's Earnings x	Next Year's Earnings y
AOL	.39	.60
EMC	1.06	1.37
GE	3.21	3.66
GM	8.43	8.15
IBM	3.91	4.49

1. Find the correlation between x and y - the two years' earnings. Find R^2 - the coefficient of variation. Interpret the results.

	y stock price	x earnings growth (%)
AOL	1	$\frac{.60 - .39}{.39} = .54$
EMC	1	
GE	3	
GM	5	
IBM	4	

2. To calculate earnings growth, subtract this year's earnings from next year's earnings and divide by the year's earnings. For AOL, earnings growth equals:

$$\frac{.60 - .39}{.39} = .54$$

Convert to % for AOL, $x = 54$

- (a) Complete the table for values of x .
- (b) Compute the correlation R and the least squares regression line $y = a + bx$,

based upon the data. Use both the formula and the TI-83 for your calculations. Interpret your results.

3.

x (education)		y (average annual salary - in thousands)
High School	0	20
Two Years College	1	25
Four Years College	2	47
Master's Degree	3	55
Law/Medical Degree	4	80

Let $x = 0$ (high school)

$x = 1$ (two years college)

$x = 2$ (four years college)

$x = 3$ (master's degree)

$x = 4$ (law/medical degree)

Calculate the correlation between your level of education and your eventual average annual salary. Also find the least squares line: $y = ax + b$. Use both the computing formula and the TI-83. Estimate y if $x = 2.5$.

4. OPEN QUESTION

Reverse x and y in problem #2. Do you get an equivalent equation with interchanged values of x and y or a slightly different equation? This problem is a BRAIN TEASER and was proposed by Dr. Ben Fusaro, the founder of the International Mathematical Modeling Competition (Florida State University).

EXPERIENCE 7

Obtain a sample of ten to thirty people and ask each to fill out the responses to two items.

For example, their level of education (x) and salary (y).

1 - no high school; 2 = high school diploma; 3 = two years college; 4 = four years college; 5 = master's degree; 6 = doctorate or law degree.

Obtain both the linear correlation and regression equation $y = a + bx$. What have you learned?

ANSWER KEY

CHAPTER ONE

1. 80%
2. 20%, 80%
3. (a) $\frac{201.6}{100,000} = .0020$
(b) $\frac{179.6}{100,000} = .0018$
(c) $\frac{307.4}{100,000} = .0031$
(d) $\frac{134.6}{100,000} = .0013$
4. $1 - .043 = .957$

CHAPTER TWO

1. $S_x = 6.099$ Sample Standard Deviation
 $\sigma_x = 5.455$ Formal Def. St. Deviation

Variance
Sample $= 6.099^2 = 37.0881$
Formal Def. $= 5.455^2 = 29.7570$

Mean $= \bar{X} = 35.2$

Use Sample St. Deviation
Coefficient of Variance $= \frac{6.099}{35.2} \cdot 100\% = 17.33\%$
2. Check
3. (a) $\bar{X} = 9.833$
(b) $S_x = 9.517$ st. dev.
 $S_x^2 = 9.517^2 = 90.573$

$$\text{coefficient of variation} = \frac{9.517}{9.833} \cdot 100\% = 96.816\%$$

$$4. \quad \bar{X} = \text{mean} = 131.6667$$

$$S_x = \text{st. dev.} = 20.2072$$

$$S_x^2 = 20.2072^2 = 408.3309$$

$$\text{coefficient of variation} = \frac{20.2072}{131.6667} \cdot 100\% = 15.347\%$$

Using the formal def. for variance will give slightly different results. The formal definition will compute the variance as 271.9. The short cut formula to compute the variance yielded an answer of 408.33. The calculus based statistics course will clarify these two seemingly different answers.

CHAPTER THREE

$$\begin{aligned} 1. \quad 99\% \text{ C.I.} &= 24.8 \pm t_{.005, 9df}(3.8816/\sqrt{10}) \\ &= 24.8 \pm 3.250 (3.8816/\sqrt{10}) \\ &= (20.811, 28.789) \end{aligned}$$

2. Check

$$3. \quad 95\% \text{ C.I.} = (362.59, 1250)$$

$$99\% \text{ C.I.} = (168.86, 1443.7)$$

$$4. \quad 95\% \text{ C.I.} = 50 \pm 1.96 (10/\sqrt{36}) = (46.733, 53.267)$$

$$99\% \text{ C.I.} = 50 \pm 2.57 (10/\sqrt{36}) = (45.707, 54.293)$$

CHAPTER FOUR

$$1. \quad Z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{55 - 50}{10/\sqrt{36}} = 3$$

Accept H_a , $3 > 1.64$

Engineers are significantly greater than population mean, $\alpha = .05$.

Accept H_a , $3 > 2.33$

Engineers are significantly greater than population mean, $\alpha = .01$.

2. $S_x = 620.263$ (calculator)

$$t = 3.878$$

$$P = .003$$

They are below the national average if $\alpha = .05$ or $\alpha = .01$.

3. $t = 1.96$

They are below the national average if $\alpha = .05$ [barely - $P = .045$]

If $\alpha = .01$, they are statistically equal to national average.

4. $Z = \frac{80 - 75}{10/\sqrt{100}} = 5$

$5 < 1.64$, $5 > 2.33$. Accept H_a , Alaska residents live longer than US mean,
 $\alpha = 5\%$ or 1%

CHAPTER FIVE

1. Redo with the TI-83

2. $Z = \frac{210 - 185}{\sqrt{50^2/100 + 40^2/60}} = 7.19$

Accept H_a - Philadelphia's mean house price is significantly greater than Atlanta -
 $\alpha = .05$.

3. $t = 2.25$

Accept H_a - Public school teachers make significantly more salary than private
school teachers - $\alpha = .05$

Accept H_0 - There is no statistically significant difference if $\alpha = .01$,
 $2.25 < t_{.01, 21df} = 2.52$.

4, $t = .87$

There is no statistically significant difference if $\alpha = .01$. $2.25 < t_{.01, 21df} = 2.52$

CHAPTER SIX

$$1. \quad Z = \frac{.625 - .5}{\sqrt{.5(.5)/40}} = \frac{.125}{.079} = 1.58$$

Accept H_0 . $P = .50$. There is no statistically significant difference from $P = .5$, $\alpha = .05$ or $\alpha = .01$

$$2. \quad P = .10$$

$$Z = \frac{.10 - 60/100}{\sqrt{.10(.9)/1000}} = \frac{.04}{.0095} = 4.21$$

Accept H_a . The claim is significantly too high. $\alpha = .05$

$$3. \quad Z = 1.82. \text{ Accept } H_a. \text{ Intel is significantly lower in rating - } \alpha = .05.$$

Accept H_0 . Intel and AOL have no significant difference in rating - $\alpha = .01$

$$4. \quad Z = 2.45$$

Accept H_a . Candidate A demonstrates significantly higher support - $\alpha = .05$ or $\alpha = .01$

CHAPTER SEVEN

$$1. \quad R = .996$$

$$R^2 = .99$$

This is extremely high direct variation.

2. (a)

	y	x
AOL	1	.54
EMC	1	.29
GE	3	.14
GM	5	-.03
IBM	4	.15

(b) $R = -.89$

$$y = -.7.5x + 4.4$$

This is an absurd equation that is indicative of the stock market crash to come.

3. $R = .98$

$$y = 15x + 15.4$$

$$\text{if } x = 2.5$$

$$y = 15(2.5) + 15.4 = 52.9$$

APPENDIX C

THE TRIAD – CONVERGENCE OF DATA MINING, QUALITATIVE RESEARCH AND THE STATISTICAL

Akira Kurosawa's masterpiece *Rashomon* depicted three dramatically different perspectives on a killing. We are left to wonder whether universal truth exists or perhaps, less profoundly, is eyewitness testimony reliable?

We propose the inverse, namely if three paths of research (data mining, qualitative research and the statistical) converge, this area of confluence is our strongest research result. This article introduces the Triad, which is the merging of data mining, statistics at all levels, and qualitative research in the form of focus groups to strengthen the confidence a fortiori in research results. If we tackle a research question with each of these relatively independent tools, and the Triad converges in supporting H_0 or in rejecting H_0 , we decrease the α level of the conclusion. We enhance the confidence in our results and establish a gold standard for psychometric research. In the words of Dr. Philip Merrifield (Professor of Educational Research – NYU), "If I was to study uncertainty (by leaving his original career in meteorology), it may as well be with people" (psychometrics).

We understand Dr. Merrifield's wise quip. Data mining, or mining large data sets for their strongest nuggets of golden insights, may test so many hypotheses that conventional alpha levels are of little value. By statistical we refer to conventional hypothesis testing of one or a few key research questions. Frequently, statistics tests the wrong hypotheses with data that do not conform to the assumptions underlying the Calculus derivation of the formula. Also, randomness is unattainable in our construction of random sample or of linear congruential

random numbers in Mathematical Modeling simulation. Qualitative research is not accepted by mainstream investigators as a valid means of inquiry. But these three tools synergistically provide the potential for a gold standard for psychometric research.

For simplicity and clarity, let us merge the two sample difference of means test (conventional statistics) with CHAID analysis (standard data mining algorithm).

STATISTICAL HYPOTHESIS TESTING

Consider the elementary two sample difference of means test. This widespread test relies upon a great many explicit and implicit assumptions, which include:

- 1) Both samples are from a well defined mathematical density function with mean and variance we are able to estimate. This is an act of faith.
- 2) Both samples are random. However, we cannot attain randomness even with the most refined linear congruential random numbers with carefully constructed modules and full periods.
- 3) Both samples are representative of each population. This assumption would take at least as much work as the typical research of a statistical hypothesis.
- 4) Both samples are taken from normal populations. Goodness of fit tests for normality is problematic, since the number of categories to measure the expected frequency of outcomes is arbitrary. Adding or subtracting numbers of categories can affect the outcome of Chi square goodness of fit tests.

One way to remedy this arbitrariness is the Kolmogorov-Smirnov Maximum Deviation Tests of a population distribution. We start with $F(x)$ as the cumulative distribution function of continuous random variable x , with n random observations. We let $S_n(x)$ be the empirical cumulative probability distribution function of the n observations. The distributions

$$A = \max [S_n(x) - F(x)]$$

$$B = \min [S_n(x) - F(x)]^*$$

are known, independent of $F(x)$, and leads to the acceptance or rejection of the hypothesized probability density function.

However, the assumptions underlying the K-S Test are randomness, continuity of sampled population, no tied observations, and a sample capable of being divided into n overlapping vertical strips each with $(1/n)$ of the distribution, with the hypothesized distribution specified completely, without regard to sample parameters so the K-S Test poses several foundational issues.

Of course, we can utilize the central limit theorem to justify normal distribution theory for n sufficiently large. However, the central limit theorem is based on $n \rightarrow \infty$. We do not know whether large samples are sufficiently large for asymptotic theory.

5) Last but not least, consider Kolmogorov's perspective on independence assumptions that underlie the moment generating function of a continuous random variable and the central limit theorem:

"Thus one comes to perceive, in the concept of independence, the first germ of the true nature of problems in probability theory."

"Dependence and the Central Limit Theorem" took Tim Sheehan and me two years to model dependence and its effect on autocorrelation for very specific probability densities and specific levels of dependence (autocorrelation). Dr. Lehmann (U of California - Berkeley) was very generous in the mathematical statistical expertise he shared with the authors. Dr. Lehmann envisioned a time where Mathematical Statisticians and Computer Scientists adjust alpha (Type I

error) levels to dependence in one or both samples. Statistical Modeling is the key to the fulfillment of Dr. Lehmann's wise vision.

DATA MINING

Data mining, or the mining of quantitative relationships from large sets of data, has spread rapidly in recent years. For clarity and brevity, let us focus on CHAID analysis. We rely on material generously sent by Dr. Jay Magidson for this section of our paper. Dr. Magidson is a nationally recognized leader in the intention, development and application of CHAID.

CHAID or CHI-squared Automatic Interaction Detection was developed in 1978 by Kass to combine categories of a given variable that do not differ significantly from each other. It separates those categories that are different. CHAID applies a bonferroni multiplier to adjust for simultaneous inference. Lastly, only statistically significant variables can split a group.*

The CHAID algorithm is described by Dr. Jay Magidson in Appendix I. It is clear that two statistical tools, foundational to CHAID are the chi-square independence test and the Bonferroni adjustment.

Consider the following sequence in the development of the statistical foundation for the chi-square independence test:

- 1) If X_1, X_2, \dots, X_n are independent random variables having standard normal distributions, then $Y = \sum_{i=1}^n X_i^2$ has the chi-square distribution with n degrees of freedom.
- 2) CHAID explores $r \times c$ contingency tables with the null hypothesis that the row variable and column variable are independent.

If P_{ij} is the probability that an element will belong to the i th row and j th column, we wish to test whether $P_{ij} = P_i \cdot P_j$ for $i = 1, 2, \dots, r$, and $j = 1, 2, \dots, c$. This alternative hypothesis is that $P_{ij} \neq$

* Magidson, Jay, "The CHAID Approach to Segmentation Marketing," prepared for Richard Bagozzi (ed.) *Advanced Marketing Research*, Blackwell forthcoming, p. 6.

$P_i \cdot P_j$. Under H_0 , we assume that the row variable and column variable are assumed independent and the expected frequencies are derived by:

Let f = grand total of frequencies

f_i = row total

f_j = column total

$l_{ij} = f_i/f \cdot f_j/f \cdot f = f_i f_j / f$

Our familiar X^2 Test for Independence statistic is computed:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c (f_{ij} - l_{ij})^2 / l_{ij}$$

We reject H_0 if computed $X^2 > X^2_{\alpha, (r-1)(c-1) \text{ df}}$.

To explore the foundations of CHAID, we need to analyze the origins of the chi-square distribution. To derive the chi-square test for goodness of fit, we start with X_1, X_2, \dots, X_k observed values of k independent random variables X_1, X_2, \dots, X_k each having binomial distributions with parameters N_1 and θ_1, N_2 and $\theta_2 \dots, N_x$ and θ_x . If n is sufficiently large, we can use

$$Z_i = \frac{X_i - n_i \theta_i}{\sqrt{n_i \theta_i (1 - \theta_i)}} \quad \text{to approximate}$$

with standard normal distributions the distribution of the independent random variable Z_i .

By mathematical statistics, we use moment generation functions to prove that if X_1, X_2, \dots, X_n are independent random variables, each with standard normal distributions, then

$$Y = \sum_{i=1}^n X_i^2 \text{ has the chi-square distribution with } v = n \text{ degrees of freedom.}$$

Therefore

$$X^2 = \sum_{i=1}^k \frac{(x_i - n_i \theta_i)^2}{n_i \theta_i (1 - \theta_i)}$$

leads to the well known chi-square goodness of fit test. By utilizing the previously discussed formula:

$$l_{ij} = \theta_i \theta_j \cdot f = f_i / f \cdot f_j / f \cdot f = \frac{f_i \cdot f_j}{f}$$

We compute the expected l_{ij} and then employ the standard X^2 test

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - l_{ij})^2}{l_{ij}}$$

for the $r \times c$ table with degrees of freedom $(r-1)(c-1)$.

It is clear that CHAID relies upon several assumptions, which include:

- 1) The independence of X_1, X_2, \dots, X_k observed values of k independent random variables X_1, X_2, \dots, X_k .
- 2) Each X can take on $k > 1$ categories and follows a multinomial distribution (if the variable is declared "FREE" or nominal) or an order-restricted multinomial distribution for X 's declared ordinal.
- 3) That n be sufficiently large. This usually means $n \rightarrow \infty$ to utilize asymptotic normality.

Like the Calculus, CHAID is beautiful and elegant. However, both suffer from typically neglected foundational issues. Few mathematicians would gain tenure at major universities by highlighting the failed effort of the 20th century in constructing the irrationals from the rationals without problematic or circular assumptions. Perhaps statisticians are more open to discussing the foundational issues in data mining, even if they are not solved.

Each of these three assumptions, which support CHAID analysis, is a potential problem. All samples are somewhat dependent. Even random numbers generated by a computer are called pseudorandom, because their linear congruential generators are intrinsically correlated.

The assumption that each set of data frequencies has a binomial distribution is based on the independence assumption and further assumes a sufficiently large sample ($n \geq 30$) for validity.

The third assumption, that $n \rightarrow \infty$, ensures asymptotic normality. However, we have little knowledge of the departure from normality if n is large (say $n = 100$).

As previously discussed, Loase and Sheehan (1989)* developed computer simulations designed to estimate enhanced alpha levels resulting from induced autocorrelation. The magnitude of enhanced alpha surprised the authors. Further, the simulation was highly specific as to the statistical distribution used and bias induced. No neat generalization emerged linking autocorrelation to enhanced alpha levels across different mathematical distributions.

The adjustment of alpha level to departures from independence, binomial distribution, and sufficiently large n in CHAID analysis would be a much more problematic endeavor. This is a futuristic concern, waiting for interdisciplinary research among Statisticians, Mathematicians and Computer Scientists.

* Loase, J. and Sheehan, T., "Dependent Random Variables and the Central Limit Theorem," Fall 1989, International Journal of Mathematical Modeling.

Bonferroni Adjustment

The Bonferroni correction may even prove a greater foundational problem to CHAID. In CHAID the computer is systematically employing a number of chi-square independence tests. If you do not adjust the alpha level (typically .05), you would reject H_0 5% of the time by chance, even if H_0 were true. Consequently, the Bonferroni method adjusts downward each individual test's alpha level to guarantee 5% Type I error for the final CHAID result.

Magidson (1992) illustrates the Bonferroni adjustment assuming a two way table and testing independence 15 times. Assuming independence, $1 - (1 - \alpha)^{15} > \alpha$ leads us to 15 as the Bonferroni multiplier. More generally $1 - (1 - \alpha)^m = m\alpha$ for small α . The adjusted p value is the conditional p value times the Bonferroni multiplier. Also, the tests, for example 15 chi-square independence tests, are not independent. Therefore, the Bonferroni adjustment is conservative.*

Another difficulty with the Bonferroni method is that we have adjusted Type I error but neglected the consequent increase in Type II error. Perhaps this objection is mitigated by CHAID's goal of obtaining the strongest rejections of the null hypothesis. Two recommended future research directions are:

- 1) To adjust alpha level to the three TRIAD outcomes.
- 2) To quantify the Type II error associated with CHAID analysis and to devise ways to lower B without compromising the alpha level of the test.

* Madison, J., "The CHAID Approach to Segmentation Modeling: CHI-Squared Automatic Interaction Detection," Chapter 4 in Richard Bagozzi (ed.) *Advanced Methods of Marketing Research*, Blackwell, 1994, pp.118-159.

Loase and Sheehan (1989) focused on a computer modeling approach to link alpha level with levels of induced autocorrelation. New thinking is necessary to explore computer modeling Type II error, since we have an infinite number of possible Z scores > 1.64 to establish H_a . Type I error is easier to model since we can assume H_0 or that the two parameters are equal. Type II error requires the researcher to select a non-zero difference between the parameters. Perchance we can fix the difference based upon an initial sample result. However, the resulting mathematical model would be specific to this restriction as well as assumption related to the density functions of the parameters in the mathematical model and the linear congruential random number generator.

QUALITATIVE RESEARCH

The subjectivity of qualitative research is well known. A recent forum of researchers interested in the qualitative dimension found a home on the internet; see www.qualitative-research.net/fqs-texte/2-03/2-03intro-l.e.htm

Focus groups traditionally consist of 8-10 people, discussing a topic of interest for 90-120 minutes. The authority role of the moderator, interaction among participants, ability to (in our case) validate prior research or plan future research, utilize non-verbal cues, and achieve 100% attention make this tool very effective.

The negatives are the possible influence of leader or one or two dominant group members, difficulty with sensitive topics, and problems with translating the qualitative into statistical measures.*

* Focus Group Research, Quirk's Marketing Research Review, June 2003, p. 1-6, www.groupsplus.com, internet paper.

We have successfully used a focus group to validate the strongest statistical findings from our data mining and statistical elements of our Triad. In one instance, the focus group led us to test several hypotheses that were indeed confirmed. It is noteworthy and remarkable that independence is required and essential in each of the components of the Triad.

CONCLUSION

In our scientific endeavor to search for truth, each of the elements of the Triad poses great promise and foundational problems. We propose the Triad as the “gold standard” for scientific research. The convergence of all three elements of the Triad may be the closest we will come to the elusive concept Truth in the research enterprise. It is our hope that future researchers will utilize the Triad and in time develop mathematical models to adjust alpha and beta to convergent findings from these three relatively independent research lines. The intersection of assumptions of different tests and the consequent influence on adjusted p value is a complex issue that may require a mathematical model. The quantification of the results of qualitative research is likewise a deep and profound issue that requires partnership across disciplines to address and solve. The synergies of the Triad provide us with great promise in tackling the complex research issues of the present and future.

APPENDIX I

(Sent to authors by Dr. Jay Magidson, Statistical Innovations, Inc., July 14, 2004)

The CHAID Algorithm

The primary statistical algorithm used in SPSS PC+ CHAID 5.0 consists of three stages – merging, splitting and applying the stopping rule. It may be formally described as follows:

STAGE 1: Merging

For each predictor, x_1, x_2, \dots, x_k

1. Form the full 2-way cross tabulation with the dependent variable.
2. For each pair of categories that is eligible to be merged together, compute chi-square statistics (which will be referred to as “pairwise chi-squares”) to test for independence in the $2 \times J$ subtable formed by that pair of categories and the dependent variable which has J categories. (If the ordinal method is selected, use the procedure described in the section “Estimation of the p-value under the ORDINAL method of Analysis” below to compute the chi-square statistics.)
3. For each pairwise chi-square, compute the corresponding “pairwise p-value.” Among those pairs that are found to be non-significant (i.e., pairwise p-value $>$ *category significance level*), merge the most similar pair (i.e., that pair having the smallest pairwise chi-square value) into a single joint category, and go to step 4. If all remaining pairs are significant, go to step 5. While the default value for the *category significance level* is .05, it may be changed globally for all predictors by setting the technical parameter “merge level.” In addition, certain predictors may be given a different *category significance level* than others by setting the predictor-specific “merge level” in the predictor’s box of the setup menu.
4. For any joint category containing 3 or more categories, test to see if any predictor category should be unmerged by testing the significance associated with that category vs. the others in that joint category. If a significant chi-square obtains, unmerge that category from the others. If more than one category is eligible to be unmerged, unmerge the one having the highest chi-square. Return to step 3.
5. If the *minimum subgroup size after split* (MINSEG) $>$ 0, merge any category having fewer than MINSEG observations with the most similar other category (as measured by the smallest chi-square).
6. Compute the Bonferroni adjusted p-value based on the “optimally merged” categories.

STAGE 2: Splitting

7. Among those predictors that have an adjusted p-value which is statistically significant, select as best that predictor with the lowest p-value and split the group on this predictor (i.e., use each of the optimally merged categories of that predictor to define a subdivision of the parent group into a new subgroup). If no predictor has a significant p-value, do not split the group. (By default, the significance level is set to $\alpha = .05$. This default level can be changed by using the TECHNICAL command with the SETUP menu.

STAGE 3: Stopping

8. Return to step 1 to analyze the *next* subgroup having at least *minimum subgroup size before split* (MINGRP) observations. Stop when all subgroups have either been analyzed or contain too few observations.