STATISTICAL METHODS FOR WSDOT PAVEMENT AND MATERIAL APPLICATIONS

WA-RD 315.1

Interim Report
February 1994

Washington State
Department of Transportation

Washington State Transportation Commission
Planning and Programming Service Center
in cooperation with the U.S. Department of Transportation
Federal Highway Administration

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO. WA-RD 315.1	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
4. TITLE AND SUBTITLE STATISTICAL METHODS FOR WSDOT PAVEMENT AND MATERIAL APPLICATIONS		5. REPORT DATE February 1994
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Joe P. Mahoney		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Washington State Transportation Center (TRAC) University of Washington, JD-10 University District Building; 1107 NE 45th Street, Suite 535 Seattle, Washington 98105-4631		10. WORK UNIT NO.
		11. CONTRACT OR GRANT NO. T9903, Task 6
12. SPONSORING AGENCY NAME AND ADDRESS Washington State Department of Transportation Transportation Building, MS 7370 Olympia, Washington 98504-7370		13. TYPE OF REPORT AND PERIOD COVERED Interim report
		14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This study was conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration. WSDOT technical contacts were Mr. R. Keith Kay and Mr. Newton C. Jackson.

16. ABSTRACT

This document is an upgraded version of Research Report WA-RD 143.2 (February 1988) and is principally the result of continuing work on the WSDOT Pavement Management System (WSPMS). Various statistical methods have been used in the WSPMS and the research team wanted to summarize some of these applications in a separate, straightforward report which may be of value to a broader group of users.

17. KEY WORDS statistics, regression, materials, pavements, models	18. DISTRIBUTION STATEMENT No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
None	None	107	

Interim Report
Research Project T9903, Task 6
Pavement Management System Modification—Phase II

STATISTICAL METHODS FOR WSDOT PAVEMENT AND MATERIAL APPLICATIONS

by

Joe P. Mahoney
Professor of Civil Engineering
121 More Hall, FX-10
University of Washington
Seattle, Washington 98195

Washington State Transportation Center (TRAC)
University of Washington, JD-10
University District Building
1107 NE 45th Street, Suite 535
Seattle, Washington 98105-4631

Washington State Department of Transportation
Technical Monitors

R. Keith Kay
Pavement Management Engineer

Newton C. Jackson
Pavement and Soils Engineering

Prepared for

Washington State Transportation Commission
Department of Transportation
and in cooperation with
U.S. Department of Transportation
Federal Highway Administration

February 1994

DISCLAIMER

TABLE OF CONTENTS

Section **Page**

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

Revised February 14, 1994

LIST OF TABLES

SECTION 1.0

INTRODUCTION

1. INTRODUCTION

This report is an updated version of WSDOT Research Report WA-RD 143.2 [1.4] (Regression Analysis for WSDOT Material Applications) which was published in February 1988. Subsequent to the first report, the following topics have been improved or added:

- expanded information on hypothesis tests (SECTION 3.0),
- expanded information on regression models and associated transformations (SECTION 4.0), and
- new section on sampling (SECTION 5.0).

The types of examples used in all sections relate to pavement performance or pavement materials.

2. DEFINITIONS

- STATISTICAL [1.1]

  "Having to do with numbers" or "drawing conclusions from numbers."

- POPULATION [after Ref. 1.2]

  All measurements or counts that are obtainable from all of the objects that process some common characteristic. Example: a "population" of data would be the pavement condition measured on <u>all</u> Interstate highways in a specific state.

- SAMPLE [1.2]

  A set of measurements or counts that constitute a part (or all) of the population.

- RANDOM SAMPLING [1.2]

  A sampling procedure whereby any one measurement in the population is as likely to be included as any other.

- BIASED SAMPLING [after Ref. 1.2]

  A sampling procedure whereby certain individual measurements have a greater chance of being included than others. Example: biased sampling would be taking density measurements only at places on a base course that appeared to be well compacted.

- MEAN

  Average of a group of measurements. The population mean is designated "μ" and a sample mean by "$\bar{x}$."

- MEDIAN [1.2]

  The number, in a set of numbers arranged in ascending order, that divides the set so that half of the numbers are higher and half are lower.

- RANGE

  The largest measurement minus the smallest measurement in a group of data.

- STANDARD DEVIATION

  A measure of variation or dispersion of a group of data. Specifically, the average of the squares of the numerical differences of each measurement (or observation) from the mean. The population standard deviation is designated by "σ" and a sample standard deviation by "s."

- HISTOGRAM

  A graphical form of data presentation. A bar chart that shows in terms of area the relative number of measurements of different classes. The width of the bar represents the class interval, the height represents the number of measurements.

- VARIABLE [1.3]

  A quantity to which any of the values in a given set may be assigned, i.e., something on which measurements are made.

- CORRELATION

  A way to measure the association between two variables.

- REGRESSION

  Goes a step further than correlation. Generates an equation that can be used to <u>predict</u> one variable from another (or others in multiple regression). The <u>predicted</u> variable is the <u>dependent</u> variable and the other variables are called <u>independent</u> variables.

- SYSTEMATIC SAMPLING [1.2]

  Selection of successive observations at uniform intervals in a sequence of time, area, etc. Example: taking pavement deflection measurements every 500 ft. on a project.

- Examples for the calculation of sample mean, sample standard deviation, range and a histogram are shown in Table 1.1 and Figure 1.1.

Table 1.1. Calculation of Sample Mean, Sample Standard Deviation and Range for Procter Density Data [after Ref. 1.2]

• Basic data (Procter density) in pcf:

107.5	100.8	107.0	101.5	107.0
112.0	111.4	124.0	103.3	101.3
104.3	109.4	103.5	114.1	98.0
106.0	99.7	110.5	105.0	93.5
101.3	102.5	95.5	94.0	110.1

• Greek symbol "Σ" indicates a summation calculation is required. To sum the 25 density test results above:

$$\sum_{i=1}^{25} x_i = 107.5 + 112.0 + 104.3 + \ldots + 93.5 + 110.1 = 2,623.2$$

• Sample mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2,623.2}{25} = 104.93 \approx 105 \text{ pcf}$$

• Sample standard deviation (s)

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

$$= \sqrt{\frac{(107.5 - 104.9)^2 + (112.0 - 104.9)^2 + \ldots + (110.5 - 104.9)^2}{25-1}}$$

$$= \sqrt{\frac{1114.27}{25-1}} = 6.8 \text{ pcf}$$

• Range

Range = largest density - smallest density

$= x_{max} - x_{min}$

= 124.0 - 93.5 = 30.5 pcf

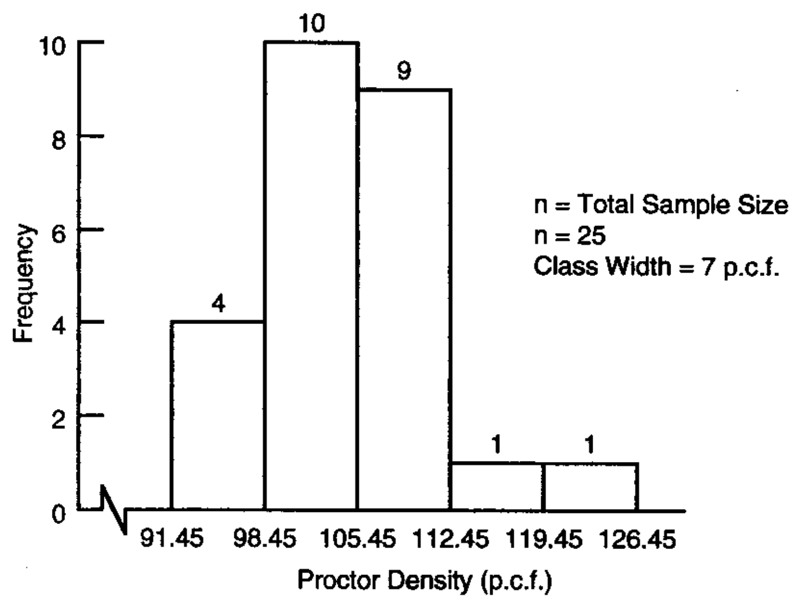


Figure 1.1. Frequency Histogram for Proctor Density Data [after Ref. 1.2]

SECTION 1.0
REFERENCES

1.1 Western Electric Co., Inc., Statistical Quality Control Handbook, Western Electric Co., Inc., AT&T Technologies, Indianapolis, Indiana, May 1985 (11th printing).

1.2 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

1.3 ASTM, Standards on Precision and Bias for Various Applications, American Society for Testing and Materials, Philadelphia, PA, 1985.

1.4 Mahoney, Joe P., "Regression Analysis for WSDOT Material Applications," Research Report WA-RD 143.2, Washington State Department of Transportation, Olympia, Washington, February 1988.

SECTION 2.0

THE NORMAL DISTRIBUTION

2.1 INTRODUCTION

The normal distribution is a data distribution that can be used to describe many types of measurements in engineering. Basically, a normal distribution is a bell shaped curve. The role of the normal distribution in statistics has been stated to be analogous to the role of the straight line in geometry. Figure 2.1 illustrates a bell curve, superimposed over a histogram of PCC compressive strength data. Such a distribution is very convenient to use because it is characterized by the mean (μ or $\bar{x}$) and standard deviation (σ or s). As Figure 2.1 shows, most of the strength measurements cluster around the mean ($\bar{x}$ = 4,824 psi), while fewer measurements are near the lowest (3,875 psi) and highest (5,975 psi) strength values.

Since the normal distribution can be defined by the mean and standard deviation, a set of measurements with equal means but differing standard deviations can be illustrated, as shown in Figure 2.2 (subgrade density measurements). In this case, the population mean is equal to 105 pcf and three different normal distributions are shown for population standard deviations of 5, 7 and 9 pcf. If you were a field inspector, which of these three distributions would you prefer?

Figure 2.2 helps to provide an answer to the above question. If the total area under the bell shaped curve is equal to 1.0, then the portion of density tests between 90 and 96 pcf is about 3.5, 8.0 and 11.0 percent for the three standard deviations of 5, 7 and 9 pct, respectively. This suggests that the distribution with σ = 5 pcf is preferable. You will see how to determine these areas later in this section.

Willenbrock [2.1] (slightly modified) helps to explain Figures 2.2 and 2.3.

> The theoretical NORMAL DISTRIBUTION extends out infinitely in both directions from a mean of 105 pcf and never quite reaches the horizontal axis...A NORMAL DISTRIBUTION has a total area under the curve of 1.00 (i.e. 100 percent of the data values are represented by the distribution). Since it extends from $-\infty$ to $+\infty$ (minus infinity to plus infinity), it encompasses all of the density results that can occur. The area under the curve within these two limits must therefore be equal to unity (i.e. 1.000 or 100 percent). For all practical purposes, however, most of the data values (actually 99.73 percent) occur between 3 σ limits below 105 pcf and 3 σ limits above 105 pcf.
>
> If the area of each NORMAL DISTRIBUTION is the same (i.e., an area equal to unity, 1.0000), then the distribution shown in Figure 2.2 that has the largest spread (i.e., the largest standard deviation, which occurs in case (c) where σ = 9 pcf) should have the shortest overall height at the average value. Normal distribution (a), on the other hand, has the smallest spread (i.e., σ = 5 pcf), so its horizontal spread is smaller than its distribution (c). Its vertical spread must therefore be larger than its distribution (c).
>
> A far more important result than those mentioned above is also related to the fact that the area under the curve is equal to 100

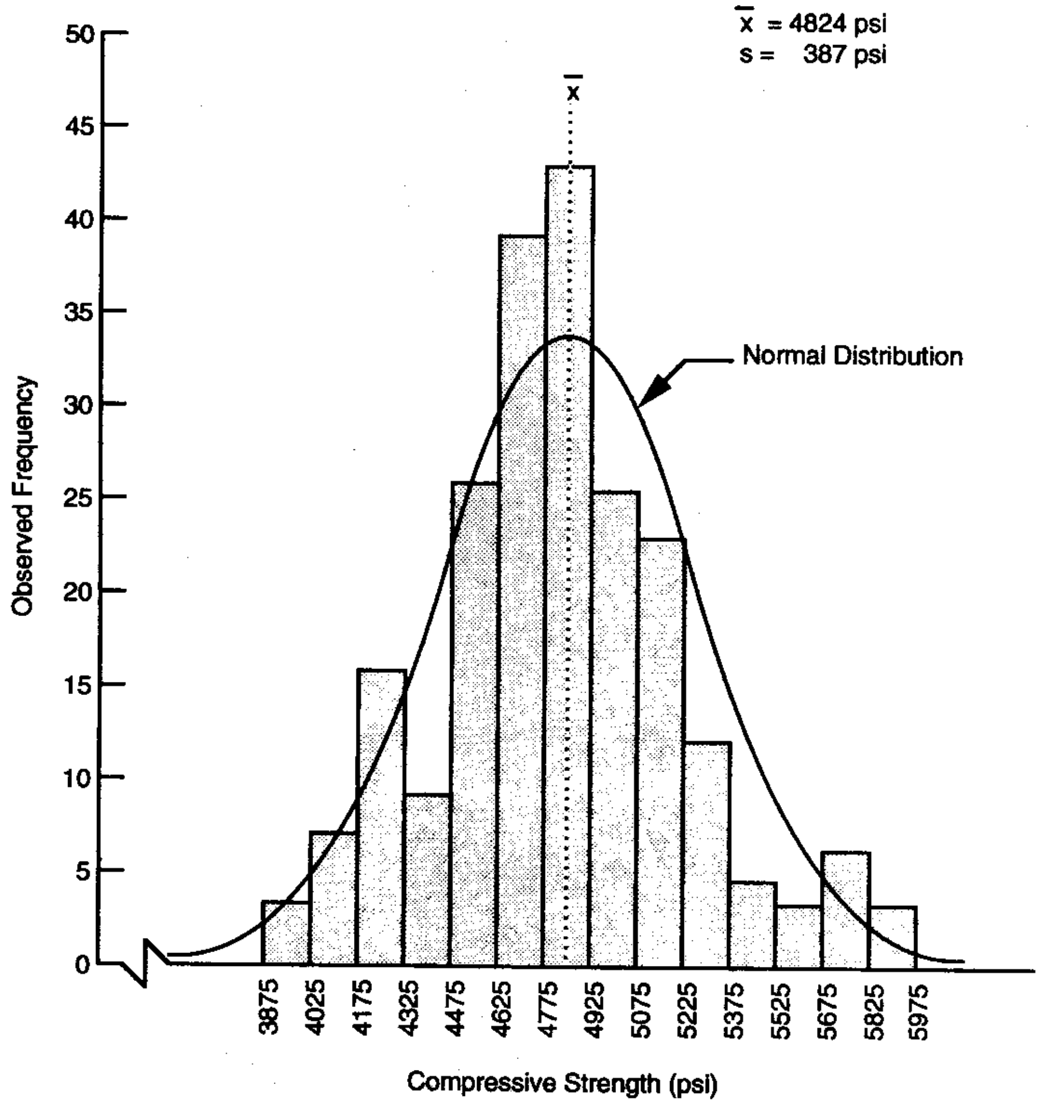


Figure 2.1. Histogram and the Normal Distribution for PCC Compressive Strength Data [after Ref. 2.1]

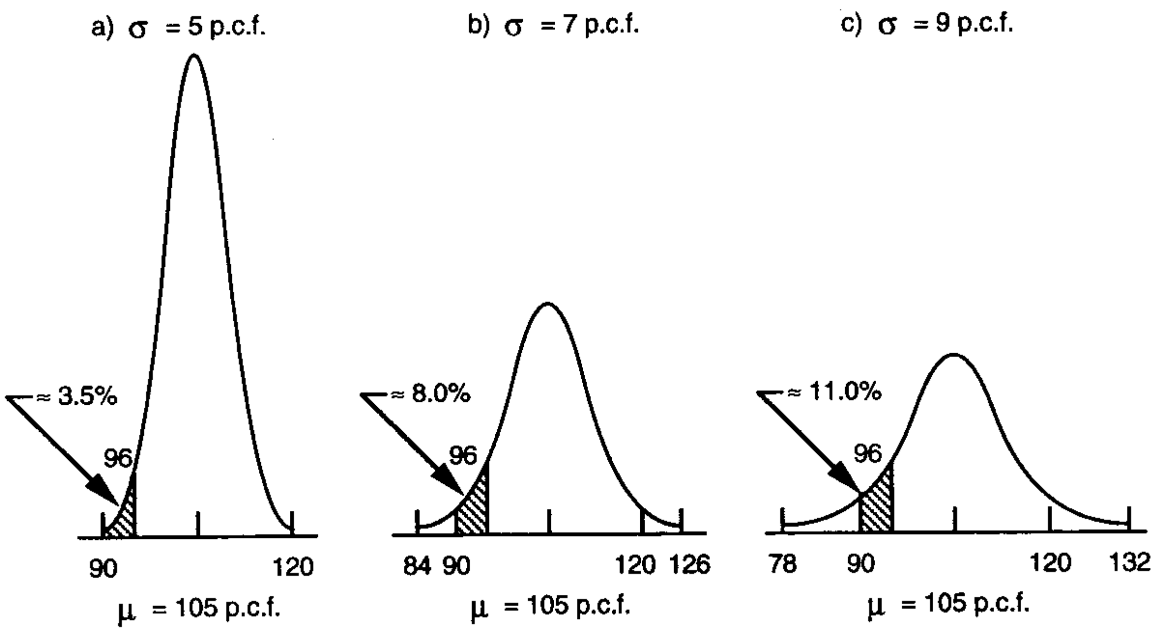


Figure 2.2. Three Normal Distributions for Procter Densities
(Same Means, Different Standard Deviations)
[after Ref. 2.1]

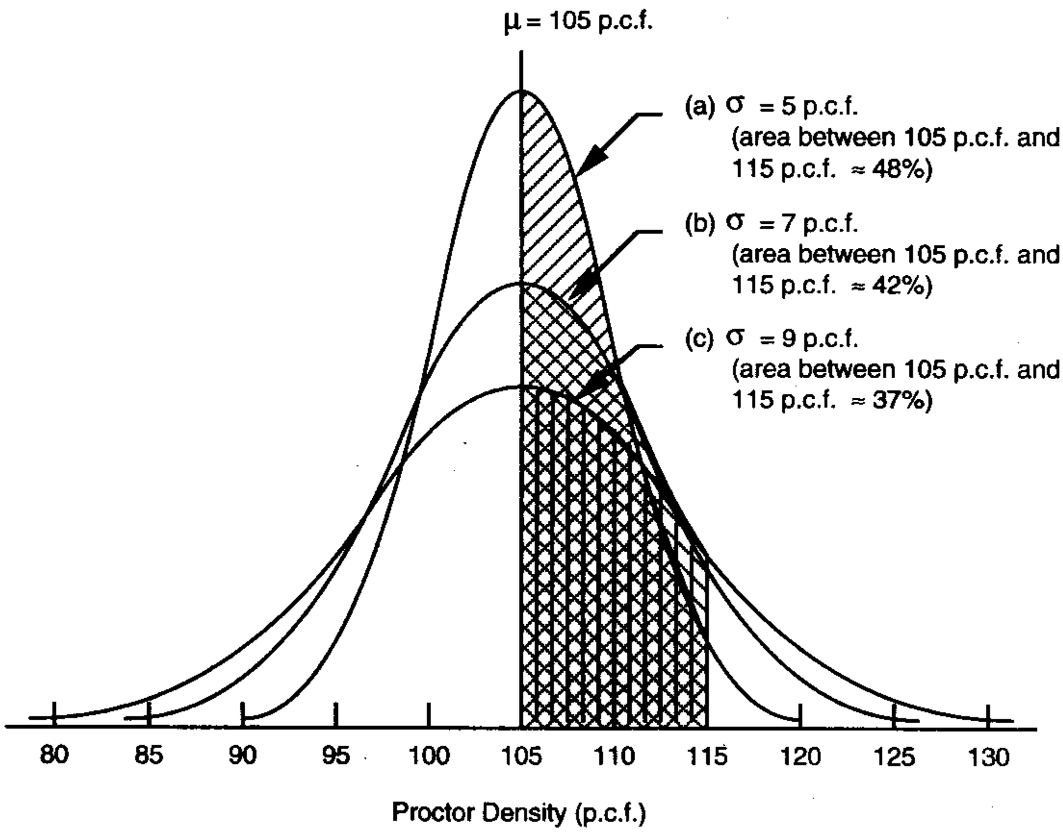


Figure 2.3. Superimposed Normal Distributions [after Ref. 2.1]

percent. Because of this, it can be stated that THE PROBABILITY OF FINDING A DATA VALUE BETWEEN 105 pcf and 115 pcf IS EQUAL TO THE AREA UNDER THE NORMAL DISTRIBUTION BETWEEN 105 pcf AND 115 pcf. For distribution (a), as shown in Figure 2.3 (i.e., $\sigma = 5$ pcf), the area between 105 pcf and 115 pcf represents about 48 percent of the area under the entire distribution.

Figure 2.4 shows two normal distributions with equal population standard deviations ($\sigma = 5$ pcf) but unequal population means ($\mu = 85$ and 105 pcf).

2.2 NORMAL DISTRIBUTION EQUATION

The height of a normal distribution (y) can be defined by its corresponding value of x (refer to Figure 2.5) by the following equation [after Ref. 2.1]:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{Eq. 2.1}$$

where

y = vertical height of a point on the normal distribution,
x = distance along the horizontal axis,
σ = standard deviation of the data distribution,
μ = mean of the data distribution,
e = constant = 2.71828 ... ,
π = constant = 3.14159 ....

To illustrate how Equation 2.1 can be used to determine area under a normal distribution, refer back to the Proctor density data (Figure 2.3). Calculate the area under the normal curve between 105 and 115 pcf for a standard deviation of 5 pcf (this is shown in Figure 2.3 to be 0.48 or 48 percent of the total area under the curve). These calculations are

$$y = \frac{1}{(5)\sqrt{2\pi}} e^{-(105-105)^2/2(5)^2} = 0.079788$$

and

$$y = \frac{1}{(5)\sqrt{2\pi}} e^{-(115-105)^2/2(5)^2} = 0.010798$$

The approximate area under the curve is about 0.45 (or 45 percent), which is close to the "theoretical" value of 48 percent (refer to sketch in Figure 2.6). The significance of this value is that the probability of a density measurement falling within the range of 105 to 115 pcf is about 0.48 (let's use the "theoretical" value).

To determine such probabilities in this manner is tedious and time consuming. There is an easier way to determine these probabilities than computing and tabulating y's for various μ's and σ's. To do this, you must convert the normal distribution to a standard normal distribution and define a variable "z," which is:

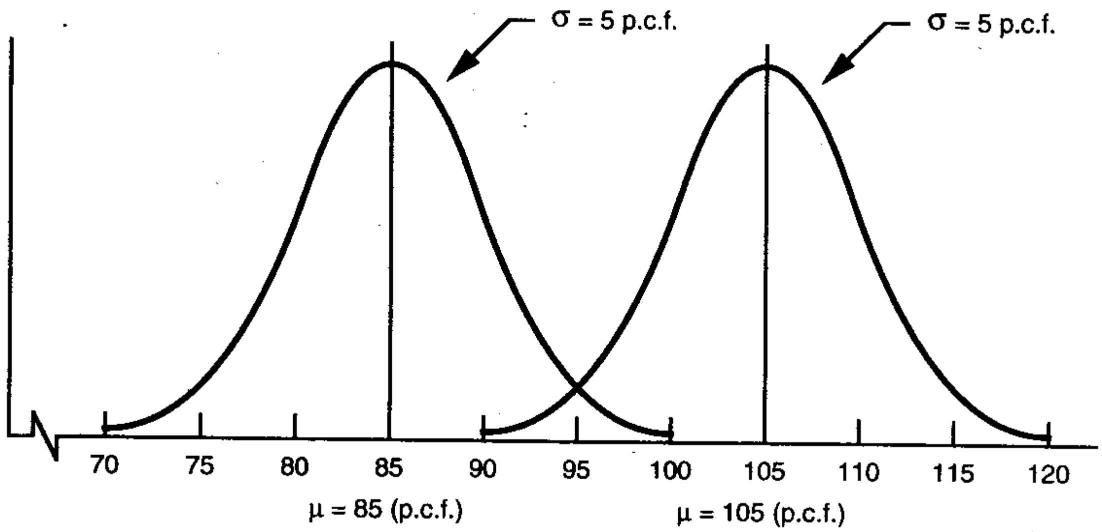


Figure 2.4. Normal Distribution (Different Means, Same Standard Deviations) [after Ref. 2.1]

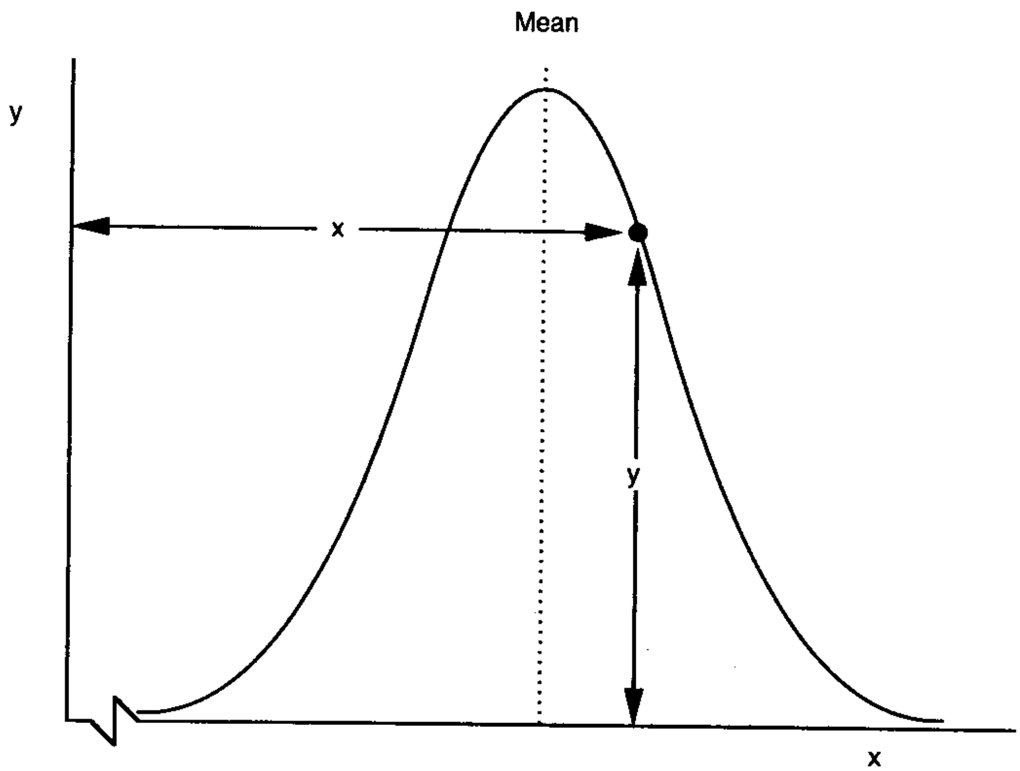


Figure 2.5. Relationship of "y" and "x" Values in the Normal Distribution

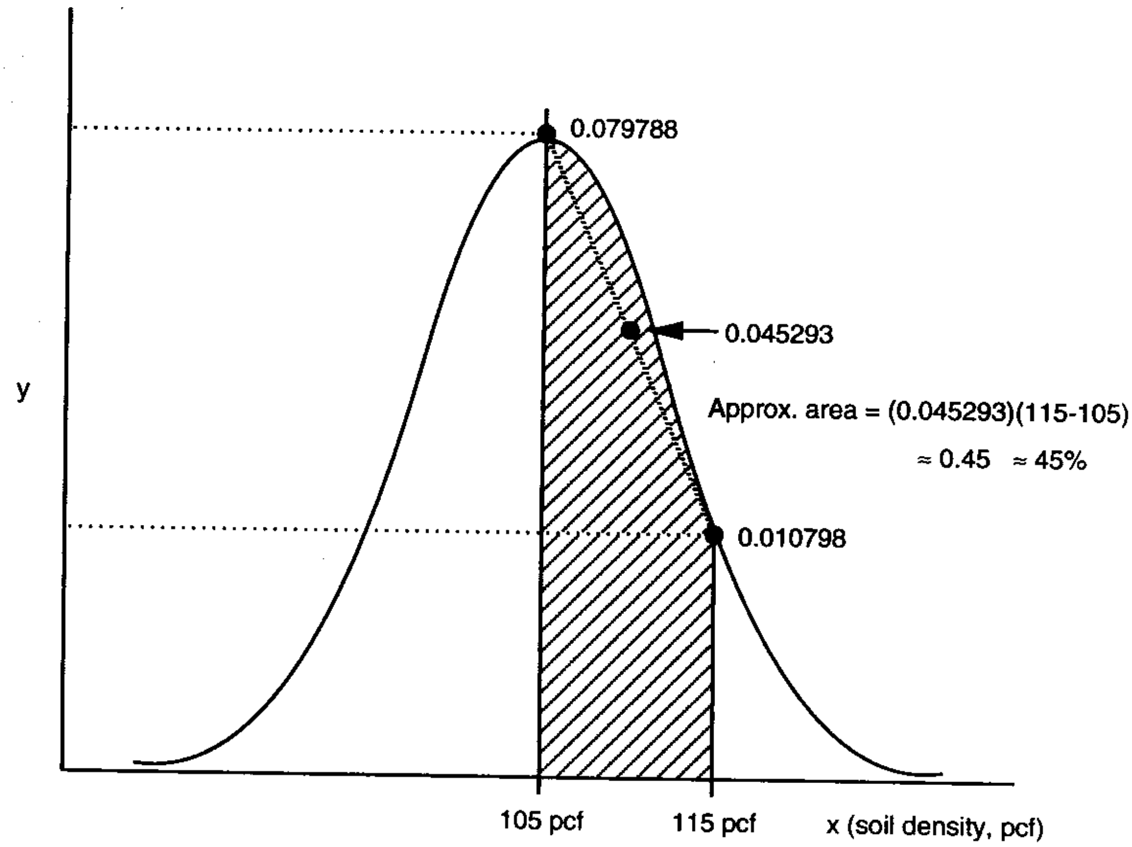


Figure 2.6. Determination of Approximate Area
Under the Normal Distribution

$$z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

If you substitute z into Equation 2.1, then following relationship results:

$$y_Z = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{Eq. 2.2}$$

where

y_Z = vertical height on the standard normal distribution, and
z = as previously defined.

Refer to Figure 2.7, which illustrates this important transformation. Thus, you can see that the probability of having a density test between 105 and 115 pcf is about 47.7 percent (or 34.1 + 13.6 percent). Fortunately, the "z-statistic" has been published in tables to allow for easy computation. Such a table is shown as Table 2.1. You can see that

mean ± 1 standard deviations ≅ 68.2% of area
mean ± 2 standard deviations ≅ 95.4% of area
mean ± 3 standard deviations ≅ 99.8% of area

Recall that all of the area under a normal distribution is 100%.

2.2.1 Example 1: Proctor density data

For normally distributed Proctor density data with μ = 105 pcf and σ = 5 pcf, what is the probability the density will be greater than 92 pcf?

First calculate z.

$$z = \frac{92 - 105}{5} = -2.6$$

Now, with z = -2.6, use a cumulative standard normal distribution table (any statistics book will have one, or use Table 2.1) to obtain the appropriate area under the curve that equals 0.0047.

Thus, P(density ≥ 92 pcf) = 1.0000 - 0.0047
= 0.9953 or 99.53 percent

2.2.2 Example 2: Portland Cement Concrete Strength

If a distribution of PCC strength data is μ = 5000 psi and σ = 500 psi, answer the following questions:

(a) What is the probability it will be more than 4,000 psi?

(b) What is the probability it will be less than 4,000 psi?

Refer to Figure 2.8 for the results.

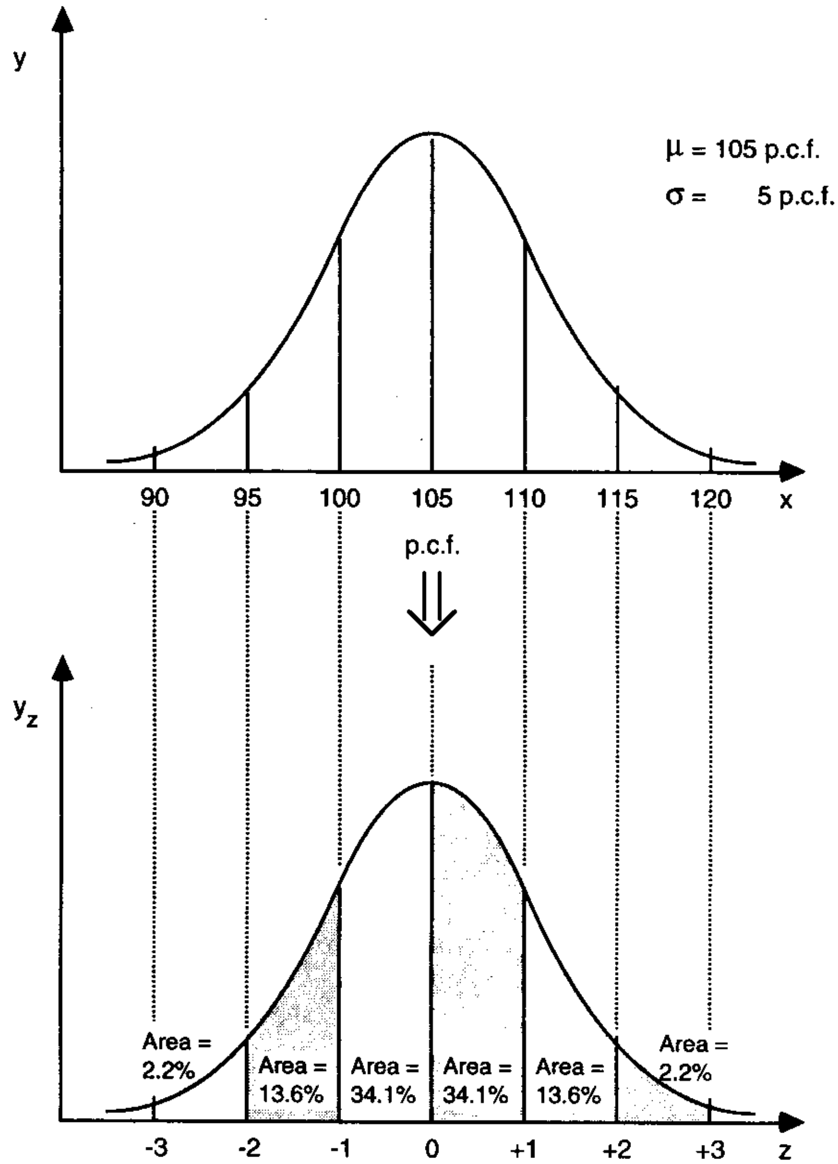
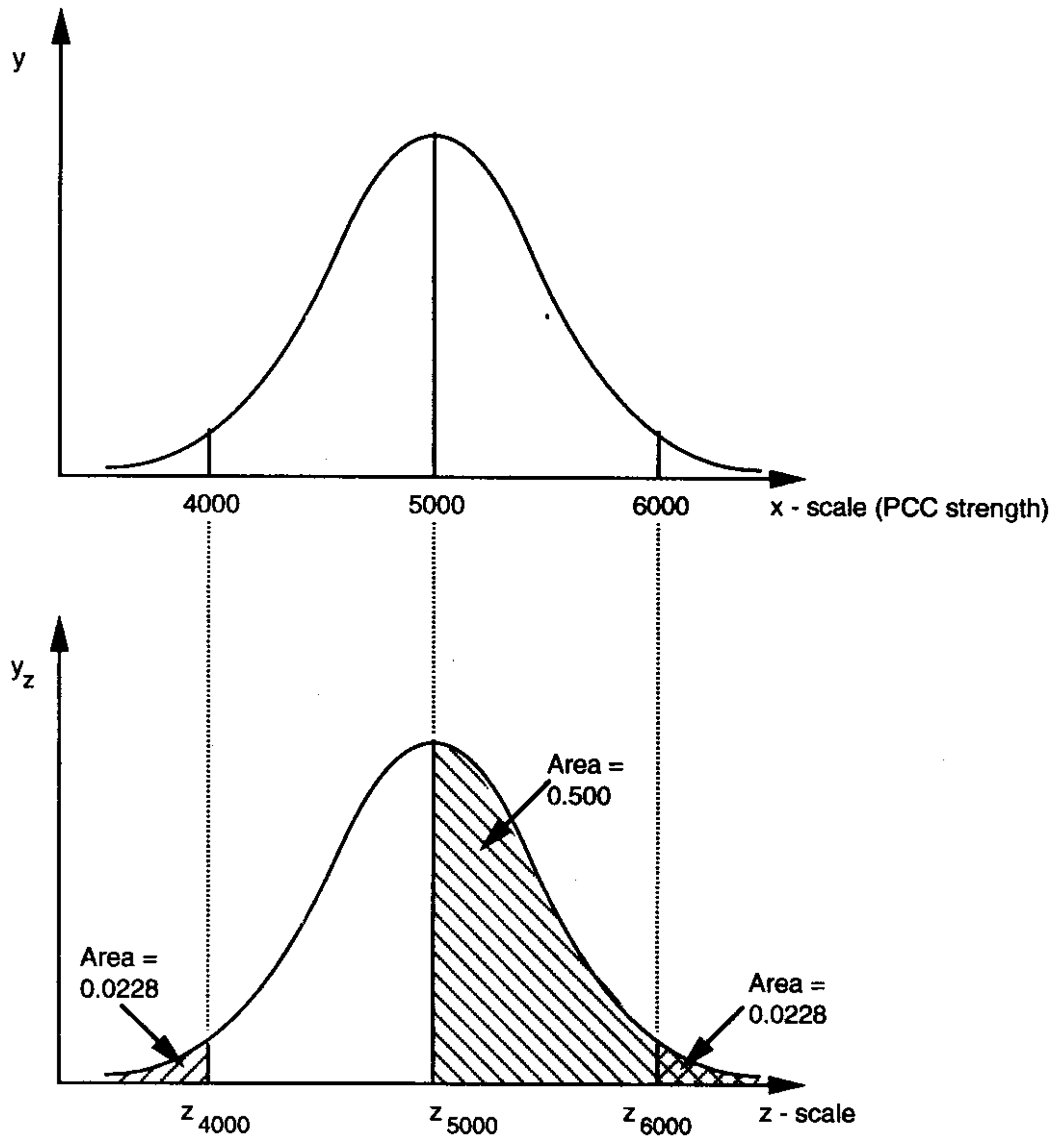


Figure 2.7. The Proctor Density Distributions (Normal and Standard Normal)



From Table 2.1

$$z_{4000} = \frac{4000 - 5000}{500} = -2$$ (area less than z = -2 equals 0.0228)

$$z_{5000} = \frac{5000 - 5000}{500} = 0$$ (area less than z = 0 equals 0.5000)

$$z_{6000} = \frac{6000 - 5000}{500} = +2$$ (area less than z = +2 equals 0.9772)

(a) P (strength $\geq$ 4000 psi) = 1.000 - 0.0228 = 0.9772 or 97.72%

(b) P (strength $\leq$ 4000 psi) = 0.0228 or 2.28%

Figure 2.8. PCC Strength Probabilities

2.2.3 Example 3: Ready-Mix Concrete

The contractor claims that the batch plant can produce PCC mix with

$\mu = 4{,}824$ psi
$\sigma = 387$ psi

Assume that a very unrealistic job specification states that an acceptable PCC must have a compressive strength no lower than 4,700 psi and no higher than 5,000 psi (after seven days of cure).

Question: If the contractor sends 50 truckloads of this mix to the job site, how many of the trucks should be rejected if you know the real, potential compressive strength of each truckload (of course you cannot do this but what the heck)?

Solution: Using z-statistic tables, you find that the total area under the standard normal distribution between 4,700 and 5,000 psi is about 0.30 (recall that the maximum is 1.0000 under the curve).

Thus, approximately 30 percent of the population will be between 4,700 and 5,000 psi. Thus, 1.0000 - 0.30 = 0.70 or about 70 percent of the 50 trucks (i.e., about 35 trucks) should be rejected.

Try to match this solution by using Table 2.1. Hint: start by computing z_{4700} and z_{5000}, then use Table 2.1.

2.2.4 The "t-distribution"

The "t-distribution" is used to test sample means when the population variance (or population standard deviation) is not known (which is usually the case for most of the data you deal with). The "t-statistic" is quite similar to the "z-statistic" but also includes consideration of the sample size (n). Table 2.2 is used to provide a partial listing of t-values.

where

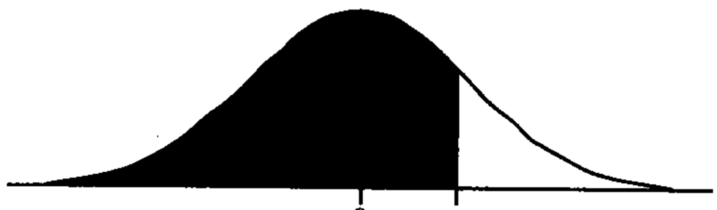
$$z = \frac{x - \mu}{\sigma} \qquad \text{Eq. 2.3}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \text{Eq. 2.4}$$

This concept will be illustrated in more detail in the next section.

Table 2.1 Normal Distribution Table [from Ref. 2.2]

Normal Distribution



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7793	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.3 Partial t-Distribution Table for Two-Tail Test (after Steel and Torrie [2.3])

Degrees of Freedom	Probability of a Larger t		
	0.10	0.05	0.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
⋮			
15	1.753	2.131	2.947
⋮			
20	1.725	2.086	2.845
⋮			
25	1.708	2.060	2.787
⋮			
30	1.697	2.042	2.750
⋮			
∞	1.645	1.960	2.576

2.3 CHI-SQUARE TEST TO CHECK NORMALITY

The χ^2 test of goodness of fit can be used to check whether a set of data is actually normally distributed. This may not seem all that important; however, the WSDOT asphalt concrete quality assurance (QA) specification, for example, uses such an assumption (normally distributed test data, that is). Such checks were made in the WSDOT report by Markey et al. [2.4] for aggregate gradations (1/2, 3/8, 1/4, #10, #40, #200), asphalt cement content and compaction on a selection of 1989 paving projects. The normality check of percent compaction will be illustrated following a bit of additional background on the χ^2 test.

2.3.1 Background Information

The χ^2 test can be used to check <u>any</u> data distribution but, of specific interest, is the normal distribution. To do this, a sample set of data frequencies are compared against a set based on some hypothesis (data frequencies are, in essence, histograms as illustrated in Figure 2.1).

The calculated results are compared to a theoretical value (from a $\chi 2$ "lookup" table).

The basic formula for the χ^2 test is (from Duncan [2.5]):

$$\sum_{i=1}^{k} \frac{(F_i - f_i)^2}{f_i} \qquad \text{Eq. 2.5}$$

where F_i = sample frequencies (or <u>actual</u> data frequencies) for k classes,

f_i = frequencies <u>expected</u> on the basis of the hypothesis that the actual data is normally distributed, and

k = number of frequencies (for example, k = 14 in Figure 2.1).

Thus, the hypothesis being tested is:

H_0 = sample data distribution = normal distribution
H_1 = sample data distribution ≠ normal distribution

2.3.2 Example

The example is based on data summarized by Markey et al. [2.4] for a selection of 1989 WSDOT paving projects. The data is combined from three "QA" paving projects, all built during the 1989 paving season. The variable being measured is "percent of Rice density" which is a measure of the degree of compaction of the in-place hot-mix asphalt concrete. All three projects had the same mix (WSDOT Class B) and specification requirements. The total data points (lots) is 201. The data is shown in Table 2.3, summarized in frequencies in Table 2.4, and the χ^2 statistic computed in Table 2.5

$\chi^2_{calculated} = 16.0$
$\chi^2_{table}\ (\upsilon = k - 3 = 10 - 3 = 7, \alpha = 0.05\%) = 14.1$
$\chi^2_{table}\ (\upsilon = 7, \alpha = 0.025\%) = 16.0$

Conclude that the Rice density data is normally distributed at an $\alpha = 0.025\%$. It fails the normal distribution at $\alpha = 0.05\%$.

Table 2.3 Rice Density Data

	Project				Project		Project
Lot	A	B	C	Lot	C	Lot	C
1	92.87	93.66	93.08	47	92.48	91	92.04
2	92.70	93.38	92.96	48	92.28	92	91.84
3	91.66	93.86	92.48	49	91.78	93	91.82
4	92.14	93.50	92.62	50	91.62	94	93.32
5	92.24	93.96	92.04	51	93.42	95	92.40
6	92.38	93.00	92.94	52	93.60	96	92.50
7	91.74	92.52	91.16	53	93.38	97	93.14
8	92.46	93.88	92.90	54	92.70	98	92.32
9	92.34	94.38	92.52	55	93.20	99	92.44
10	93.74	94.56	91.68	56	92.60	100	93.20
11	92.76	95.16	92.22	57	93.28	101	92.94
12	92.82	95.52	92.06	58	92.58	102	91.60
13	92.76	95.02	93.44	59	92.42	103	92.62
14	93.04	94.32	91.22	60	92.38	104	90.30
15	93.06	95.10	92.36	61	91.78	105	92.18
16	93.90	93.26	92.12	62	93.18	106	92.28
17	92.88	93.40	92.68	63	94.12	107	88.64
18	93.18	93.94	92.52	64	93.06	108	88.90
19	93.12	94.20	91.58	65	93.04	109	92.82
20	92.86	94.66	92.26	66	92.62	110	92.18
21	93.24	91.82	91.68	67	92.90	111	92.62
22	92.56	92.96	91.04	68	92.06	112	92.10
23	93.64	91.42	93.40	69	91.86	113	92.84
24	94.82	92.20	93.38	70	92.38	114	91.78
25		92.16	92.96	71	91.86	115	91.98
26		92.78	93.72	72	93.26	116	93.76
27		92.38	94.42	73	92.80	117	92.90
28		92.48	93.58	74	93.88	118	92.48
29		92.10	93.12	75	94.12	119	92.12
30		91.80	94.34	76	92.30	120	93.06
31		93.48	92.50	77	92.20	121	92.48
32		92.46	93.14	78	92.14	122	93.70
33		92.60	92.98	79	93.50	123	92.50
34		93.54	92.30	80	91.86	124	91.90
35		91.24	92.36	81	92.62	125	92.99
36		93.44	91.14	82	92.16	126	92.34
37		92.52	91.16	83	92.64	127	91.60
38		93.78	93.12	84	91.30	128	92.60
39		93.68	92.70	85	93.70	129	91.66
40		94.60	92.30	86	92.80	130	92.10
41		94.18	92.24	87	92.58	131	92.28
42		94.64	92.76	88	91.38		
43		92.80	92.84	89	92.26		
44		93.80	92.48	90	93.76		
45		92.24	93.00				
46		93.20	92.02				

Table 2.4 Rice Density Data — Frequencies, Basic Statistics

From the data contained in Table 2.3

Class Interval[1]	Number of Lots Within Interval	
88.5 - 89.0	2	
89.0 - 89.5	0	
89.5 - 90.0	0	
90.0 - 90.5	1	
90.5 - 91.0	0	$\overline{x} = 92.73\%$
91.0 - 91.5	9	s = 0.95%
91.5 - 92.0	21	n = 201
92.0 - 92.5	52	
92.5 - 93.0	47	
93.0 - 93.5	33	
93.5 - 94.0	20	
94.0 - 94.5	8	
94.5 - 95.0	4	
95.0 - 95.5	3	
95.5 - 96.0	1	

Note
1. Class Interval = 0.5% in Rice Density

Table 2.5. Calculation of χ^2 for Rice Density Data

(1) Upper Limit of Class Interval	(2) Limits in Standardized Units (z)	(3) Relative Frequency from -∞ to z	(4) Relative Frequency of Cell	(5) Absolute Theoretical Frequency (f)	(6) Actual Frequency (F)	(7) χ^2
91.0 or less	-1.82	0.0344	0.0344	6.9	3	2.2
91.5	-1.29	0.0985	0.0641	12.9	9	1.2
92.0	-0.77	0.2206	0.1221	24.5	21	0.5
92.5	-0.24	0.4052	0.1846	37.1	52	6.0
93.0	+0.28	0.6103	0.2051	41.2	47	0.8
93.5	+0.81	0.7910	0.1807	36.3	33	0.3
94.0	+1.34	0.9099	0.1189	23.9	20	0.6
94.5	+1.86	0.9686	0.0587	11.8	8	1.2
95.0	+2.39	0.9916	0.0230	4.6	4	0.1
∞	∞	1.0000	0.0084	1.7	4	3.1
			1.0000	200.9	201	16.0

Note

1. Column 1 from Table 2.4.

2. Column 2: $z = \frac{x - \overline{x}}{s}$ where $\overline{x} = 92.73\%$
   $s = 0.95\%$
   $n = 201$

3. Column 3: Use z (Col. 2) and Table 2.1 to obtain relative frequency.

4. Column 4: Relative frequency of cell obtained from Col. 3 (interval value - preceding interval value)

5. Column 5: f = (Relative Freq. Cell)(number of lots)
   = (Col. 4)(201)

6. Column 6: from lot data shown in Table 2.4.

7. Column 7: $\chi^2 = \frac{(F - f)^2}{f}$

SECTION 2.0
REFERENCES

2.1 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

2.2 Ulberg, Cy, "Highway Research Manual," Draft Final Report, Research Project Y3399, Task 29, Washington State Department of Transportation, Olympia, WA, August 1987.

2.3 Steel, Robert G.D., and Torrie, James H., Principles and Procedures of Statistics," McGraw-Hill, New York, 1960.

2.4 Markey, Stephen J., Mahoney, Joe P., Gietz, Robert M., "An Evaluation of the WSDOT Quality Assurance Specification for Asphalt Concrete," WSDOT Research Report WA-RD XXX.X, Washington State Department of Transportation, Olympia, Washington, October 1993.

2.5 Duncan, Acheson, J., Quality Control and Industrial Statistics, Fourth Edition, Richard D. Irwin, Inc., Homewood, Illinois, 1974.

SECTION 3.0

TESTS OF HYPOTHESES

3.1 INTRODUCTION

Hypothesis testing is a way in which statistical methods can be used to help in the decision making process. Such testing considers the mean and standard deviation of a group of data, the confidence level (a probability statement) and something about the population being sampled. Hypothesis testing is extremely helpful in performing multiple regression analysis and hence it is important for you to understand the basics.

3.2 HYPOTHESES

Webster's Seventh New Collegiate Dictionary defines hypothesis as "...a tentative assumption made in order to draw out and test its logical or empirical consequences....an assumption or concession made for the sake of argument..." You can begin to see the problem in explaining hypothesis testing.

There are always two hypotheses for any statistical test [3.1]. These hypotheses are

H_0 = null hypothesis (most important)

H_1 = alternative hypothesis

What is about to be presented is one of the fundamental problems in statistics which is the use of "double negative" statements. Any hypothesis must be tested statistically to be rejected or not rejected (this is a statistical way of accepting something).

The hypotheses (H_0 or H_1) can result in two types of errors if the wrong one is selected, as shown in Table 3.1. The probability of the Type I and II errors is very important, since it determines how carefully you must distinguish between true and false hypotheses.

(This is an area in which statistical "games" can be played, so you need to be very careful). These probabilities are [after Ref. 3.1]

$P(\text{Type I error}) = \alpha$

$P(\text{Type II error}) = \beta$

Table 3.1. Types of Hypothesis Errors

"The Truth"	"The Actual Decision"	
	Reject H_0	Accept H_0
H_0 true	Type I Error (α)	Correct!
H_0 false	Correct!	Type II Error (β)

The general form for calculating the z-statistic for hypothesis testing is

$$z_{calc} = \frac{\text{(sample mean) - (hypothesized value)}}{\text{standard error}}$$

where sample mean = $\bar{x}$

hypothesized value = μ (sometimes assumed to = 0 in regression hypothesis testing)

standard error = $\frac{\sigma}{\sqrt{n}}$ = standard deviation of means of random samples of size n from a "parent" population with standard deviation σ. Standard error is sometimes designated $\sigma_{\bar{x}}$.

sample size = n

The same general form applies to the t-statistic for hypothesis testing when none of the population statistics (μ, σ) are known:

$$t_{calc} = \frac{\text{(sample mean) - (hypothesized value)}}{\text{standard error}}$$

where sample mean = $\bar{x}$

hypothesized value = μ (again, sometimes an assumed or stated value)

standard error = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$

3.2.1 Example 1: PCC Mix

For this example, use the data shown in Figure 2.1. This contractor states that the batch plant has produced a mix in the past of

μ = 4,824 psi

σ = 387 psi

(Since these are population statistics, you can assume that these data were collected over a long period of time)

On the job you take six samples (cylinders) with the result that $\bar{x}$ = 4,549 psi.

Question: Is the contractor correct?

Solution: Assume that the data are normally distributed and use hypothesis testing.

H_0: μ = 4,824 psi
H_1: μ < 4,824 psi

$$z_{calc} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4549 - 4824}{387/\sqrt{6}} = -1.74$$

$z_{critical} = -1.65$ (for Type I error (or α) = 5%) (Refer to Table 2.1 to determine z.)

since $z_{calc} > z_{critical}$, reject H_0.

Thus for your job, you must judge the contractor's claim to be incorrect. Refer to Figure 3.1 for an illustration of this process.

3.2.2 Example 2: PCC Mix [after Ref 3.2]

The PCC mix contractor claims the following:

PCC mix ≥ 4,000 psi (28-day compressive strength)

You take a random sample of five cylinders and cure them for 28 days (n = 5).

The results:

$\bar{x}$ = 3,740 psi
s = 390 psi
n = 5

Question: If you are willing to accept a 5 percent chance of a Type I error (i.e., rejecting a true H_0), should you believe the contractor?

Solution

H_0: $\mu \geq 4{,}000$ psi (null hypothesis)

H_1: $\mu < 4{,}000$ psi (alternative hypothesis)

$$t_{calc} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3740 - 4000}{390/\sqrt{5}} = -1.48$$

$t_{critical}$ (@ 5%) < -2.13 (one-tail α = 5% with $\nu = n - 1 = 4$ degrees of freedom)

Therefore, you accept H_0, since

$t_{calc} = -1.48 < -2.13 = t_{critical}$

You have no "statistical" reason to doubt the contractor's claim. Refer to Figure 3.2(a) which further illustrates this example. However, note that if the Type I error (rejecting a true H_0) were reduced to a 1 percent chance, then

$t_{critical}$ (@1%) > -3.747 (one-tail $\propto$ =1% with 4 degrees of freedom)

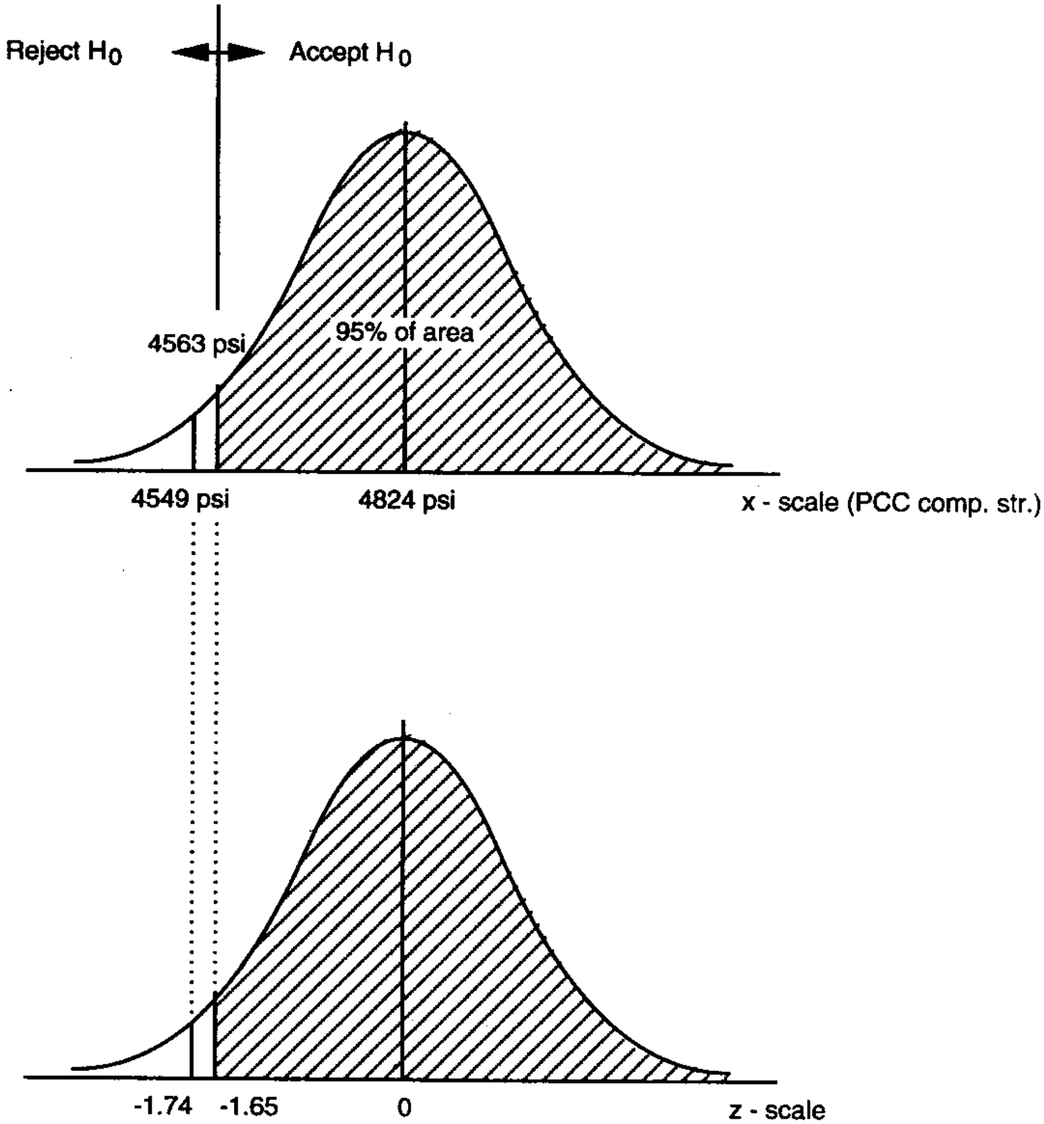


Figure 3.1. Hypothesis Testing with Population Mean and Standard Deviation Known

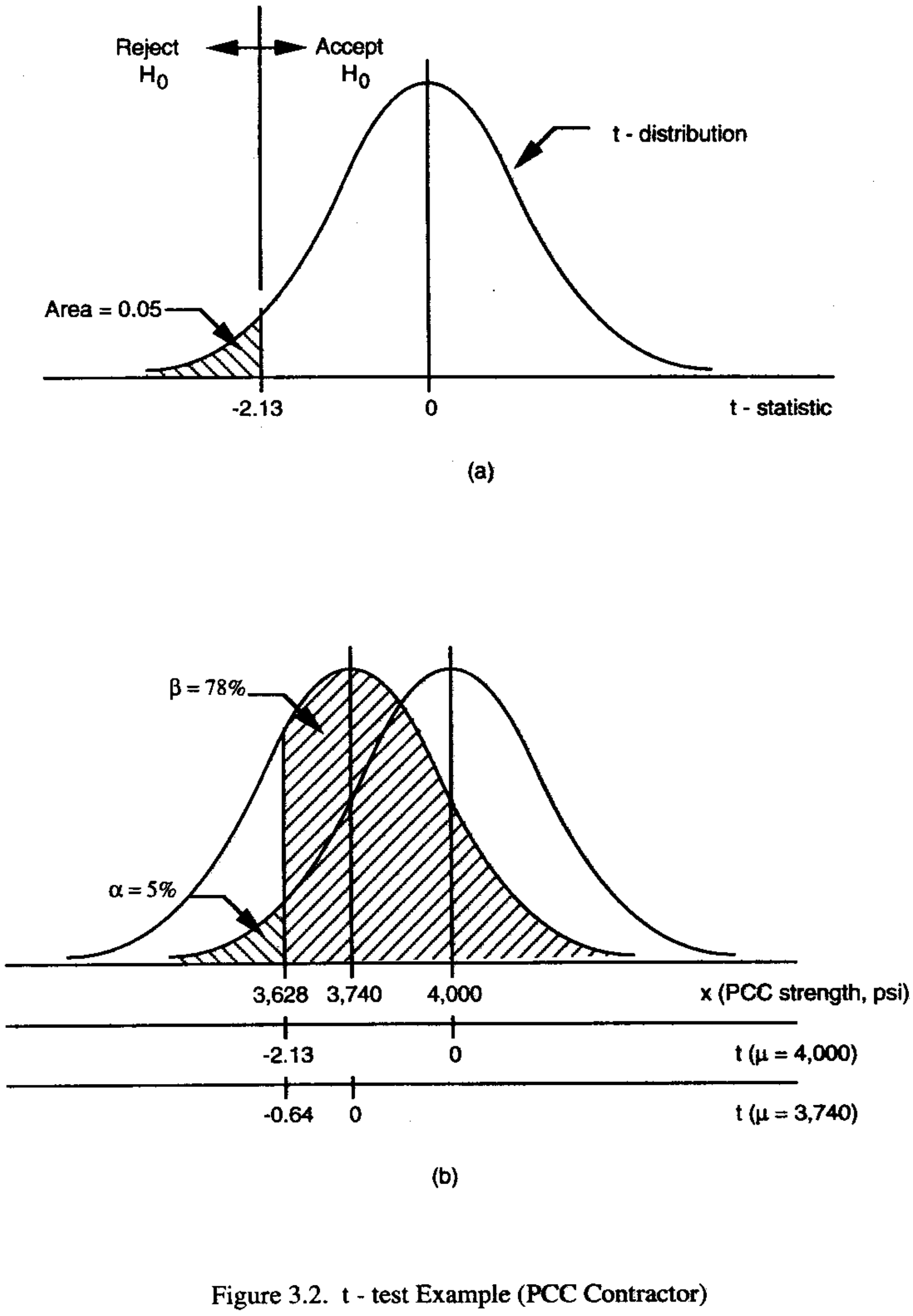


Figure 3.2. t - test Example (PCC Contractor)

Thus, you are even more unwilling to accept the alternate hypothesis (H_1) that the contractor's claim was incorrect. Note that the Type I error protects against rejecting a true null hypothesis. In other words, you can select a low Type I error level so that it is difficult to reject the null hypothesis. However, as the Type I error level decreases, the Type II error level increases (not rejecting a false H_0). It is not easy to illustrate the calculation of the Type II error (β), but this example is a good case since a Type I error level of about 11 percent would be needed to reject the null hypothesis.

Often the Type I error is termed the "seller's risk" and the Type II error the "buyer's risk." For the example, the lower the Type I error the lower the risk of the contractor. Correspondingly, the Type II error increases the risk of the DOT accepting PCC of lower than specified quality (again not rejecting a false H_0). Needless to say, a balance between Type I and Type II errors is needed (but not necessarily the same number or value because one error type may be more important than another) in developing statistically based materials "acceptance plans."

At least one illustration of ignoring the Type II error (β) is appropriate The PCC mix contractor example will be used.

Assume that the sample results,

$\bar{x}$ = 3,740 psi
s = 390 psi

actually represent the true population (i.e., μ = 3,740 psi instead of μ = 4,000 psi). First, calculate the value of $\bar{x}$, which corresponds to t = -2.13 (α = 5%). This value is the one that separated the rejection and acceptance region for H_0 ($\mu \geq$ 4,000 psi). (Refer to Figures 3.2(a) and (b).)

$$\frac{\bar{x} - 4000}{\frac{390}{\sqrt{5}}} = -2.13$$

$$\therefore \bar{x} = 3{,}628 \text{ psi}$$

The value of the Type II error (β) is the area under the curve (or distribution) with μ = 3,740 psi and σ = 390 psi but within the acceptance region of the original H_0 distribution (μ = 4,000 psi). This is illustrated in Figure 3.2(b).

$$t = \frac{3{,}628 - 3{,}740}{\frac{390}{\sqrt{5}}} = -0.64$$

(Use a t-table, available in most statistics books.)

$$\therefore \beta = P(t \geq -0.64) = 0.78 \text{ (or 78\%)}$$

Therefore, there is a 78 percent chance of accepting a false H_0 ($\mu = 4{,}000$) if the true population mean (μ) = 3,740 psi. You can see that the β will change as the correct population mean changes. Clearly, this level of β is quite high. Thus, the DOT's risk (the "buyer's risk") is too high. A balance in setting α, β and sample size levels is very important in the proper use of these kinds of statistical tests. This leads to another important area of statistics, generally called "acceptance testing" and "operating characteristic curves," which is not appropriate for discussion in these notes.

3.2.3 Example 3: WSDOT/Industry PCC Testing

WSDOT and industry representatives jointly tested fresh concrete delivered to the South Seattle Community College (the testing site) on December 5, 1989. One purpose of this activity was to see how test results compared for different testing teams. The tests performed by all of the teams included both fresh and hardened mix properties: slump, air content, unit weight and compressive strength. The ready-mix was specified to conform to a standard WSDOT mix ("AX" mix, the WSDOT Standard Specifications have since been changed). This mix was delivered to the testing site as a 5.2 sack mix with fly ash and 3/4 in. maximum coarse aggregate size; additives included an air entraining agent and a water reducer. The per cubic yard batch weights were:

- cement : 488 lb
- Fly ash : 152 lb
- Fine aggregate : 1,130 lb
- Coarse aggregate : 1,864 lb
- Water : 32 gal
- Water reducer : 24.4 oz
- Air entraining agent : 6.4 oz

The specified 28-day design compressive strength was 4,000 psi with a maximum slump of 3 in. (vibrated concrete) and an air content of 5 percent (± 1 1/2 percent).

Each team picked a number at random (1 through 20). Team 1 would then obtain their PCC from the truck, Team 2 next and so forth (i.e., there should be no bias as to when the test teams received their material for testing). The team test results are shown in Table 3.2 with basic summary statistics in Table 3.3. The hypothesis test used is a means test for two independent samples with the population standard deviation unknown and for small samples. A small sample implies the number of testing teams were fewer than 30. The hypothesis test formulas used are shown in Table 3.4 and the hypothesis results in Table 3.5.

The results shown in Table 3.5 also includes a comparison of Tests 1-8 and 9-20. This was done since the PCC mix, as discharged from the truck, apparently had somewhat different fresh mix properties as characterized by slump. The results shown in Table 3.5 indicate that there were no significant test differences between WSDOT and industry test teams for measurements of slump, air content and unit weight. There were significant differences for the compressive strength results. The strength tests were organized as follows:

Table 3.2. Test Summary — WSDOT/Industry Concrete Testing Program

Tester No.[a]	Affiliation	Results						
		Slump (in.)	Air Content (%)	Unit Weight (pcf)	Compressive Strength (psi)			
					Commercial Lab		WSDOT Lab	
					No. 1	No. 2	No. 1	No. 2
1	Industry	4.75	4.9	148.45	4280	4040	4680	5000
2	Industry	4.75	4.5	147.20	4360	4210	4700	4750
4	Industry	5.25	5.4	147.24	4180	4310	4650	5050
5	WSDOT	4.75	4.9	148.69	4340	4390	4770	5190
6	WSDOT	5.00	5.6	147.72	4530	4400	4700	4740
7	Industry	4.75	5.5	147.44	4620	4400	4700	4700
8	WSDOT	5.50	5.0	148.15	4740	4730	4770	4730
9	WSDOT	4.00	5.3	149.39	4590	4590	4980	5160
10	Industry	4.00	5.4	146.12	4280	4320	4520	4670
11	Industry	4.50	4.6	148.39	4320	4280	4050	3920
12	WSDOT	4.25	4.8	136.90	4570	4410	5140	5120
13	Industry	4.00	5.0	147.68	4390	4420	4610	5110
14	WSDOT	4.00	5.1	163.33	4890	4670	5390	5250
15	Industry	3.75	4.8	146.70	4350	4380	5110	5180
16	WSDOT	4.00	5.0	147.83	4500	4590	4820	5050
17	WSDOT	4.00	5.0	148.05	4700	4730	5280	5640
20	Industry	4.25	4.7	146.20	4710	4780	4680	5050

Note:
(a) No testers selected numbers 3, 18, and 19.

Table 3.3. Basic Statistics For PCC Test Results

Test	Data Set		Sample Size (n)	Mean ($\bar{x}$)	Standard Deviation (s)	Coefficient of Variation ($s/\bar{x}$)100
Slump (in.) (Basic Groupings)	WSDOT		8	4.44	0.58	13.0%
	Industry		9	4.44	0.48	10.8%
	Overall		17	4.44	0.51	11.5%
Slump (in.) (Sequential Groupings)	Tests 1-8		7	4.96	0.30	6.1%
	Tests 9-20		10	4.08	0.21	5.1%
Air (%) (Basic Groupings)	WSDOT		8	5.09	0.25	5.0%
	Industry		9	4.98	0.37	7.5%
	Overall		17	5.03	0.32	6.3%
Air (%) (Sequential Groupings)	Tests 1-8		7	5.11	0.40	7.8%
	Tests 9-20		10	4.97	0.25	5.1%
Unit Weight (pcf)	WSDOT		8	148.8	7.1	4.8%
	Industry		9	147.3	0.8	0.6%
	Overall		17	148.0	4.8	3.3%
Unit Weight (pcf) with "outliers" removed	WSDOT		6	148.3	0.6	0.4%
	Industry		9	147.3	0.8	0.6%
	Overall		15	147.7	0.9	0.6%
Compressive Strength (psi) (Basic Groupings)	WSDOT Cyl/WSDOT Lab		16	5046	275	5.5%
	WSDOT Cyl/Commercial Lab		16	4586	153	3.3%
	Industry Cyl/ WSDOT Lab		18	4729	340	7.2%
	Industry Cyl/Commercial Lab		18	4368	181	4.1%
	Overall/WSDOT Lab		34	4878	346	7.1%
	Overall/Commercial Lab		34	4471	199	4.5%
	Overall		68	4674	347	7.4%
Compressive Strength (psi) (Sequential Groupings)	Tests 1-8	WSDOT Lab	14	4795	163	3.4%
		Commercial Lab	14	4395	202	4.6%
	Tests 9-20	WSDOT Lab	20	4936	425	8.6%
		Commercial Lab	20	4524	183	4.1%

Table 3.4. Formulas Used For Basic Statistical Analysis [after References 3.1 and 3.2]

1. Statistical tests reported are "means test for two independent samples with population standard deviation unknown and small samples" (small sample implies that the WSDOT and Industry testers were less than 30, which was the case).

2. Null hypothesis is H_0: $\mu_W = \mu_I$

   Alternative hypothesis is H_1: $\mu_W \neq \mu_I$

   where: μ_W = population mean for a specific test for WSDOT testers
   μ_I = population mean for a specific test for Industry testers

3. t-statistic

$$t = \frac{\bar{x}_W - \bar{x}_I}{s_d}$$

$$v = n_W + n_I - 2$$

where: $\bar{x}_W$ = sample mean for a specific test for WSDOT testers

$\bar{x}_I$ = sample mean for a specific test for Industry testers

n_W = sample size (WSDOT)

n_I = sample size (Industry)

s_d = standard deviation of the difference of the sample means

$$= \left[\left(\frac{s_W^2(n_W - 1) + s_I^2(n_I - 1)}{(n_W - 1) + (n_I - 1)}\right)\left(\frac{n_W + n_I}{n_W\, n_I}\right)\right]^{1/2}$$

v = degrees of freedom = $(n_W - 1) + (n_I - 1) = n_W + n_I - 2$

Reference 3.1

Leland Blank, *Statistical Procedures for Engineering, Management, and Science*, McGraw-Hill, 1980, p. 381-383.

Reference 3.3

Robert Steel and James Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, 1960, p. 74.

Table 3.5. Results Of Hypothesis Testing For Various PCC Tests

Test	Comparison	t-statistic: Calculated	t-statistic: Critical Region [a] ($\alpha = 0.05$)	Conclusion[b]
Slump (Basic)	WSDOT = Industry	-0.023	-2.131	No significant difference
Slump (Sequential)	Tests 1-8 = Tests 9-20	+7.228	+2.131	Significant difference
Air Content (Basic)	WSDOT = Industry	+0.637	+2.131	No significant difference
Air Content (Sequential)	Tests 1-8 = Tests 9-20	+0.892	+2.131	No significant difference
Unit Weight	WSDOT = Industry	+0.632	+2.131	No significant difference
Unit Weight without Outliers	WSDOT = Industry	+2.597	+2.160	Significant difference
Compressive Strength (WSDOT Lab)	WSDOT = Industry	+2.964	+2.038	Significant difference
Compressive Strength (Commercial Lab)	WSDOT = Industry	+3.766	+2.038	Significant difference
Compressive Strength (WSDOT Lab)	Tests 1-8 = Tests 9-20	-1.178	-2.038	No significant difference
Compressive Strength (Commercial Lab)	Tests 1-8 = Tests 9-20	-1.939	-2.038	No significant difference
Compressive Strength (All Tests)	WSDOT Lab = Commercial Lab	+5.946	+2.000	Significant difference

Notes

(a) Critical region defined by the t-statistic for a two tail Type I error of 5% ($\alpha = 0.05$) for ν degrees of freedom
(b) Conclusion based on hypothesis test described in Table 3.3

- Cylinders prepared by all teams and tested at the WSDOT lab.
- Cylinders prepared by all teams and tested at a commercial lab.

The results in Table 3.5 show a significant difference between the two laboratories (the WSDOT results were higher).

A final set of hypothesis tests were performed to compare the test results against "fixed" values (or limits). This was done for slump, air content and compressive strength. The associated and necessary formulas are shown in Table 3.6 with the results in Table 3.7. The hypotheses are (illustrated in Figure 3.3):

- Slump

  H_0 : μ = 3 in. (specified maximum slump for "AX" vibrated PCC)
  H_1 : μ > 3 in. (i.e., critical condition is <u>more</u> slump not <u>less</u>)

- Air content

  H_0 : μ = 5% (specification target air content)
  H_1 : $\mu \neq$ 5%

- Compressive strength

  H_0 : μ = 4000 psi (specification minimum 28-day compressive strength)
  H_1 : μ < 4,000 psi

The results in Table 3.7 indicate that the air content is within the acceptance region which is to say that the null hypothesis (H_0) is accepted (i.e., there is a statistical basis for accepting the fact that the air content is, in essence, 5 percent). The compressive strength is also in the acceptance region (i.e., accept H_0). Naturally, this can be observed by inspection but was included to illustrate the calculation process (i.e., all means were greater than 4,000 psi). Finally, the slump results are in the critical region (accept H_1). This indicates that the slump results exceed the maximum value of 3 in. Again, by inspection of the data, this result is rather obvious.

What the above hypothesis tests <u>do not do</u> is indicate whether the difference between a 3 in. slump or say a 4.44 in. slump is structurally important.

3.2.4 Example 4: Friction Number Data — Paired t-test

The friction number data are for SR-82, an Interstate highway in District 5. The milepost locations indicate that the friction testing was performed on portland cement concrete surfaces only. The friction tests were essentially obtained every one-half mile in all available lanes. The friction number data are shown in Table 3.8. Lane 1 is the "outside" or "curb" lane in all cases.

Table 3.6. Formulas Used For Comparing Test Results To Fixed Values [after References 3.1 and 3.3]

1. Statistical tests reported are "means test for one sample with population standard deviation unknown and a small sample."

2. Null hypothesis is H_0: μ = fixed value

   Alternative hypothesis is H_1: $\mu_w \neq$ fixed value (or < or > fixed value)

   where: μ = population mean for a specific test and tester group

   fixed value = 3 in. for maximum specified slump (H_1 > 3 in.)

   = 5% for air content (assumed based on WSDOT Spec. 6-02.3(2)A for cast-in-place concrete above the finished ground line) ($H_1 \neq 5\%$)

   = 4,000 psi for 28-day compressive strength (H_1 < 4,000 psi)

3. t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$v = n - 1$$

   where: $\bar{x}$ = sample mean for a specific test and tester groups

   μ_0 = stated mean population value in H_0

   s = standard deviation of the sample

   n = sample size

   $s/\sqrt{n}$ = standard error (sometimes designated $s_{\bar{x}}$)

Reference 3.1

Leland Blank, Statistical Procedures for Engineering, Management, and Science, McGraw-Hill, 1980, p. 377.

Reference 3.3

Robert Steel and James Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, 1960, p. 19.

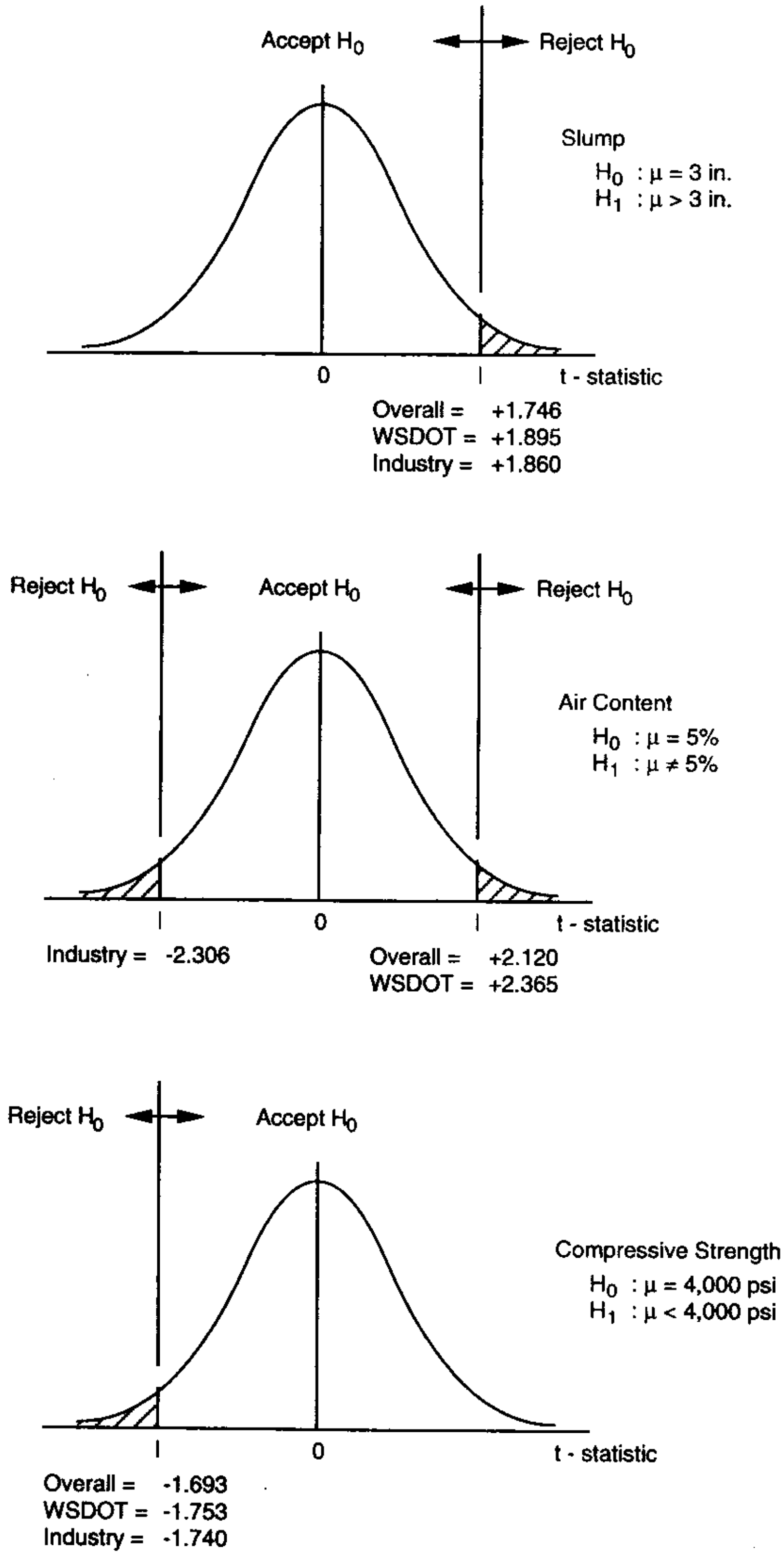


Figure 3.3. t - test Acceptance and Critical Regions for PCC Test Data Compared to Specification Values

Table 3.7. Results Of Hypothesis Testing for Fixed "Population" Values

Test	Comparison	t-statistic Calculated	t-statistic Critical Range[a] ($\alpha = 0.05$)	Conclusion[b]
Air Content (Basic)	Overall = 5%	+0.389	+2.120	No significant difference
	WSDOT = 5%	+1.018	+2.365	No significant difference
	Industry = 5%	-0.161	-2.306	No significant difference

Notes

(a) Critical region defined by the t-statistic for a two tail Type I error of 5% ($\alpha = 0.05$) for $\nu = n-1$ degrees of freedom

(b) Conclusion based on hypothesis test described in Table 3.6

Test	Comparison	t-statistic Calculated	t-statistic Critical Range[c] ($\alpha = 0.05$)	Conclusion
Slump (Basic)	Overall = 3 in.	+11.604	> + 1.746	Significant difference
	WSDOT = 3 in.	+7.025	> + 1.895	Significant difference
	Industry = 3 in.	+9.006	> + 1.860	Significant difference
Compressive Strength[d]	Overall = 4000 psi	+14.796	< -1.693	No significant difference
	WSDOT = 4000 psi	+15.215	< -1.753	No significant difference
	Industry = 4000 psi	+9.097	< -1.740	No significant difference

Notes

(c) Critical region defined by the t-statistic for a one tail Type I error of 5% ($\alpha = 0.05$) for $\nu = n-1$. A one-tail $\alpha = 0.05$ is equal to a two-tail $\alpha = 0.10$. Thus, can use Table 2.2

(d) Based on WSDOT lab results

Table 3.8. Friction Number Data, SR 82 (MP 4.0 - 14.0)

Mileposts	Friction Numbers			
	Eastbound		Westbound	
	Lane 1	Lane 2	Lane 1	Lane 2
4.0	40.4	50.6	38.5	46.4
4.5	36.2	50.5	41.0	47.2
5.0	38.3	44.3	40.6	47.7
5.5	40.1	46.0	38.7	46.8
6.0	41.3	47.4	39.6	44.4
6.5	42.6	45.9	40.4	47.8
7.0	35.4	43.9	39.5	47.7
7.5	37.9	48.1	41.1	43.3
8.0	41.0	48.6	43.3	55.3
8.5	41.6	45.5	41.7	49.2
9.0	41.0	44.3	38.8	51.7
9.5	40.0	48.5	40.8	46.1
10.0	39.9	50.3	38.6	46.5
10.5	42.9	47.3	43.6	49.4
11.0	41.8	46.2	41.0	50.7
11.5	43.8	46.5	41.7	49.5
12.0	40.9	55.9	42.1	47.0
12.5	41.5	52.9	41.5	53.4
13.0	43.4	50.3	40.8	52.5
13.5	46.3	52.6	40.9	46.4
14.0	40.4	54.0	41.4	46.5

Note: Lane 1: "Outside" or "curb" lane
Lane 2: "Inside" or "median" lane

Table 3.9 presents a few basic statistical measures of the friction number data. This includes mean, standard deviation, the number of data points and the coefficient of variation. The coefficient of variation is a dimensionless number that is the standard deviation divided by the mean multiplied by 100 (to convert to a percent). Stated another way, this value is used to express the standard deviation as a percentage of the mean.

A review of the information in Table 3.9 shows that, in general, the mean (or average) friction number is highest for the "inside" or "median" lanes. This is not surprising since the traffic in the inside lanes is generally lower than that in the "outside" lanes. These differences are about 7 friction numbers. Further, the coefficient of variation is generally higher for the inside lanes. This might suggest that studded tire wear (higher wear exposure in the outside lanes) results in more uniform but lower friction numbers.

For this analysis, the proposed hypothesis (or "null" hypothesis) is that there are no statistically significant differences among the differences between the mean friction numbers of the lanes. The mean friction number differences were based on "paired" observations, i.e., at a specific milepost the difference between any two friction numbers from two separate traffic lanes was calculated. The probability level chosen for this comparison was 95 percent. This implies that there is only a 5 percent chance that the true null hypothesis will be rejected. If the null hypothesis was accepted, then the last column of Table 3.10 would show "no significant difference." This would suggest that there would be only a 5 percent chance of the conclusion being wrong, and one could conclude (for the specific test data used) that there was no real difference in friction numbers for the two lanes compared. On the other hand, if the last column in Table 3.10 showed "significant difference," then the null hypothesis would be rejected. This would result in the conclusion that there was a real difference in friction numbers for the specific lanes and route being compared. This is further illustrated in Figure 3.4.

A review of the conclusions drawn in Table 3.10 (and the formulas shown in Table 3.11 and the calculations in Table 3.12) suggest that the inside lane has "significantly" higher friction numbers than the other lanes. This indicates that the lower exposure to vehicle traffic (and studded tires) results in higher pavement friction resistance (not a surprising finding). At the time the Friction Numbers were taken, this section of SR-82 was 15 years old.

3.2.5 Example 5: Pavement Structural Condition (visual pavement distress) — Paired t-test

A reasonable statistical test of PSC visual condition "paired observations" can be made (for example, if rating results from the 20 or so sample units are collected in a district and the "matching" information from the annual survey). These paired observations along with hypothesis testing using the t-statistic can be used. The mean differences of PSC (as one measure) will be tested. The proposed hypothesis (null hypothesis) is that there are no statistically significant differences for identical sections. The specific

Table 3.9. Basic Statistics for Friction Number Data, SR 82

Route	Mileposts	Parameter	Lane Direction and Number			
			WB		EB	
			1	2	1	2
SR-82	4.0 - 14.0	Mean ($\bar{x}$) Std. Dev. (s) Data Points (n) Coeff. of Var.	40.7 1.4 21 3.4%	48.4 3.0 21 6.2%	40.8 2.5 21 6.1%	48.6 3.4 21 7.0%

Table 3.10. Results of Hypothesis Testing for Friction Number Data, SR 82

Route	Mileposts	Lane Comparison (Dir: Lane Nos.)	t - statistic		Conclusion (c) (Hypothesis)
			Calculated (a)	Critical Region (b) ($\alpha = 0.05$)	
SR-82	4.0 - 14.0	WB: 1-2 EB: 1-2	-12.587 -9.540	-2.086 -2.086	Significant Difference Significant Difference
	Lane 1 is "Outside" lane of two lanes in one direction. Lane 2 is "Inside" lane of two lanes in one direction.				

Notes:

(a) t-statistic calculated from paired friction number observations for lane directions and numbers shown.

(b) Critical region defined by the t-statistic for a Type I error of 5% ($\alpha = 0.05$) for n-1 degrees of freedom
(number of data points minus one).

(c) Conclusion based on the initial hypothesis that there are no significant differences in friction numbers for the cases shown.

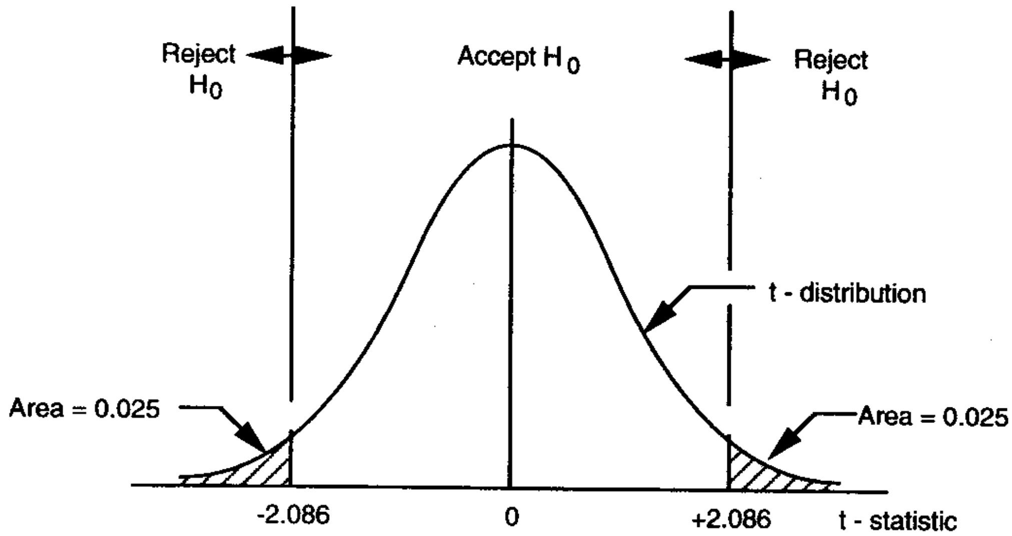


Figure 3.4. t - test Example (SR-82 Friction Numbers)

equations needed for paired t-tests were previously shown in Table 3.11. An example using paired observations is shown in Table 3.13.

Additional analysis must be done in order to check the β (Type II error). Recall that the basic errors in hypothesis testing for this condition are:

- Type I error (α): risk of rejecting a true null hypothesis (H_0) and thus concluding the annual and sample surveys are different when in fact they are not.

- Type II error (β): risk of accepting a false null hypothesis (H_0) and thus concluding the annual and sample surveys are the same when in fact they are not.

To check the potential for a Type II error, an operating characteristic curve was used for a two-sided t-statistic and α = 0.05 (from p. 368, Blank [3.1]). The Type II error decreases as the difference between the means of the annual survey and sample survey increases. Below are listed the sample size, β and 1 - β (defined as the "power" of a test which is the probability of rejecting the null hypothesis (no difference between means) when it is false). The ΔPSC is the difference between the mean of the annual survey and the sample survey.

	β			1 - β		
ΔPSC	n = 20	n = 25	n = 30	n = 20	n = 25	n = 30
2.5	0.82	0.79	0.75	0.18	0.21	0.25
5.0	0.52	0.42	0.35	0.48	0.58	0.65
7.5	0.17	0.10	0.05	0.83	0.90	0.95
10.0	0.02	0.01	~0	0.98	0.99	~1.00

The larger the power the better. Thus, sample sizes of 20 have reasonable power in detecting mean differences in PSC of about 7.5 points or more. A sample size of 30 is almost adequate to detect differences in PSC of 5.0 points. It would take a sample size of greater than 100 to detect mean differences as small as 2.5 points (β = 0.25, 1 - β = 0.75, n = 100) and a sample size of 75 to detect mean differences of 5.0 points (β = 0.05, 1 - β = 0.95, n = 75).

The bottom line is that the paired t-test with the range of sample sizes of 20 or more within a district should provide reasonable probabilities against rejecting a true hypothesis (no mean differences between the two surveys) and accepting a false hypothesis (the two surveys are different) for differences in PSC of say 7.5 or more points.

3.2.6 Nonparametric Statistic

The paired t-test discussion in the previous paragraph assumed the samples are to be taken from a normal distribution of PSC differences. If this is not the case, nonparametric ("distribution free") statistics can be used. Nonparametric tests assume that the sampled data are simply from a continuous variable.

Table 3.11. Formulas Used For Comparing Test Results To Fixed Values [after References 3.1 and 3.3]

1. Statistical tests reported are "means test for one sample with population standard deviation unknown and a small sample."

2. Null hypothesis is H_0: μ = fixed value

   Alternative hypothesis is H_1: $\mu_w \neq$ fixed value (or < or > fixed value)

   where: μ = population mean for a specific test and tester group

   fixed value = 3 in. for maximum specified slump ($H_1 > 3$ in.)
   = 5% for air content (assumed based on WSDOT Spec. 6-02.3(2) A for cast-in-place concrete above the finished ground line) ($H_1 \neq 5\%$)
   = 4,000 psi for 28-day compressive strength ($H_1 < 4{,}000$ psi)

3. t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$\nu = n - 1$$

   where: $\bar{x}$ = sample mean for a specific test and tester groups

   μ_0 = stated mean population value in H_0

   s = standard deviation of the sample

   n = sample size

   $s/\sqrt{n}$ = standard error (sometimes designated $s_{\bar{x}}$)

Reference 3.1

Leland Blank, *Statistical Procedures for Engineering, Management, and Science*, McGraw-Hill, 1980, p. 377.

Reference 3.3

Robert Steel and James Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, 1960, p. 19.

Table 3.12. Friction Number Example – Calculation of t-statistics

1. Summation of differences between Lane 1 and Lane 2:

$\sum d_i = -162.9$ (eastbound) $\sum d_i = -159.9$ (westbound)

To illustrate: $\sum d_i = (40.4 - 50.6) + (36.2 - 50.5) + \ldots + (40.4 - 54.0) = -162.9$ (for eastbound)

2. Summation of squared differences between Lane 1 and Lane 2:

$\sum d_i^2 = 1541.3$ (eastbound) $\sum d_i^2 = 1371.2$ (westbound)

To illustrate: $\sum d_i^2 = (40.4 - 50.6)^2 + (36.2 - 50.5)^2 + \ldots + (40.4 - 54.0)^2 = 1541.3$ (for eastbound)

3. Calculate mean of friction number differences:

$$\bar{d} = \frac{\sum d_i}{n} = -\frac{162.9}{21} = -7.757 \quad \text{(eastbound)}$$

$$\bar{d} = \frac{\sum d_i}{n} = -\frac{159.9}{21} = -7.614 \quad \text{(westbound)}$$

4. Calculate standard deviation of the differences:

$$s_d = \left[\frac{\sum d_i^2}{n-1} - \frac{n}{n-1}\bar{d}^2\right]^{1/2}$$

$$s_d = \left[\frac{1541.3}{20} - \left(\frac{21}{20}\right)(-7.757)^2\right]^{1/2} = 3.726 \text{ (eastbound)}$$

$$s_d = \left[\frac{1371.2}{20} - \left(\frac{21}{20}\right)(-7.614)^2\right]^{1/2} = 2.772 \text{ (westbound)}$$

5. Calculate t-statistic:

$$t_{calc} = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$t_{calc} = \frac{(-7.757) - (0)}{\frac{3.726}{\sqrt{21}}} = -9.540 \quad \text{(eastbound)}$$

$$t_{calc} = \frac{(-7.614) - (0)}{\frac{2.772}{\sqrt{21}}} = -12.587 \quad \text{(westbound)}$$

6. Determine t value

$t_{critical} = \pm 2.086$ (for two tail Type I error = 5% with $v = n - 1 = 20$ degrees of freedom)

(the $t_{critical}$ concept is illustrated in Figure 3.4)

Table 3.13. Illustrative Example of Comparison of Two Different Survey Methods

Assume the measured variable is PSC measured by the annual survey and PSC by the district sample

Segment No.	Survey Method: Annual	Survey Method: Sample	d_i	d_i^2
1	100	90	10	100
2	100	85	15	225
3	60	40	20	400
4	50	40	10	100
5	70	50	20	400
6	35	30	5	25
7	80	65	15	225
8	40	20	20	400
9	50	30	20	400
10	60	50	10	100
11	80	80	0	0
12	50	55	-5	25
13	60	70	-10	100
14	70	75	-5	25
15	80	65	15	225
16	90	95	-5	25
17	95	95	0	0
18	85	80	5	25
19	65	40	25	625
20	80	50	30	900

$\Sigma d_i = 195,\ \bar{d} = 9.75,\ \Sigma d^2 = 4325$

$$s_d = \left(\frac{\Sigma d_i^2}{n-1} - \frac{n}{n-1}\bar{d}^2\right)^{1/2}$$

$$= \left(\frac{4325}{19} - \frac{20}{19}(9.75)^2\right)^{1/2} = 11.3$$

$$t_{calculated} = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{9.75}{\frac{11.3}{\sqrt{20}}} = +3.859$$

$t_{critical^*} = +2.093$

$^*\alpha = 5\%,\ n = 20$

Thus, the conclusion is that there is a significant difference between the annual and sample surveys run on the same 20 segments (since $t_{calculated} > t_{critical}$).

A "two-sample signed rank test for means" will be used (referred to as the Wilcoxon test—refer to Blank [3.1], p. 456). The Wilcoxon is the nonparametric equivalent of the paired t-test. Tables 3.14, 3.15, and 3.16 are used to describe the Wilcoxon procedure and show an example problem (example uses same data as used in Table 3.13).

Table 3.14. Wilcoxon Paired Section Analysis
(after Blank [3.1])

1. Null hypothesis is H_o: $\delta = \delta_o$

   H_1: $\delta \neq \delta_o$

   Where:

   δ = mean difference between paired measurements

   $\delta_o = 0$

2. T statistic

   T = smaller sum of absolute value of signed ranks of differences
   = $\min[D_1, D_2]$

   If $T \leq T_0$, reject H_0

   If $T > T_0$, accept H_0
   where T_0 = table value of T

3. Procedure

   (a) Subtract paired samples to obtain differences

   $$d_i = X_{i1} - X_{i2}$$

   and discard all $d_i = 0$ values and reduce n accordingly.

   (b) Arrange absolute values, $| d_i |$, in increasing order.

   (c) Assign ranks (1, 2, ..., n) to the ordered differences. "Ties" are given the average of the assigned ranks.

   (d) Compute the sums

   $$D_1 = \Sigma d_i \text{ for all } d_i > 0$$

   $$D_2 = \Sigma d_i \text{ for all } d_i < 0$$

   (e) Compute T test statistic

   $$T = \min [D_1, D_2]$$

   (f) Refer to Table 3.9 for table T values

Table 3.15. Typical Two-Tail T Distribution Values (from Blank [3.1])

n	α level = 0.05
6	0
7	2
8	4
9	6
10	8
11	11
12	14
13	17
14	21
15	25
16	30
17	35
18	40
19	46
20	52
21	59
22	66
23	73
24	81
25	89

Note: T values shown can be used for sample sizes up to 25. For larger sample sizes, use a standard normal distribution and

$$z = \frac{T - \mu_T}{\sigma_T}$$

$$\mu_T = \frac{n\,(n + 1)}{4}$$

$$\sigma_T = \left[\frac{n\,(n + 1)(2n + 1)}{24}\right]^{1/2}$$

Table 3.16. Illustrative Example of Wilcoxon Hypothesis Test

1. Use the example data in Table 3.7 (recall eliminate all $d_i = 0$)

d_i	$\lvert d_i \rvert$ ordered	Ranks Plus	Ranks Minus	
10	5	3		recall, all repeated ranks given the avg rank, i.e., ranks 1, 2, 3, 4, 5 equal 15 and avg = 3
15	5		3	
20	5		3	
10	5		3	
20	5	3		
5	10	7.5		
15	10	7.5		
20	10	7.5		
20	10		7.5	
10	15	11		
0	15	11		
-5	15	11		
-10	20	14.5		
-5	20	14.5		
15	20	14.5		
-5	20	14.5		
0	25	17		
5	30	18		
25				
30		154.5	16.5	

2. $n = 18$, $T_0 = 40$ for $\alpha = 0.05$ (Table 6)

3. $T = \min [150.5, 16.5]$

4. Since $(T = 16.5) < (T_0 = 40)$, conclude that H_0 is false, i.e., there is a significant difference between sample means.

SECTION 3.0
REFERENCES

3.1 Blank, Leland, Statistical Procedures for Engineering, Management, and Science, McGraw - Hill Book Company, 1980.

3.2 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

3.3 Steel, Robert and Torrie, James, Principles and Procedures of Statistics, McGraw-Hill, 1960.

SECTION 4.0

REGRESSION ANALYSIS

4.1 INTRODUCTION

Recall from SECTION 1.0 that regression analysis can be used to generate an equation to predict one variable from another (or others, which constitutes multiple regression). The predicted variable is the "dependent variable" and the other variables are called "independent variables."

Sir Francis Galton (England) apparently first used the term "regression" in the context of statistics in the late 1800s [4.1]. He was studying the inheritance of human characteristics and noted that offspring tend to "revert" (regress) toward "mediocrity." What he was trying to say was that children's heights, as they grow into adults, tend toward an average or median height.

4.2 CORRELATION

In statistics, there are several ways two variables can be evaluated so that their collective association can be measured. A common measure of association is correlation. A few of the more significant points about correlation include [after Ref. 4.2] the following.

(a) The correlation coefficient is designated by "r."

(b) The correlation coefficient can range between -1 and +1. If the two variables whose association is being measured are designated as "y" and "x", then the correlation coefficient is positive if an increase in y corresponds to an increase in x.

(c) The correlation coefficient equals 1.0 if all of the y and x values fall on a straight line.

(d) When the correlation coefficient approaches 0.0, then there is little (if any) association between y and x (however, there are exceptions - refer to Figure 10.2 (p. 221) of the MINITAB manual [4.2]).

The basic equation for determining the correlation coefficient is

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\left[\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2\right]^{1/2}} \qquad \text{Eq 4.1}$$

The "computation" formula for r is

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\left[\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)\right]^{1/2}} \qquad \text{Eq.4.2}$$

4.2.1 Example — calculation of the correlation coefficient

An example of correlation will be a comparison of asphalt concrete resilient modulus values obtained in the same laboratory by two different technicians [4.3]. The laboratory data are for a gravel aggregate specimens tested at 77°F. They are shown on Table 4.1.

To simplify the calculations, units of $X10^6$ psi will be used, along with Equation 4.2.

y	x	y^2	x^2	xy
0.204	0.195	0.0416	0.0380	0.0398
0.231	0.207	0.0543	0.0428	0.0478
0.227	0.198	0.0515	0.0392	0.0449
0.228	0.204	0.0520	0.0416	0.0465
0.261	0.229	0.0681	0.0524	0.0598
0.195	0.180	0.0380	0.0324	0.0351
0.225	0.206	0.0506	0.0424	0.0464
0.216	0.202	0.0467	0.0408	0.0436
0.205	0.182	0.0420	0.0331	0.0373
0.232	0.235	0.0538	0.0552	0.0545
0.205	0.186	0.0420	0.0346	0.0381
0.261	0.237	0.0681	0.0562	0.0619
$\Sigma y = 2.690$	$\Sigma x = 2.461$	$\Sigma y^2 = 0.6078$	$\Sigma x^2 = 0.5087$	$\Sigma xy = 0.5557$

After Equation 4.2:

$$r = \frac{0.5557 - \frac{(2.461)(2.690)}{12}}{\left[\left(0.5087 - \frac{(2.461)^2}{12}\right)\left(0.6078 - \frac{(2.690)^2}{12}\right)\right]^{1/2}}$$

$$r = \frac{0.00403}{0.00437} = +0.92$$

4.3 LINEAR REGRESSION WITH ONE INDEPENDENT VARIABLE

4.3.1 Basic Regression Model

First, examine a basic regression <u>model</u> (or <u>equation</u>, which in this case indicates the same thing):

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{Eq. 4.3}$$

where y_i = value of the dependent variable for the i^{th} data point,

x_i = value of the independent variable for the i^{th} data point,

b_0, b_1 = constants (regression parameters),

Table 4.1. Asphalt Concrete Resilient Modulus Data by Two Technicians - Gravel Aggregate at 77°F

Specimen No.	Resilient Modulus (psi) at 77°F	
	Operator y	Operator x
1	204,000	195,000
2	231,000	207,000
3	227,000	198,000
4	228,000	204,000
5	261,000	229,000
6	195,000	180,000
7	225,000	206,000
8	216,000	202,000
9	205,000	182,000
10	232,000	235,000
11	205,000	186,000
12	261,000	237,000

ϵ_i = random error term, and

$i = 1, 2, 3, ..., n$.

The above model is a simple, linear model. It is simple since there is only one independent variable (x). It is linear since both the parameters (b_0, b_1) and the independent variable (x) are not power functions. (A non-linear model is one where the regression parameters ("constants") appear as exponents or when multiplied or divided by other parameters. Further, other types of non-linear models are ones where the independent variable(s) are second order powers (or higher). Non-linear models will be illustrated in Sections 4.3.7, 4.3.8, and 4.3.9.)

The regression parameters (b_0, b_1) are usually called regression coefficients. The coefficient b_1 is the slope of the regression line and the coefficient b_0 is the intercept of the regression line. This is illustrated in Figure 4.1, which is a plot of Friction Number at 40 mph versus ADT per lane for pavement field data for a select type of limestone rock surface course. The resulting equation, based on eight data points, is

$$FN_{40} = 56.9 - 0.00666 \text{ (ADT LANE)}$$

The intercept (b_0) at zero ADT per lane is 56.9. This is analogous to a new pavement surfacing that has received no traffic. The slope (b_1) is 0.00666, which means that the Friction Number is reduced by 0.00666 for each increase of 1 ADT per lane (or more understandably, the Friction Number is reduced by about 6.7 for each increase of 1,000 ADT per lane).

4.3.2 Method of Least Squares

The best relationship (or line) to use to predict some y from x is one that minimizes the differences between the regression line and the actual data. In Figure 4.2 (a), a clear association is shown and one not so clear is shown in Figure 4.2 (b). Thus, Figure 4.2 (a) probably comes closest to minimizing the differences between the line and the plotted y and x data points.

The minimization of the differences between the regression line and the actual data points is illustrated in Figure 4.3, i.e., the differences between the fitted data on the regression line ($\hat{y}_i$) and the actual data points (y_i) are minimized . More specifically, the squared differences are minimized, i.e., a minimum of $\Sigma(\hat{y}_i - y_i)^2$. The squared term results from the derivation, which is based on calculus. From this basic idea come the following equations, which are used to obtain the regression coefficients for a simple, linear regression model:

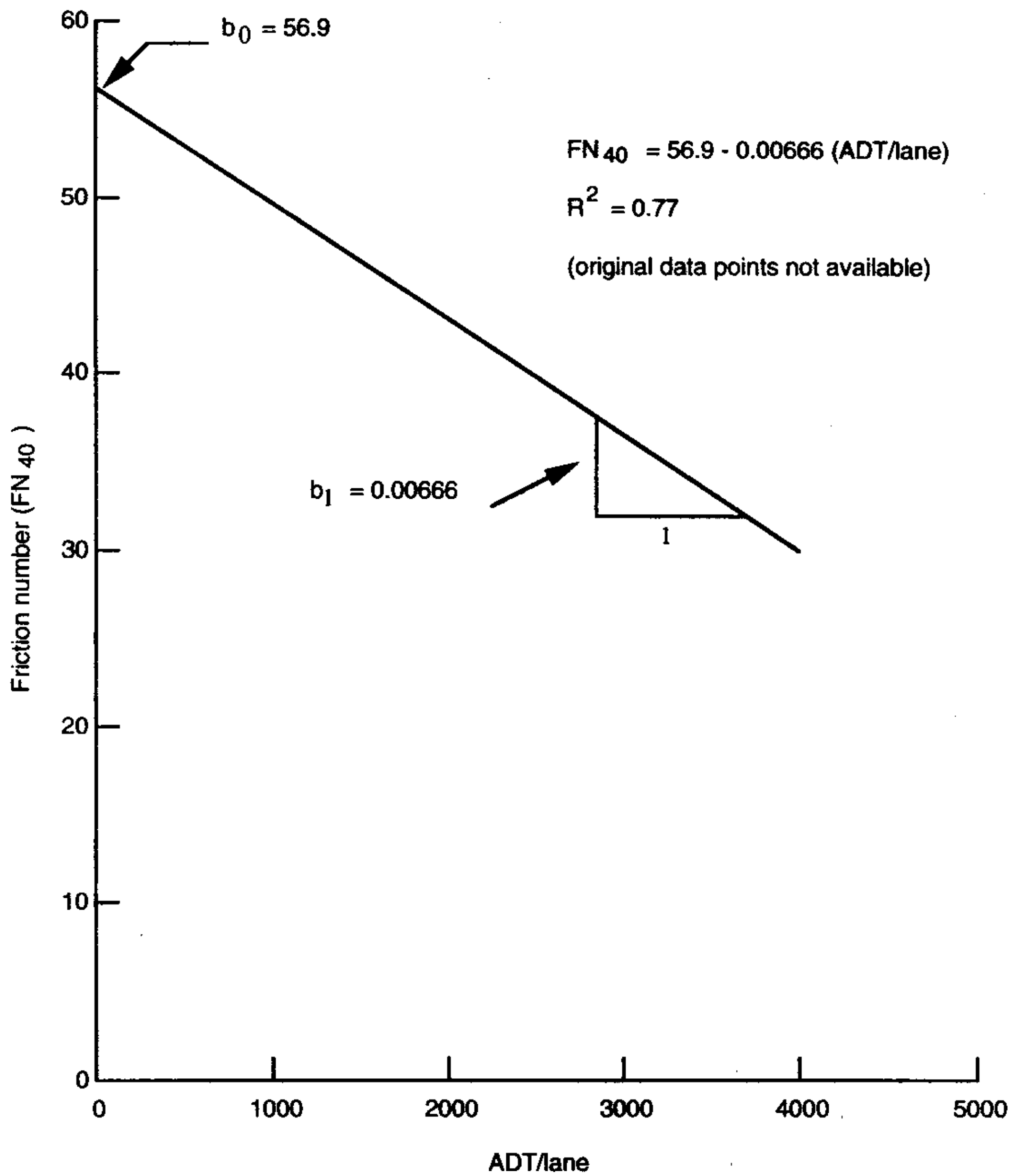
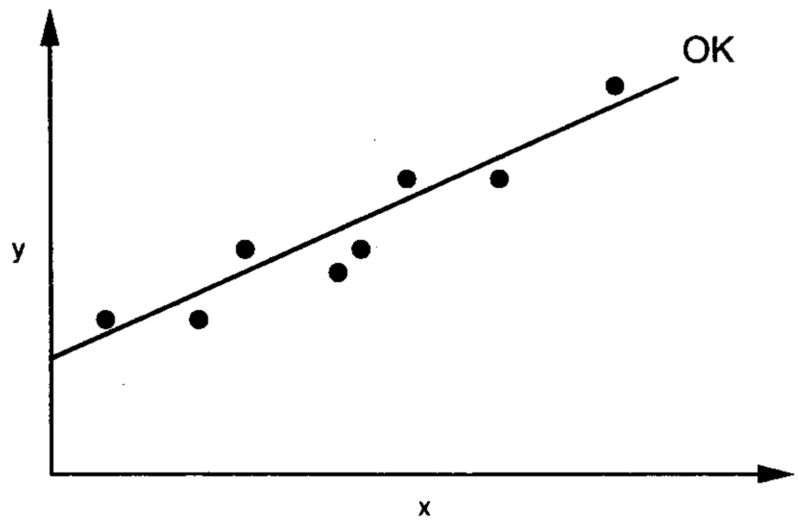
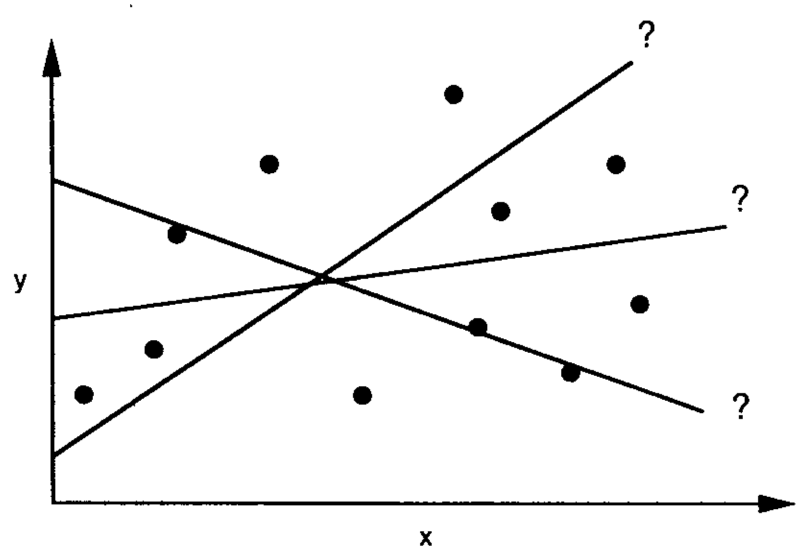


Figure 4.1. Friction Numbers versus ADT per Lane for Select Limestone Rock Surfaces



(a) A clear association between the line and the data points



(b) Unclear associations between lines and the data points

Figure 4.2. Data Fits

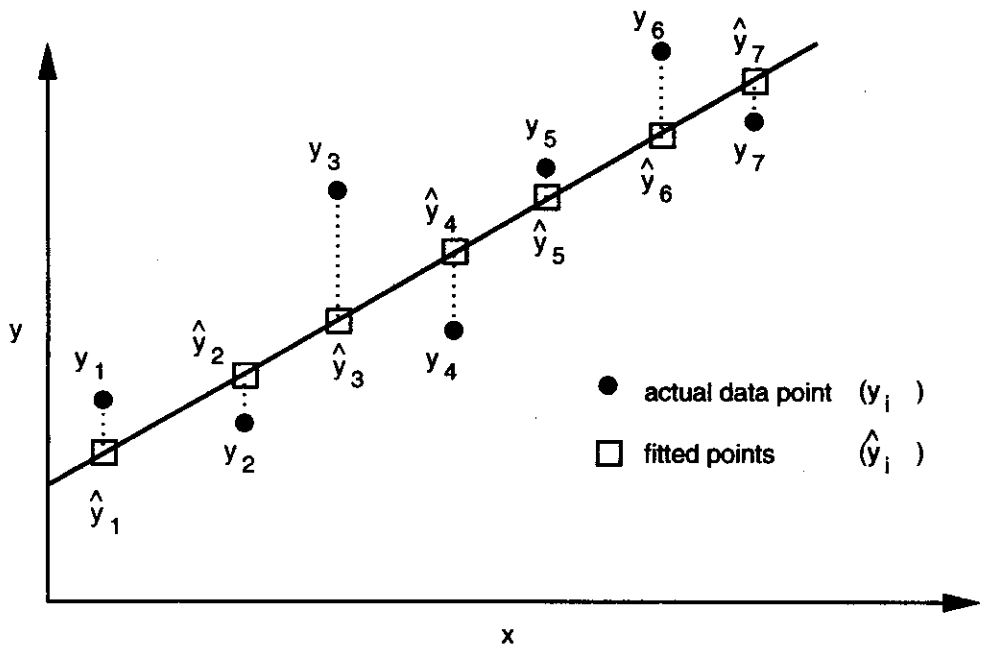


Figure 4.3. Illustration of Minimization of Differences Between the Regression Line and the Data Points

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} \qquad \text{(basic)} \qquad \text{Eq. 4.4}$$

$$b_1 = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} \qquad \text{(computational)} \qquad \text{Eq. 4.5}$$

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \text{(basic)} \qquad \text{Eq. 4.6}$$

$$b_0 = \frac{1}{n} (\Sigma y_i - b_1 \Sigma x_i) \qquad \text{(computational)} \qquad \text{Eq. 4.7}$$

4.3.3 Example — determination of a regression line by the method of least squares

Use the data from Table 4.1 to apply an "adjustment" to the resilient modulus data obtained by Operator "x". In other words, predict the modulus values that would be obtained by Operator "y" from what you know about the results obtained by Operator "x." Determine the appropriate regression line (the basic information needed to determine b_0 and b_1 is contained in SECTION 4.3.2).

$$b_1 = \text{(from Eq. 4.5)} = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$= \frac{0.5557 - \frac{(2.461)(2.690)}{12}}{0.5087 - \frac{(2.461)^2}{12}} = \frac{0.00403}{0.00399}$$

$$= 1.0100$$

$$b_0 = \text{(from Eq. 4.7)} = \frac{1}{n} (\Sigma y_i - b_1 \Sigma x_i)$$

$$= \frac{1}{12} (2.690 - (1.0100)(2.461))$$

$$= 0.0170$$

The following regression equation results:

$$\hat{y} = 0.0170 + 1.0100 \, (x)$$

The results of this equation (predicted y) are shown for each of the original "y" and "x" data points in Table 4.2.

Table 4.2. Actual and Predicted Resilient Modulus Values for Operator Y

Specimen No.	Resilient Modulus (psi) at 77°F		
	Actual Operator y	Predicted Operator y*	Operator x
1	204,000	214,000	195,000
2	231,000	226,000	207,000
3	227,000	217,000	198,000
4	228,000	223,000	204,000
5	261,000	248,000	229,000
6	195,000	199,000	180,000
7	225,000	225,000	206,000
8	216,000	221,000	202,000
9	205,000	201,000	182,000
10	232,000	254,000	235,000
11	205,000	205,000	186,000
12	261,000	256,000	237,000

* Rounded to nearest 1,000 psi

4.3.4 Sum of squares (or the basic information needed to evaluate how "good" a regression line "fits" the data)

4.3.4.1 Total sum of squares (SSTO) [after Ref. 4.5]

If all y_i data points were identical, then all $y_i = \bar{y}$, which would mean there would be no need for any statistical measure (regression line, etc). However, this is rarely the case with real data. To best evaluate the "fit" of a regression line to actual data, three types of sum of squares measures will be examined, the first of which is illustrated in Figure 4.4 (a) and is denoted the total sum of squares (SSTO):

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

If SSTO = 0, then all data points must have the same value. The larger the SSTO, the greater the difference between the y_i data points. The calculation of SSTO for the five data points shown in Figure 4.4 (a) is

$$SSTO = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + (y_4 - \bar{y})^2 + (y_5 - \bar{y})^2$$

$$= \sum_{i=1}^{5} (y_i - \bar{y})^2$$

Use squared terms, since the sum of deviations of $y_i - \bar{y}$ by definition is zero (which gets us nowhere).

4.3.4.2 Error sum of squares (SSE)

The "fit" of the regression line to actual data results in some error ("lack-of-fit"), as shown in Figure 4.4 (b). This error is defined as

$$y_i - \hat{y}_i$$

Thus, the error sum of squares (SSE) is defined as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

or for Figure 4.4 (b), as

$$SSE = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2 + (y_5 - \hat{y}_5)^2$$

Stated another way, the SSE is the amount of the sum of squares best explained by the mean ($\bar{y}$) of the dependent variable. The SSE = SSTO when all $\bar{y} = \hat{y}$.

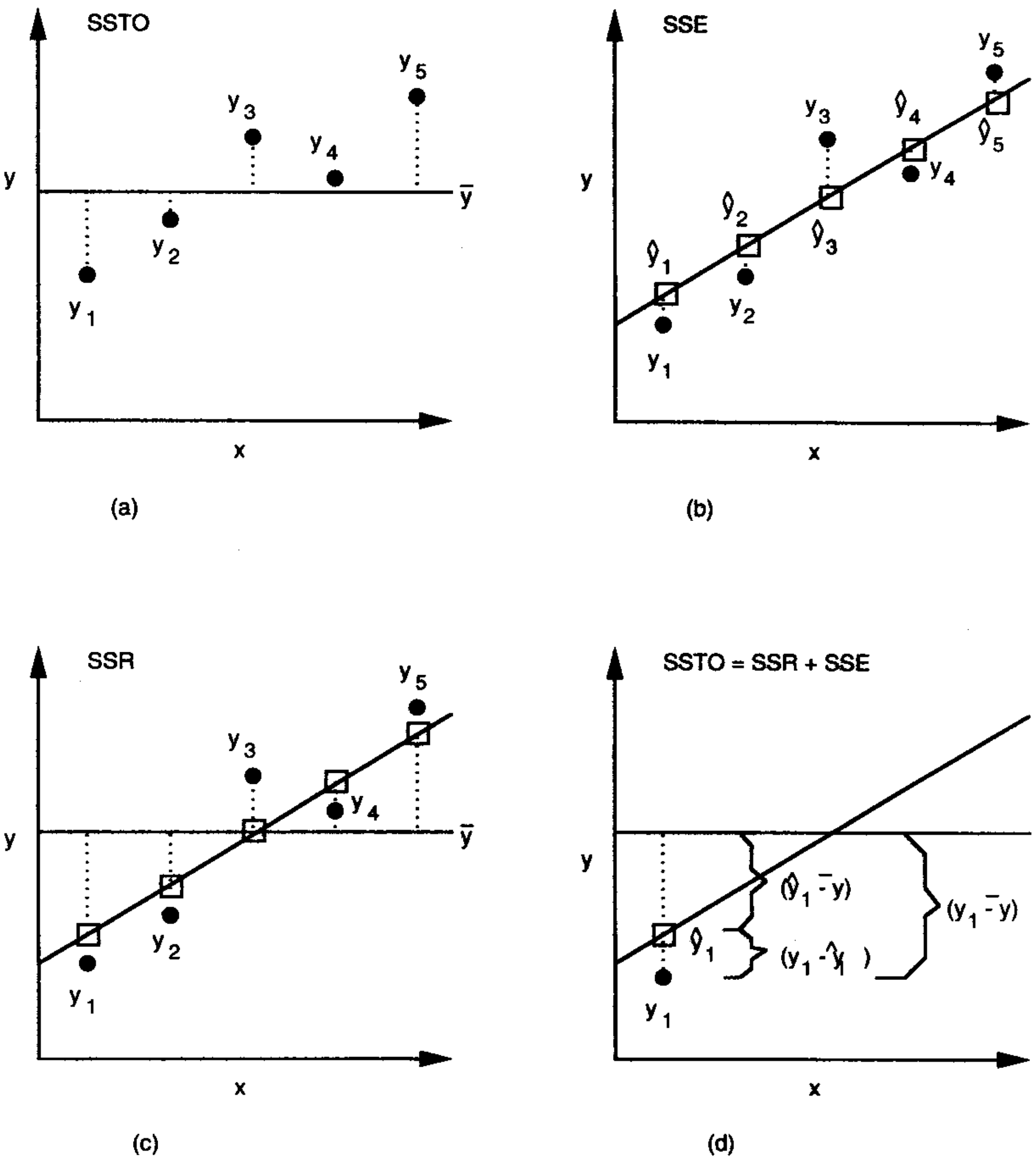


Figure 4.4. Illustration of Deviations Used to Determine Sum of Squares

4.3.4.3 Regression sum of squares (SSR)

Hopefully, for a regression line you wish to develop, the SSTO is much larger than the SSE. The difference is termed the regression sum of squares (SSR):

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

These deviations are illustrated by the dashed lines shown in Figure 4.4 (c). Since the SSR is composed of deviations between the "fitted" regression line and the mean of the data points, the larger the SSR the better the fit of the regression line to the data. Stated another way, the SSR is the amount of the sum of squares explained by the regression equation. For Figure 4.4 (c), SSR is

$$SSR = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + (\hat{y}_3 - \bar{y})^2 + (\hat{y}_4 - \bar{y})^2 + (\hat{y}_5 - \bar{y})^2$$

4.3.4.4 Final overview of sum of squares

From the previous sections, we can see that

$y_i - \bar{y}$	=	$\hat{y}_i - \bar{y}$	+	$y_i - \hat{y}_i$
total deviation		deviation of fitted regression value about the mean		deviation around the regression line

or $\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma (y_i - \hat{y}_i)^2$

$$SSTO = SSR + SSE$$

4.3.5 Regression line "goodness-of-fit"

4.3.5.1 Coefficient of determination (R^2)

The R^2 value explains how much of the total variation in the data is explained by the regression line. Stated another way, the R^2 measures the reduction in the total variation for "y" associated with the use of "x". The $R^2 = 1.0$ when all data points fall on the regression line, as shown in Figure 4.5(a). The $R^2 = 0$ when the regression line matches the average (or mean) of the data points, as illustrated in Figure 4.5 (b). In other words the mean of the data points is as good a predictor of "y" as any line fit through the data points.

For example, if $R^2 = 0.20$, then the total variation in y is reduced by only 20 percent when x is used (on the other hand $r = \sqrt{R^2} = \sqrt{0.2} = 0.45$).

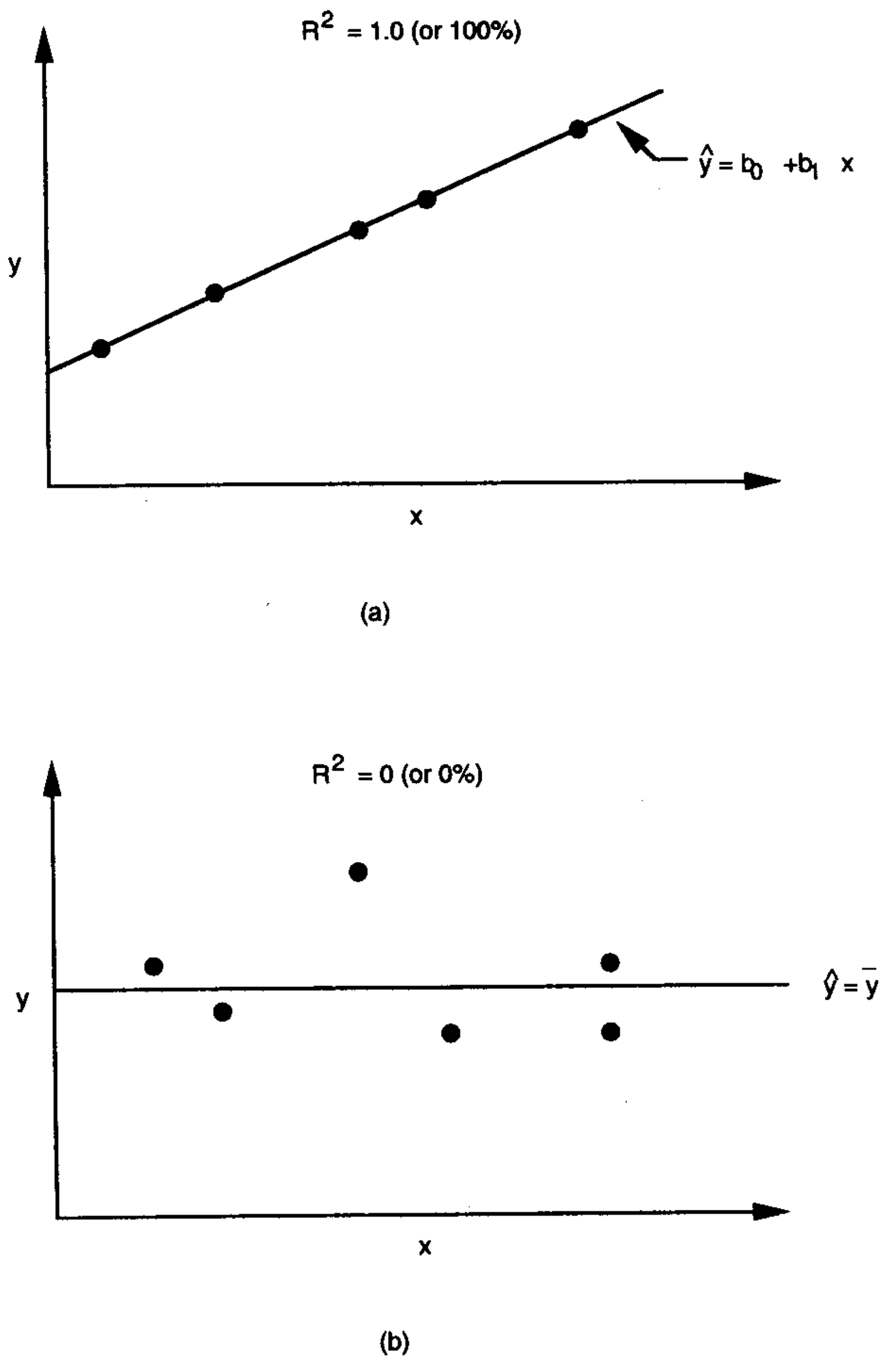


Figure 4.5. Illustrations of the Coefficient of Determination (R^2)

4.3.5.2 Mean square error (MSE) and root mean square error (RMSE)

The mean square error is calculated as follows:

$$MSE = \frac{SSE}{\text{error degrees of freedom}} = \frac{SSE}{n-2}$$

The root mean square error is simply the square root of MSE:

$$RMSE = \sqrt{MSE}$$

The RMSE is the standard deviation of the distribution of $\hat{y}$ for a specific x. Stated another way, the RMSE is the standard deviation of the regression line. The <u>larger</u> the RMSE for a specific regression equation, the <u>poorer</u> the associated predictions.

4.3.6 Example — Straight Line Fit (PSC vs. Age)

An example with six data points is shown in Figure 4.6. Assume that these data resulted from measurement of pavement distress (hence Pavement Structural Condition (PSC)) over a 15-year period. The regression equation would allow the estimation of PSC at any age. First, attempt a "straight line" fit of the data.

(a) Determine the regression coefficients (b_0, b_1)

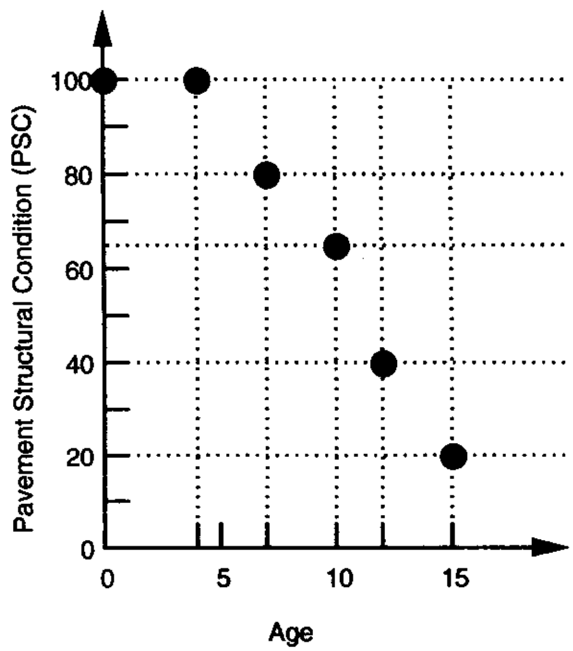
Data Point	y (PSC)	x (AGE)	x^2	xy
1	100	0	0	0
2	100	4	16	400
3	80	7	49	560
4	65	10	100	650
5	40	12	144	480
6	20	15	225	300
	405	48	534	2,390
	Σy	Σx	Σx^2	Σxy

$$\bar{y} = \frac{405}{6} = 67.5 \text{ and } \bar{x} = \frac{48}{6} = 8.0$$

$$b_1 = (\text{from Eq. 4.5}) = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{(\Sigma x^2) - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{2{,}390 - \frac{(48)(405)}{6}}{534 - \frac{(48)^2}{6}}$$

$$= -5.667$$



Data Point	y (pavement condition)	x (age)
1	100	0
2	100	4
3	80	7
4	65	10
5	40	12
6	20	15

Figure 4.6 Example Data - Pavement Condition versus Age

$b_0 = \text{(from Eq. 4.7)} = \frac{1}{n}(\Sigma y - b_1 \Sigma x)$

$= \frac{1}{6}(405 - (-5.667)(48))$

$= 112.836$

The following regression equation results (and refer to Figure 4.7a):

$$\hat{y} = 112.836 - 5.667\,(x)$$

or

PSC = 113 - 5.7 (AGE)

This equation results in the following predicted values for PSC:

Data Point	Predicted PSC	Actual PSC	Actual AGE
1	112.836	100	0
2	90.169	100	4
3	73.169	80	7
4	56.169	65	10
5	44.836	40	12
6	27.836	20	15

(b) Total sum of squares (SSTO)

$$SSTO = \sum_{i=1}^{6} (y_i - \bar{y})^2$$

$= (100-67.5)^2 + (100-67.5)^2 + (80-67.5)^2 + (65-67.5)^2 + (40-67.5)^2 + (20-67.5)^2$

$= 1{,}056.25 + 1{,}056.25 + 156.25 + 6.25 + 756.25 + 2{,}256.25$

$= 5{,}287.50$

(c) Error sum of squares (SSE)

$$SSE = \sum_{i=1}^{6} (y_i - \hat{y}_i)^2$$

$= (100-112.836)^2 + (100-90.169)^2 + (80-73.169)^2 + (65-56.169)^2 + (40-44.836)^2 + (20-27.836)^2$

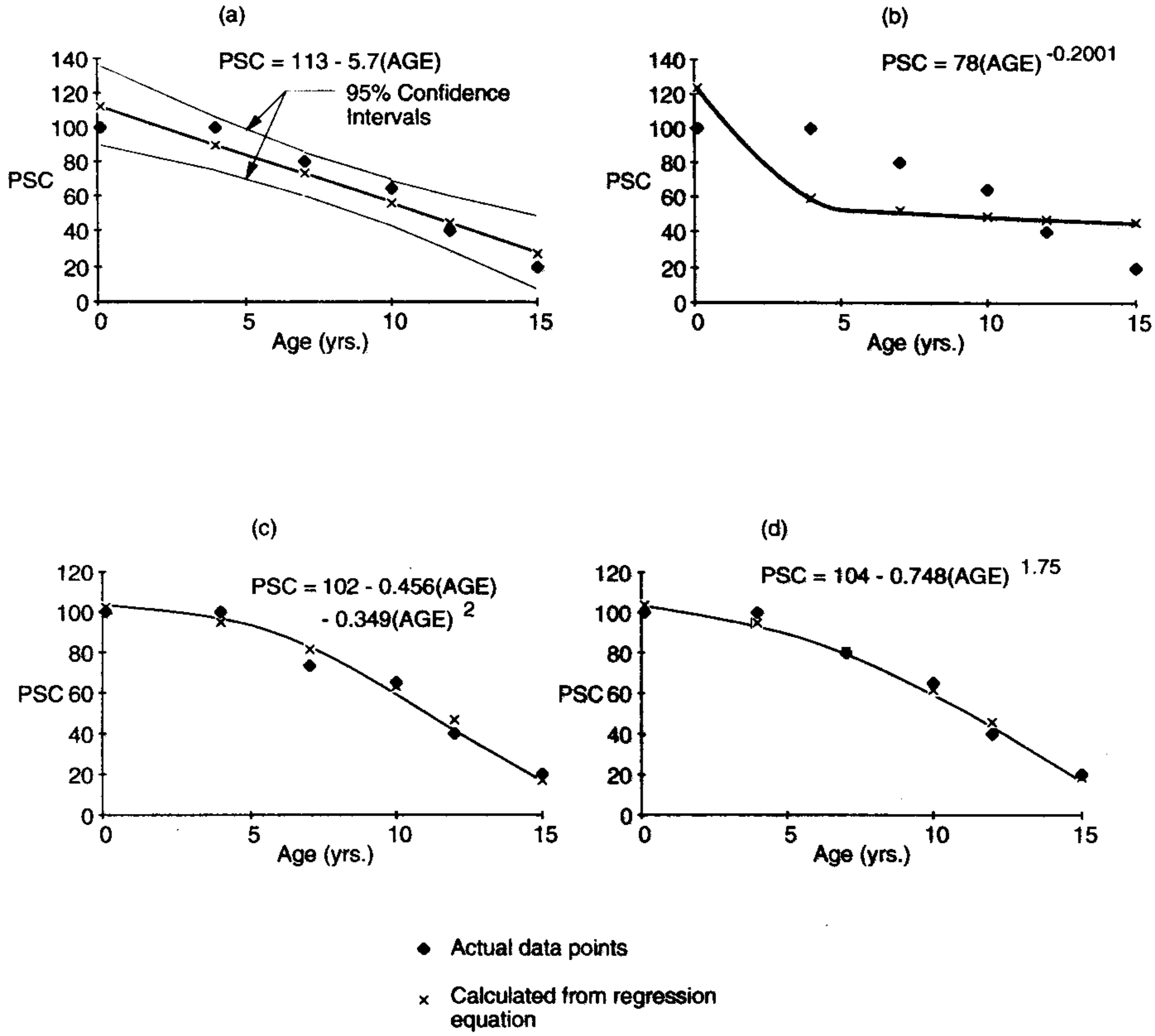


Figure 4.7. Fitted Regression Equations for PSC versus AGE Examples

$= 164.763 + 96.649 + 46.663 + 77.987$
$+ 23.389 + 61.403$

$= 470.854$

(d) Regression sum of squares (SSR)

$$SSR = \sum_{i=1}^{6} (\hat{y}_i - \bar{y})^2)$$

$= (112.836\text{-}67.5)^2 + (90.169\text{-}67.5)^2 + (73.169\text{-}67.5)^2 + (56.169\text{-}67.5)^2 +$
$(44.836\text{-}67.5)^2 + (27.836\text{-}67.5)^2$

$= 2{,}055.353 + 513.884 + 32.138 + 128.392 +$
$513.657 + 1{,}573.233$

$= 4{,}816.657$

(e) SSTO = SSR + SSE = 4,816.657 + 470.854 = 5,287.511 (checks with (b))

(f) Coefficient of determination (R^2)

$$R^2 = \frac{SSR}{SSTO} = \frac{4{,}816.657}{5{,}287.500} = 0.911 \text{ (or } 91.1\%)$$

(g) Mean square error (MSE) and root mean square error (RMSE)

$$MSE = \frac{SSE}{df_{error}} = \frac{SSE}{n-2} = \frac{470.854}{4} = 117.714$$

$$RMSE = \sqrt{117.714} = 10.850$$

The RMSE is the standard deviation of the distribution of y for a fixed x. If you wish to determine the interval estimate of y for a given x, as illustrated in Figure 4.7a, first calculate [after Ref. 4.6]:

$$s_{\hat{y}} = RMSE\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}\right]^{1/2}$$

The interval estimate is then

$$\hat{y} \pm ts_{\hat{y}}$$

This interval is the "narrowest" for $\bar{x}$ and gets larger as one moves further away from $\bar{x}$ (either larger or smaller). For our initial example of PSC versus AGE:

Data Point	Actual x (AGE)	Actual y (PSC)	Predicted y ($\hat{y}$)	$(x_i - \bar{x})^2$	Standard Deviation ($s_{\hat{y}}$)
1	0	100	112.84	64	8.4
2	4	100	90.17	16	5.7
3	7	80	73.17	1	4.5
4	10	65	56.17	4	4.8
5	12	40	44.84	16	5.7
6	15	20	27.84	49	7.6
				150	

If you wish to place a 95% confidence interval estimate for the mean values of $\hat{y}$ for specific x values on the regression line, then

$$t_{table} = 2.776 \text{ for two-tail } \alpha = 5\%, \nu = n - 2 = 6 - 2 = 4$$

Thus, the confidence intervals are as follows:

Data Point	Predicted y ($\hat{y}$)	Standard Deviation ($s_{\hat{y}}$)	95% Confidence Interval ($\hat{y} \pm ts_{\hat{y}}$)
1	112.8	8.4	89.5, 136.1
2	90.2	5.7	74.4, 106.0
3	73.2	4.5	60.7, 85.7
4	56.2	4.8	42.9, 69.5
5	44.8	5.7	29.0, 60.6
6	27.8	7.6	6.7, 48.9

This interval estimate is plotted in Figure 4.7(a). It illustrates the uncertainty associated with the regression line, particularly as one moves away from the $\bar{y}$ and $\bar{x}$ values.

The MSE (or RMSE) can also be used for testing whether the regression coefficients (b_0, b_1) are significantly different from zero. This is illustrated for the PSC vs. AGE example:

H_0: $b_0 = 0$ (null)

H_1: $b_0 \neq 0$ (alternative)

$$t_{calc} = \frac{b_0 - 0}{s_{b_0}}$$

where $$s_{b_0} = \left[MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right)\right]^{1/2}$$

$$= \left[117.7\left(\frac{1}{6} + \frac{8^2}{150}\right)\right]^{1/2} = 8.4$$

$$\therefore t_{calc} = \frac{112.8 - 0}{8.4} = 13.4$$

$$\text{If } t_{calc} > t_{table}\,(1 - \propto, n - 2), \text{ conclude } H_1$$

$$t_{calc} = 13.4 > 2.776 = t_{table}$$

Since $t_{calc} > t_{table}$, you can conclude that b_0 is <u>different</u> from zero.

For b_1:

H_0: $b_1 = 0$ (null)
H_1: $b_1 \neq 0$ (alternative)

$$t_{calc} = \frac{b_1 - 0}{s_b}$$

where

$$s_b = \left[MSE \left(\frac{1}{\Sigma (x_i - \bar{x})^2} \right) \right]^{1/2}$$

$$= \left[117 \left(\frac{1}{150} \right) \right]^{1/2} = 0.9$$

$$\therefore t_{calc} = \frac{-5.667 - 0}{0.9} = -6.3$$

Since $t_{calc} = -6.3 > -2.776 = t_{table}$, conclude H_1, i.e., b_1, is <u>different</u> from zero.

As a rough rule-of-thumb, the t_{calc} values automatically calculated from statistical software for the regression b values should equal or exceed the following (for $\propto = 5\%$):

Number of Data Points	t_{calc} (equal to or greater)
3	13
4	4
5-7	3
8-13	2.5
14 or more	2

(h) Since we have already spent so much time on this example, we will introduce the concept of "residuals." A residual is simply the difference between the observed value (actual data point) and the fitted (predicted) value:

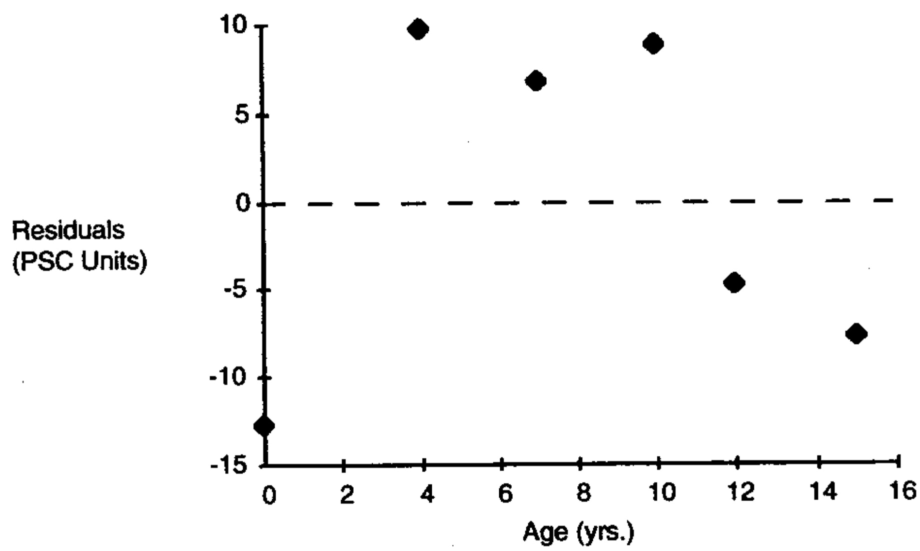
$$\text{Residual} = y_i - \hat{y}$$

Generally, residuals are examined by plotting them versus the independent variable (x). If the residuals generally fall within a band centered on Residuals = 0, then this suggests the model is appropriate. A definite pattern in the plotted residuals indicates a poor model. The significance of this is that a random distribution of the plotted residuals indicates that there is no systematic error in predicting y (PSC in this case) due to x (Age). A systematic residual trend indicates that x is producing some systematic prediction error.

The residuals for this example are:

Data Point	(y_i) Actual (PSC)	($\hat{y}$) Predicted PSC	Residual
1	100	112.836	-12.836
2	100	90.169	+9.831
3	80	73.169	+6.831
4	65	56.169	+8.831
5	40	44.836	-4.836
6	20	27.836	-7.836

The plot of the residuals versus AGE follows:



This plot suggests a better model should be attempted since there is a bit of a trend apparent from AGE = 4 to 15 years.

4.3.7 Example — Power Fit (PSC vs. Age)

Since we can detect a modest "curve" to the data points in Figure 4.6, try an equation with the following form (someone told us a "power" (or non-linear) model would probably work best):

$$PSC = b_0 \,(AGE)^{b_1}$$

The log transformation process is required to obtain the regression constants, in the case of b_0 and b_1. (Additional information on transformations is presented in Section 4.4.)

(a) Transform the data to logs (base 10):

Data Point	y (PSC)	x (AGE)	log y (log (PSC))	log x (log (AGE))
1	100	0*	2.0000	-1.0000
2	100	4	2.0000	0.6021
3	80	7	1.9031	0.8451
4	65	10	1.8129	1.0000
5	40	12	1.6021	1.0792
6	20	15	1.3010	1.1761

* set AGE = 0 to 0.1 (cannot take log of 0.0)

(b) Determine regression constants (b_0, b_1):

Data Point	log (PSC)	log(AGE)	$(\log(AGE))^2$	(log(PSC))(log(AGE))
1	2.0000	-1.0000	1.0000	-2.0000
2	2.0000	0.6021	0.3625	1.2042
3	1.9031	0.8451	0.7142	1.6083
4	1.8129	1.0000	1.0000	1.8129
5	1.6021	1.0792	1.1647	1.7290
6	1.3010	1.1761	1.3832	1.5301
	10.6191	3.7025	5.6246	5.8845
	Σy	Σx	Σx^2	Σxy

$$\bar{y} = \frac{10.6191}{6} = 1.7699 \text{ and } \bar{x} = \frac{3.7025}{6} = 0.7838$$

$$b_1 = \text{(from Eq. 4.5)} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{(\Sigma x^2) - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{5.8845 - \frac{(3.7025)(10.6191)}{6}}{5.6246 - \frac{(3.7025)^2}{6}} = \frac{-0.6684}{3.3398}$$

$$= -0.2001$$

$$b_0 = \text{(from Eq. 4.7)} = \frac{1}{n}(\Sigma y - (b_1)(\Sigma x))$$

$$= \frac{1}{6}(10.6191 - (-0.2001)(3.7025))$$

$$= 1.8933$$

The following regression equation results (and refer to Figure 4.7b):

$$\log \hat{y} = b_0 + b_1 (\log x)$$

$$\log \hat{y} = 1.8933 - 0.2001 (\log x)$$

$$\hat{y} = (10^{b_0})(x)^{b_1} = (10^{1.8933})(x)^{-0.2001}$$

$$\hat{y} = 78.2168(x)^{-0.2001}$$

or

$$PSC \cong 78.2 (AGE)^{-0.2001}$$

(c) The basic regression equation statistics are (without calculations shown):

$$R^2 = 35.8\%$$

$$RMSE = 0.2448$$

As one can see, the RMSE value is significantly smaller than shown in Paragraph 4.3.6(g). The question arises "How does one interpret RMSE in this case?" Since the variable transformation was in terms of logs, so is the RMSE. For example, if AGE = 10 years, then the predicted PSC is:

$$\log PSC = 1.8933 - 0.2001 (\log 10)$$

$$= 1.6932$$

and the log PSC ± 1 RMSE is:

$$1.6932 + 1(0.2448) = 1.9380$$

$$1.6932 - 1(0.2448) = 1.4484$$

thus, PSC ± 1 RMSE is:

$$PSC + 1\ RMSE = 10^{1.9380} = 86.7$$

$$PSC - 1\ RMSE = 10^{1.4484} = 28.1$$

Thus, a "small" RMSE in this case is not so small!

Therefore, the "power fit" is not an improvement over a straight line fit (as determined in Section 4.3.6). Refer to Figure 4.7b for the plotted curve. The equation $PSC = 78.2168 (AGE)^{-2.001}$ results in the following <u>predicted</u> values for PSC:

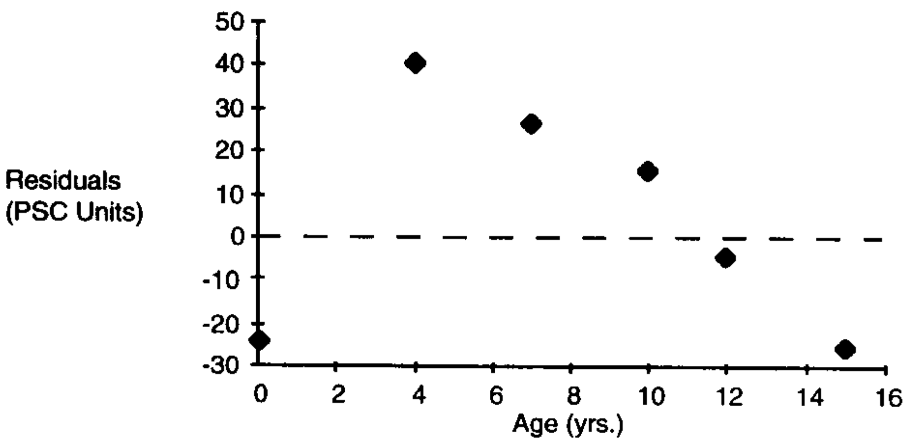
Data Point	Predicted PSC	Actual PSC	Actual AGE
1	124.0	100	0.1
2	59.3	100	4
3	53.0	80	7
4	49.3	65	10
5	47.6	40	12
6	45.5	20	15

Note that if AGE = 0 in this equation, then PSC = 0. Thus, the "form" of the equation is essentially useless (the R^2 is quite low as well).

(d) Check the residuals (refer back to Paragraph 4.3.6(h))

Data Point	Actual PSC	Predicted PSC	Residual
1	100	124.0	-24.0
2	100	59.3	+40.7
3	80	53.0	+27.0
4	65	49.3	+15.7
5	40	47.6	-7.6
6	20	45.5	-25.5

Now plot the residuals versus AGE:



The above residual plot shows a clear systematic trend about the Residual = 0 line, thus a better model is needed (however the low R^2 suggested that conclusion already).

4.3.8 Example — Polynomial Fit (PSC vs. Age)

Well, our attempt at a "curve" fit did not work out too well. This time, try a <u>polynomial</u> fit of the PSC versus AGE data. (Refer to Section 4.4.3 for more details on polynomials.) We will try a <u>second degree polynomial</u> with the following general form:

$$y = b_0 + b_1 (x) + b_2 (x)^2$$

or

$$PSC = b_0 + b_1(AGE) + b_2 (AGE)^2$$

Since this equation has two independent variables (AGE and AGE^2), the calculations get a bit more complex. To keep this example short, a computer program (MINITAB) was used to estimated b_0, b_1, b_2 and R^2 and RMSE. The following equation results:

$$PSC = 102.23 - 0.456 (AGE) - 0.349 (AGE)^2$$

(b_0) (b_1) (b_2)

$R^2 = 98.2\%$

RMSE = 5.56

The above equation is a good fit of the data.

The equation

$$PSC = 102.23 - 0.456 (AGE) - 0.349 (AGE)^2$$

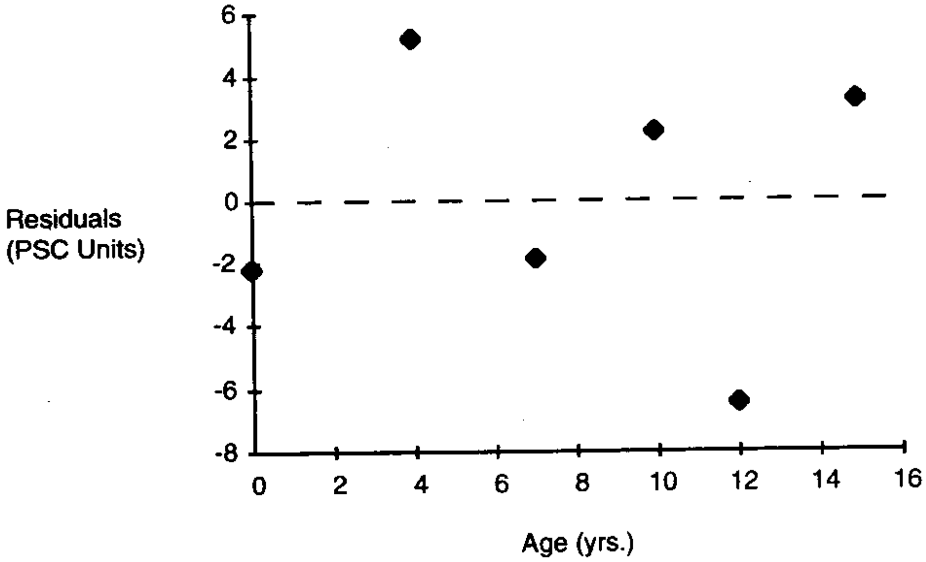
results in the following predicted values for PSC (plotted in Figure 4.7c):

Data Point	Predicted PSC	Actual PSC	Actual AGE
1	102.2	100	0
2	94.8	100	4
3	81.9	80	7
4	62.8	65	10
5	46.5	40	12
6	16.8	20	15

(a) Check the residuals:

Data Point	Actual PSC	Predicted PSC	Residual
1	100	102.2	-2.2
2	100	94.8	+5.2
3	80	81.9	-1.9
4	65	62.8	+2.2
5	40	46.5	-6.5
6	20	16.8	+3.2

Now plot the residuals versus AGE:



The above residual plot shows no systematic trend about the Residual = 0 line, thus it is an improved model over the "straight line" fit of Paragraph 4.3.6.

4.3.9 Example — WSDOT Power Model (PSC vs. Age)

The polynomial fit was good but we do not like the type of equation. Another approach which uses regression analysis which "fixes" the form of the equation is used by the WSDOT Pavement Management System (WSPMS). This form is:

$$PSC = b_0 - b_1 (AGE)^{Power}$$

whereby the "Power" is selected, then b_0, b_1 determined. The power is varied in increments of 0.25 (starting with a Power = 1.0 ranging up to a Power = 3.0) until the best fit is obtained. For example, we will <u>select</u> a Power = 2.5 for illustration purposes.

(a) Power = 2.5, determine b_0, b_1

Data Point	PSC	AGE	$(AGE)^{2.5}$
1	100	0	0.0
2	100	4	32.0
3	80	7	129.6
4	65	10	316.2
5	40	12	498.8
6	20	15	871.4

The resulting b_0, b_1 values are:

$b_0 = 96.95 \cong 97.0$
$b_1 = 0.0956$

The equation becomes:

$PSC = 97.0 - 0.0956\ (AGE)^{2.5}$
$R^2 = 96.3\%$
$RMSE = 7.01$

(b) A summary of various power levels results in the following:

Power = 1.50
$PSC = 106 - 1.48\ (AGE)^{1.50}$
$R^2 = 97.6\%$
$RMSE = 5.58$

Power = 1.75
$PSC = 104 - 0.748\ (AGE)^{1.75}$
$R^2 = 98.4\%$
$RMSE = 4.64$

Power = 2.00
$PSC = 101 - 0.377\ (AGE)^{2.00}$
$R^2 = 98.2\%$
$RMSE = 4.88$

Power = 2.25
$PSC = 98.9 - 0.190\ (AGE)^{2.25}$
$R^2 = 97.4\%$
$RMSE = 5.80$

Thus, the use of a Power = 1.75 produces the "best" fit for our example data.

(c) As a check on our "best fit" equation, the predicted values for PSC are:

Data Point	Predicted PSC	Actual PSC	Actual AGE
1	103.6	100	0
2	95.1	100	4
3	81.0	80	7
4	61.5	65	10
5	45.7	40	12
6	18.1	20	15

These results are plotted in Figure 4.7d.

(d) To verify the process of calculating b_0, b_1, R^2 and RMSE, all necessary calculations will be shown.

Data Point	y (PSC)	AGE	x $(AGE)^{1.75}$	x^2 $((AGE)^{1.75})^2$	xy $((AGE)^{1.75})(PSC)$
1	100	0	0	0	0
2	100	4	11.31	127.92	1,131.00
3	80	7	30.12	907.21	2,409.60
4	65	10	56.23	3,161.81	3,654.95
5	40	12	77.37	5,986.12	3,094.80
6	20	15	114.33	13,071.35	2,286.60
	405		289.36	23,254.41	12,576.95
	Σy		Σx	Σx^2	Σxy

$$\bar{y} = \frac{405}{6} = 67.50 \text{ and } \bar{x} = \frac{289.36}{6} = 48.23$$

$$b_1 = \text{(from Eq. 4.5)} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{12{,}576.95 - \frac{(289.36)(405)}{6}}{23{,}254.41 - \frac{(289.36)^2}{6}}$$

$$= -0.7479$$

$$b_0 = \text{(from Eq. 4.7)} = \frac{1}{n}(\Sigma y - b_1 \Sigma x)$$

$$= \frac{1}{6}(405 - (-0.7479)(289.36)$$

$$= 103.5687$$

This confirms the regression equation obtained by use of MINITAB in that the rounded b_0 and b_1 result in:

$$PSC = 103.6 - 0.748(AGE)^{1.75}$$

This equation results in the following predicted values for PSC:

Data Point	Predicted PSC	Actual PSC	Actual AGE
1	103.57	100	0
2	95.11	100	4
3	81.04	80	7
4	61.51	65	10
5	45.70	40	12
6	18.06	20	15

Total sum of squares (SSTO)

$$SSTO = \sum_{i=1}^{6} (y_i - \bar{y})^2$$

$$= (100 - 67.5)^2 + \ldots + (20 - 67.5)^2$$

= 5,287.50 (same as calculated in Section 4.3.6)

$$\text{Error sum of squares (SSE)} = \sum_{i=1}^{6} (y_i - \hat{y}_i)^2$$

$$= (100 - 103.57)^2 + (100 - 95.11)^2 + (80 - 81.04)^2 + (65 - 61.51)^2 + (40 - 45.70)^2 + (20 - 18.06)^2$$

$$= 12.74 + 23.91 + 1.08 + 12.18 + 32.49 + 3.76$$

$$= 86.16$$

Regression sum of squares (SSR)

$$SSR = \sum_{i=1}^{6} (\hat{y}_i - \bar{y})^2$$

$$= (103.57 - 67.50)^2 + (95.11 - 67.50)^2 + (81.04 - 67.50)^2 + (61.51 - 67.50)^2 + (45.70 - 67.50)^2 + (18.06 - 67.50)^2$$

$$= 1{,}301.04 + 762.31 + 183.33 + 35.88 + 475.24 + 2{,}444.31$$

$$= 5{,}202.11$$

SSTO = SSR + SSE = 5,202.11 + 86.16 = 5,288.27 (checks approximately with SSTO = 5,287.50)

Coefficient of determination

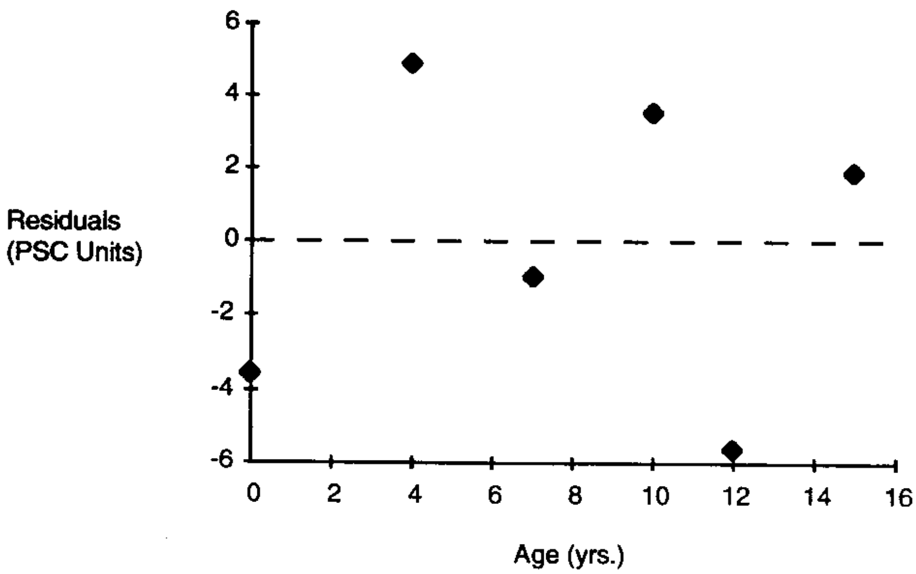
$$R^2 = \frac{SSR}{SSTO} = \frac{5{,}202.11}{5{,}287.50} = 0.9839 \cong 98.4\% \text{ (checks)}$$

$$RMSE = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{86.16}{4}} = 4.6411 \cong 4.64 \text{ (checks)}$$

(d) Check the residuals:

Data Point	Actual PSC	Predicted PSC	Residual
1	100	103.6	-3.6
2	100	95.1	+4.9
3	80	81.0	-1.0
4	65	61.5	+3.5
5	40	45.7	-5.7
6	20	18.1	+1.9

Now plot the residuals versus AGE:



The above residual plot shows no clear systematic trend about the Residual = 0 line, thus it is an improved model over the "straight line" fit of Paragraph 4.3.6.

4.3.10 Example — Nonlinear Fit Using Natural Logarithm (PSR vs. IRI)

If a relationship between PSR (Present Serviceability Rating) and IRI (International Roughness Index) is desired, then the following model should work well:

$$PSR = (5)\, e^{b_1 (IRI)}$$

or transformed for the purpose of performing the regression:

$$\ln\left[\frac{PSR}{5}\right] = (b_1)(IRI)$$

This model is a bit different in that it has a "fixed" intercept (b_0) which is set at 5 (the upper limit of PSR). To develop an example, we will use data from Paterson [4.8] developed for the World Bank.

PSR	IRI (m/km)
4.2	1.0
2.5	3.8
2.0	5.0
1.5	6.6

(a) Determine regression coefficient (b_1):

Data Point	(PSR)	(IRI)	y (PSR/5)	x (IRI)	transformed y $\ln\left(\frac{PSR}{5}\right)$	xy	x^2
1	4.2	1.0	0.84	1.0	-0.17435	-0.17435	1.00
2	2.5	3.8	0.50	3.8	-0.69315	-2.63397	14.44
3	2.0	5.0	0.40	5.0	-0.91629	-4.58145	25.00
4	1.5	6.6	0.30	6.6	-1.20397	-7.94620	43.56
						-15.33597	84.00

and $b_1 = \frac{\sum x_i y_i}{\sum x_i^2}$ (from p. 156, Reference 4.5)

$$b_1 = \frac{-15.33597}{84.00} = -0.18257 \simeq -0.183$$

thus, $\ln\frac{PSR}{5} = -0.183\ (IRI)$

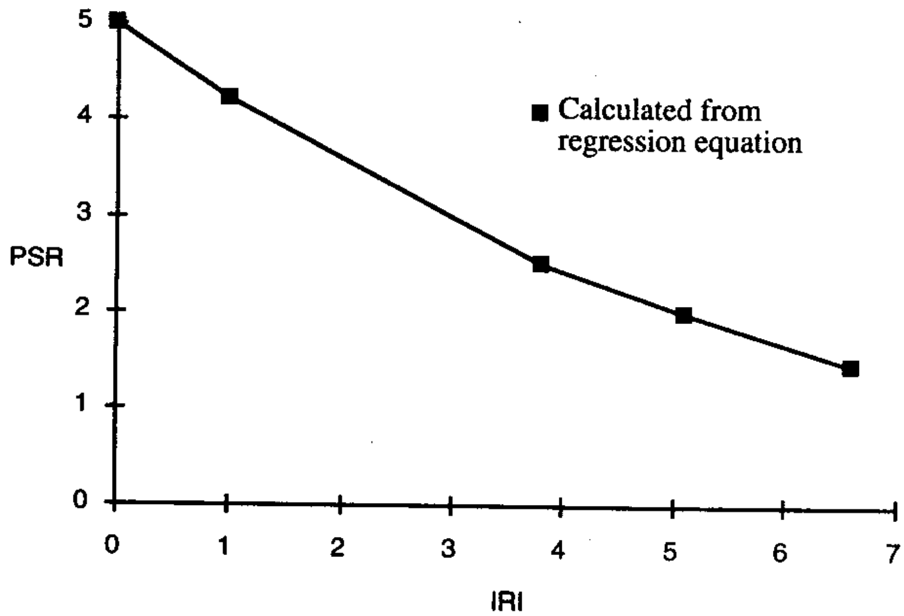
or

$$\frac{PSR}{5} = e^{-0.183\ (IRI)}$$

or

$$PSR = 5e^{-0.183\ (IRI)}$$

The plot of PSR vs. IRI is



(b) Now, check the regression statistics:

Data Point	Predicted PSR	Actual PSR	Actual IRI
1	4.16	4.2	1.0
2	2.49	2.5	3.8
3	2.00	2.0	5.0
4	1.49	1.5	6.6

(i) Total sum of squares (SSTO)

$$SSTO = \sum_{i=1}^{4} (y_i - \bar{y})^2$$

where $$\bar{y} = \frac{(-0.17435) + (-0.69315) + (-0.91629) + (-1.20397)}{4}$$

$$= \frac{-2.98776}{4}$$

$$= -0.74694$$

$$SSTO = (-0.17435 - (-0.74694))^2 + (-0.69315 - (-0.74694))^2$$

$$+ (-0.91629 - (-0.74694))^2 + (-1.20397 - (-0.74694))^2$$

$$= 0.32786 + 0.00289 + 0.02868 + 0.20888$$

$$= 0.56831$$

(ii) Error sum of squares (SSE)

$$SSE = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ = -0.183 (IRI)

Data Point	IRI	$\hat{y}_i$)
1	1.0	-0.1830
2	3.8	-0.6954
3	5.0	-0.9150
4	6.6	-1.2078

$$SSE = (-0.17435 - (-0.183))^2 + (-0.69315 - (-0.6954))^2 + (-0.91629 - (-0.915))^2 + (-1.20397 - (-1.2078))^2$$

$$= 0.000070 + 0.000005 + 0.000002 + 0.000015$$

$$= 0.000092$$

(iii) Regression sum of squares (SSR)

$$SSR = \sum_{i=1}^{4} (\hat{y}_i - \bar{y})^2$$

$$= (-0.183 - (-0.74694))^2 + (-0.6954 - (-0.74694))^2 + (-0.915 - (-0.74694))^2 + (-1.2078 - (-0.74694))^2$$

$$= 0.31803 + 0.00266 + 0.02824 + 0.21239$$

$$= 0.56132$$

(iv) SSTO = SSR + SSE = 0.56132 + 0.000092 = 0.56141 (approx. checks)

(v) Coefficient of determination (R^2)

R^2 = SSR/SSTO = 0.56132/0.56831 = 0.988 (or 98.8%)

(vi) Root Mean Square Error

$$MSE = SSE/df_{error} = \frac{SSE}{n - 1} = 0.000092/3 = 0.000031$$

$$RMSE = \sqrt{0.000031} = 0.00554$$

4.3.11 Exercise — Power Fit (Resilient Modulus vs. Bulk Stress)

An example of a regression analysis one can try involves the use of $\log_{10}$ transformations to estimate the resilient modulus of a subgrade soil from laboratory triaxial tests. The goal is to develop a relationship between resilient modulus and bulk stress (sum of the principal stress, σ_1, $\sigma_2 + \sigma_3$) in the following form:

$$E_R = K_1 (\theta)^{K_2}$$

where E_R = resilient modulus (psi)

θ = Bulk Stress (psi)

K_1, K_2 = regression constants (better known in these notes as b_0, b_1)

The data follow:

Data Point	Resilient Modulus, psi	Bulk Stress, psi
1	13,600	13
2	15,100	14
3	19,100	16
4	21,700	18
5	24,300	20
6	13,600	7
7	15,100	8
8	19,100	10
9	22,900	12
10	25,400	14
11	17,000	4
12	17,000	5
13	19,100	7
14	21,700	9
15	25,400	11

Required:

Estimate K_1, K_2, R^2 and RMSE

Answer:

$K_1 = 12{,}092$

$K_2 = 0.1929$

$R^2 = 17.2\%$

RMSE = 0.0886

Thus, equation is: $E_R = 12{,}092(\theta)^{0.1929}$

(Similar to Paragraph 4.3.7(c), the interpretation of RMSE in this case can be illustrated as follows:

General Form: $\log E_R = \log K_1 + K_2 (\log \theta)$

$\log E_R = \log (12{,}092) + 0.1929 (\log \theta)$

Use $\theta = 10$ psi

then $\log E_R = 4.0825 + 0.1929 (\log 10)$

$= 4.2754$

now calculate $E_R \pm 1$ RMSE

thus $4.2754 + 0.0886 = 4.3640$

or $E_R = 10^{4.3640} = 23{,}121$ psi (+ 1 RMSE)

and $4.2754 - 0.0886 = 4.1868$

or $E_R = 10^{4.1868} = 15{,}374$ psi (- 1 RMSE)

Thus, the interpretation of RMSE is a function of the variable transformation used to develop the "model" being used.)

Can you reproduce the above results? Review Section 4.4.2 for the necessary transformations. Clearly, there is a large amount of scatter in the data which even a regression equation cannot accommodate.

4.4. TRANSFORMATIONS

4.4.1 Transformation of variables

Variable transformations can be used for at least three reasons:

(a) to make data distributions more symmetric (or normal),

(b) to simplify a regression equation that relates two or more variables, and

(c) to create a regression equation form that is more "theoretically" correct.

The three most commonly used transformations are:

(a) square root (e.g. $\sqrt{x}$),

(b) logarithm (e.g. $\log_{10}(x)$),

(c) negative reciprocal (e.g. - 1/x).

The Minitab Handbook (p. 72) is an excellent information source.

4.4.2 Transformation of equations

Transformations of equations can be quite helpful. For example, a commonly used relationship to describe the stress sensitivity of unstabilized soils is (refer back to the exercise in Section 4.3.10):

$$E_R = K_1 (\theta)^{K_2} \qquad \text{Eq. 4.8}$$

(coarse-grained)

or

$$E_R = K_3 (\sigma_d)^{K_4} \qquad \text{Eq. 4.9}$$

(fine-grained)

where E_R = resilient modulus (psi),
θ = bulk stress (psi),
σ_d = deviator stress (psi), and

K_1, K_2, K_3, K_4 = regression constants.

To obtain the regression constants, a transformation of the entire equation is needed:

$$E_R = K_1 (\theta)^{K_2}$$

becomes

$$\log E_R = \log K_1^* + K_2 (\log \theta)$$

and

$$E_R = K_3 (\sigma_d)^{K_4}$$

becomes

$$\log E_R = \log K_3^* + K_4 (\log \sigma_d)$$

To obtain a regression equation in the form of

$$y = b_0 + b_1(x)$$

then $y = \log E_R$

$b_0 = \log K_1^*$ (or $\log K_3^*$)

$b_1 = K_2$ (or K_4)

$x = \log \theta$ (or $\log \sigma_d$)

To convert out of $\log_{10}$, convert from

$$\log E_R = \log K_1^* + K_2(\log \theta)$$

to

$$E_R = 10^{K_1^*}\, \theta^{K_2} = K_1 \theta^{K_2}$$

(letting $10^{K_1^*} = K_1$)

Clearly, a similar transformation is appropriate for Equation 4.9.

As a reminder, helpful $\log_{10}$ relationships include the following:

$$\log xy = \log x + \log y$$

$$\log \frac{x}{y} = \log x - \log y$$

$$\log x^n = n \log x$$

$$\log \sqrt[n]{x} = \frac{\log x}{n}$$

4.4.3 Types of regression models

(a) Linear (refer to Figure 4.8(a))

$$y = b_0 + b_1(x)$$

(b) Exponential (refer to Figure 4.8(b))

$$y = b_0(b_1)^x \text{ (original)}$$

$$\log y = \log b_0 + x \log b_1 \text{ (transformed)}$$

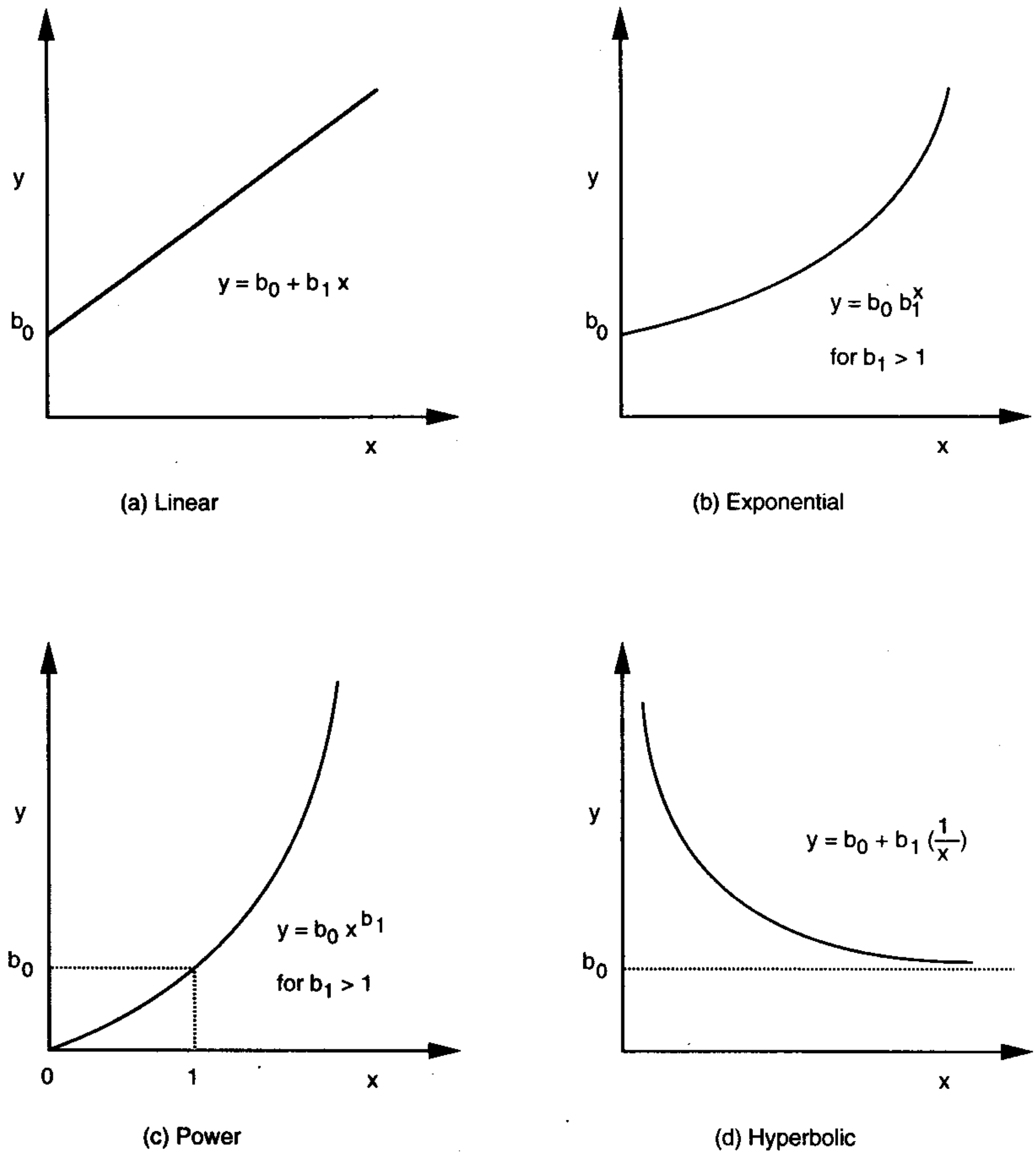


Figure 4.8. Types of Regression Models [after Ref. 4.6]

(c) Power (refer to Figure 4.8(c))

$$y = b_0(x)^{b_1} \text{ (original)}$$

$$\log y = \log b_0 + b_1 \log(x) \text{ (transformed)}$$

(d) Hyperbolic (refer to Figure 4.8(d))

$$y = b_0 + b_1\left(\frac{1}{x}\right) \text{(original)}$$

$$y = b_0 + b_1 (x^1) \text{ (transformed)}$$

$$\text{where } x^1 = \frac{1}{x}$$

(e) Polynomial

$$y = b_0 + b_1x + b_2x^2 + \ldots + b_kx^k \text{ (general form)}$$

(i) First degree polynomial (straight line)

$$y = b_0 + b_1x \text{ (refer to Figure 4.9(a))}$$

(ii) Second degree polynomial (parabola)

$$y = b_0 + b_1x + b_2x^2 \text{ (refer to Figure 4.9(b))}$$

(iii) Third degree polynomial

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 \text{ (refer to Figure 4.9(c))}$$

Note: If you choose a polynomial form, always attempt to obtain the "best fit" with the lowest degree polynomial equation.

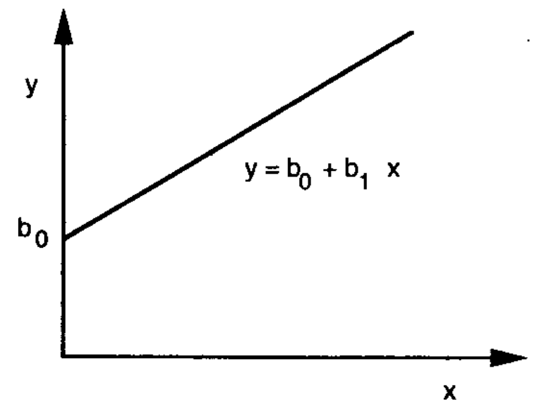
4.5 USE OF REGRESSION EQUATIONS

Any regression equation is only as good as the data used to develop it. Further, such equations should never be used beyond the range of the data from which they were developed. This is particularly important for multiple regression equations.

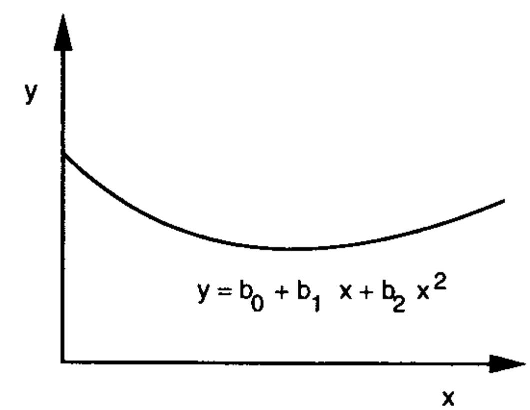
4.5.1 Illustration of Multiple Regression Equation

The originally reported regression equation which correlated Pavement Serviceability Rating (PSR) to Pavement Serviceability Index (PSI) for the AASHO Road Test experiment is as follows [from Ref. 4.7]:

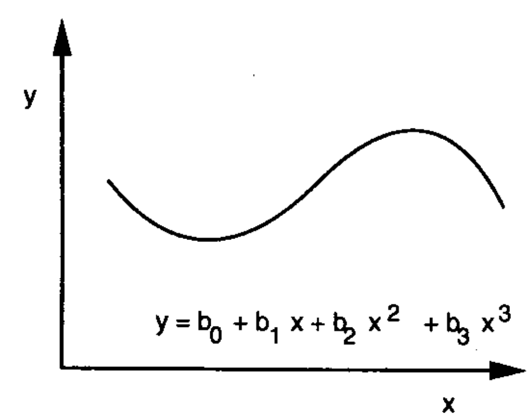
$$PSI = 5.41 - 1.80 \log_{10}(1+\overline{SV}) - 0.09\sqrt{C+P} \qquad \text{Eq. 4.10}$$



(a) First Degree (straight line)



(b) Second Degree (parabola)



(c) Third Degree

Figure 4.9. Polynomial Regression Models [after Ref. 4.6]

where PSI = dependent variable whose source data was the mean PSR ratings from a panel of raters. Mean PSR based on about 10 individual ratings.

$\overline{SV}$ = wheelpath roughness measured by the Road Test longitudinal profilometer in terms of slope variance (mean of the two wheelpaths)

C = length of Class 3 and sealed cracks in feet per 1,000 ft^2

P = patched area in ft^2 per 1,000 ft^2

The summary statistics for Equation 4.10 are:

$R^2 =$ 91.6%

RMSE = 0.32

n = 49 data points

The original data used to develop Equation 4.10 is shown in Table 4.3. To illustrate more examples of regression analysis, the MINITAB computer program was used to see if we could obtain the same result. Using the same independent variable transformations, we get

$$PSI = 5.40 - 1.79 \log_{10} (1 + \overline{SV}) - 0.09 \sqrt{C + P}$$

$R^2 = 91.7\%$

RMSE = 0.32

n = 49 data points.

Thus, we get essentially the same result; however, it might be interesting to develop some different regression equations using the same basic data as follows:

(a) PSR vs. $\log_{10} (1 + \overline{SV})$ Eq. 4.11

$PSI = 5.92 - 2.61 \log_{10} (1 + \overline{SV})$

$R^2 = 88.8\%$

RMSE = 0.37

n = 49

(b) PSR vs. $\sqrt{C + P}$ Eq. 4.12

$PSI = 4.01 - 0.23 \sqrt{C + P}$

$R^2 = 81.2\%$

RMSE = 0.48

n = 49

Table 4.3. Data Used in the Development of AASHO Road Test Correlation Between PSR and PSI for Rigid Pavements [from Ref. 4.7]

State	Mean Panel PSR	Mean Slope Variance in Wheelpath $(\overline{SV})$	Cracking -- Class 3 and Sealed cracks ft/1000 ft^2 (C)	Patching -- Patched Area ft/1000 ft^2 (P)	Transformations log $(1 + (\overline{SV}))$	Transformations $\sqrt{C + P}$
Illinois	2.0	52.0	53	8	1.72	7.8
	4.2	6.5	4	0	0.88	2.0
	2.6	22.2	42	11	1.37	7.3
	2.3	26.2	46	7	1.44	7.3
	1.2	47.8	102	28	1.69	11.4
	2.8	25.5	15	1	1.42	4.0
	4.4	3.2	0	0	0.63	0.0
	1.1	50.8	65	5	1.71	8.4
	0.9	76.8	74	85	1.89	12.6
Minnesota	1.3	43.3	40	59	1.65	10.0
	1.8	24.2	23	66	1.40	9.4
	2.1	24.7	47	41	1.41	9.4
	4.1	2.4	4	0	0.54	2.0
	3.8	4.0	2	0	0.70	1.4
	3.0	7.8	14	1	0.95	3.9
	3.0	7.5	22	0	0.93	4.7
	2.9	9.7	14	0	1.03	3.7
	2.5	17.6	34	0	1.27	5.8
	1.4	59.2	16	12	1.78	5.3
	4.3	3.0	0	0	0.60	0.0
	4.3	4.0	0	0	0.70	0.0
	3.7	5.3	0	0	0.80	0.0
	3.6	4.4	0	0	0.73	0.0
	3.9	5.3	0	0	0.80	0.0
	3.9	6.3	0	0	0.87	0.0
	1.3	32.3	76	1	1.52	8.8
	1.2	27.8	64	0	1.46	8.0
	2.2	25.6	97	1	1.42	9.9
	4.4	4.0	0	0	0.70	0.0
Indiana	4.0	6.6	0	0	0.88	0.0
	3.8	6.6	11	0	0.88	3.3
	3.6	6.8	2	0	0.89	1.4
	3.2	9.8	1	2	1.03	1.7
	2.6	14.6	72	0	1.19	8.5
	2.8	10.4	70	1	1.06	8.4
	1.8	49.4	41	29	1.70	8.4
	1.8	54.5	42	37	1.74	8.9
	2.1	36.6	50	29	1.58	8.9
	2.2	25.1	86	33	1.42	10.9
	1.8	45.4	40	65	1.67	10.2
	2.7	9.9	81	5	1.04	9.3
	4.2	6.1	0	0	0.85	0.0
	4.3	5.2	0	0	0.79	0.0
	4.3	7.1	0	0	0.91	0.0
	1.2	81.9	54	219	1.92	16.5
	2.2	32.2	36	0	1.52	6.0
	4.3	4.6	0	0	0.75	0.0
	2.8	12.6	5	13	1.13	4.2
	2.7	17.8	5	16	1.27	4.6

(c) PSR vs. $\overline{SV}$ Eq. 4.13

$PSI = 3.84 - 0.05\ \overline{SV}$

$R^2 = 73.9\%$
RMSE = 0.57
n = 49

(d) PSR vs. C Eq. 4.14

$PSI = 3.64 - 0.03C$

$R^2 = 59.4\%$
RMSE = 0.71
n = 49

(e) PSR vs. P Eq. 4.15

$PSI = 3.08 - 0.02P$

$R^2 = 26.2\%$
RMSE = 0.95
n = 49

(f) PSR vs. (C + P) Eq. 4.16

$PSI = 3.55 - 0.02\ (C + P)$

$R^2 = 61.2\%$
RMSE = 0.69
n = 49

(g) PSR vs. log [(C + 1) + (P + 1)] Eq. 4.17

$PSI = 4.61 - 1.41 \log_{10} [(C + 1) + (P + 1)]$

$R^2 = 84.1\%$
RMSE = 0.44
n = 49

(h) PSR vs. $\log_{10} (1 + \overline{SV})$, $\log_{10}[(C + 1) + (P + 1)]$ Eq. 4.18

$PSI = 5.56 - 1.64 \log_{10} (1 + \overline{SV}) - 0.62 \log_{10}[(C + 1) + (P + 1)]$

$R^2 = 92.7\%$
RMSE = 0.30
n = 49

(i) PSR vs. $\log_{10} (1 + \overline{SV})$, $\sqrt{C + P}$ Eq. 4.19

Illinois data only:

$PSI = 5.73 - 1.73 \log_{10} (1 + \overline{SV}) - 0.13\sqrt{C + P}$

$R^2 = 95.5\%$
RMSE = 0.31
n = 9

Minnesota data only: Eq. 4.20

$$PSI = 5.48 - 2.08 \log_{10} (1 + \overline{SV}) - 0.08\sqrt{C + P}$$

$R^2 = 94.5\%$
RMSE = 0.28
n = 20

Indiana data only: Eq. 4.21

$$PSI = 5.37 - 1.62 \log_{10} (1 + \overline{SV}) - 0.09\sqrt{C + P}$$

$R^2 = 95.2\%$
RMSE = 0.23
n = 20

From the prior equations, we can conclude:

- If only one independent variable is to be used $\overline{SV}$ (or its transformation) is the single best predictor of PSR (Eq. 4.11 and Eq. 4.13). The worst single predictor is Patching (Eq. 4.15). The combined C + P term as a log (Eq. 4.17) can predict PSR almost as well as $\log_{10}\overline{SV}$ (Eq. 4.11)

- The fit of the data can be improved (Eq. 4.18) over the originally developed equation (Eq. 4.10); however, the improvement is rather small.

- The three separate equations (Eq. 4.19, 4.20, 4.21) based on data from the three states in the form of the original equation (Eq. 4.10), all fit the data rather well.

SECTION 4.0
REFERENCES

4.1 Mills, Blake D., "The Origin of the Term Regression as Applied to Plotted Curves," Personal Memorandum, Mechanical Engineering Department, University of Washington, Seattle, WA, September 1979.

4.2 Ryan, Barbara F., Joiner, Brian L., and Ryan, Thomas A., Minitab Handbook, Second Edition, PWS Publishers, Boston, MA, 1985.

4.3 Button, Joe W. and Mahoney, Joe P., "Statistical Summary of Resilient Modulus Measurements," Research Brief No. 2, Texas Transportation Institute, Texas A&M University, College Station, Texas, July 1977.

4.4 Mahoney, Joe P., "A Study of the Skid Performance of Limestone Rock Asphalt Surface Courses," Report submitted to the Institute of Statistics, Texas A&M University, College Station, Texas, May 1978.

4.5 Neter, John, and Wasserman, William, Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, Illinois, 1974.

4.6 Blank, Leland, Statistical Procedures for Engineering, Management, and Science, McGraw - Hill Book Company, 1980.

4.7 Highway Research Board, "The AASHO Road Test — Report 5 — Pavement Research," Special Report 61E, Highway Research Board, Washington, D.C., 1962, p. 298.

4.8 Paterson, William D.O., "Road Deterioration and Maintenance Effects," Published for the World Bank by the Johns Hopkins Press, Baltimore, Maryland, 1987.

SECTION 5.0

SAMPLING

5.1 INTRODUCTION

The purpose of a sample survey is to make inferences about the sampled "population" [5.1]. For many WSDOT applications, the population would be the WSDOT maintained highway network.

In any sampling process, two factors affect the usefulness of the data contained in the sample: the size of the sample and the variability of the data within the sample. The goal of most sampling surveys is to keep the sample size as low as possible while keeping the variability of the data below some maximum acceptable limit.

To accomplish the above goal, careful consideration should be given to the sample survey design. Such surveys are generally inexpensive when compared to other data collection procedures but can still represent a significant investment. Enough emphasis cannot be placed on the design of a sampling survey in order to minimize costs while maximizing the information gained with the survey. Some of the sample survey methods available are [5.1, 5.2, 5.3]:

- simple random sampling
- stratified random sampling
- one-stage cluster sampling
- multi-stage cluster sampling (multi-stage sampling)
- systematic sampling

A brief description of each of the above sampling methods follows:

(a) **Simple random sampling**. This method provides that every sample has an equal probability of being chosen from a population.

> **Example**: If all highways in a given geographic area were divided into equal lengths (segments or sections), then each highway segment would have an equal chance of being chosen for the required sample size.

(b) **Stratified random sample**. This is the sampling process whereby a population is divided into strata and then random samples are obtained within the described strata.

> **Example**: Divide the WSDOT route system into the six districts and data estimates were required for each district, then each district could be considered a stratum and each individual highway segment could be randomly selected within each district (thus six strata).

(c) **One-stage cluster sampling**. This process first groups elements within a population together and then the elements are randomly sampled.

Example: If data estimates are required for the entire state route system, counties could be randomly selected throughout the state. Within each selected county all highway segments would be sampled. The pavement segments surveyed are considered to be "clustered" within the selected counties.

(d) **Multi-stage cluster sampling (multi-stage sampling)**. This method is similar to one-stage cluster sampling but takes the process further. Multi-stage clustering allows for larger areas to be clustered together and then randomly sampled. The elements within these clusters are also randomly sampled.

Example: Again, as for the previous example, if data estimates are required for the state route system, then counties within a district can be randomly selected and within those selected counties pavement segments may be randomly selected. This would constitute a two-stage cluster sample if all data within the pavement segment are sampled. If the data are only sampled within the pavement segment, this is simply referred to as a two-stage sample. A three-stage sample would be randomly selecting WSDOT districts within the state, then counties within the selected districts, then pavement sections within the selected counties.

(e) **Systematic sampling**. This process samples every K-th element of a set of data.

Example: If data estimates are required for the state route system and with 39 counties in the state, then every fourth county from a listing of all counties could be selected for a total of ten counties. Within each county selected all highway segments would be sampled in the data collection effort.

In addition to the above sampling methods, combinations of the five presented can be created. For example, a stratified two-stage cluster sample can be taken. Other combinations are possible.

A properly designed highway sample survey can provide the following:

- A less expensive indication of statewide, district or county pavement trends (as opposed to a network-wide survey of the population).
- Year-to-year differences in pavement trends.
- Valuable research tool for various statistical pavement experiments.
- Expansion or reduction to accommodate changing needs.
- More detailed objective data may be obtained since the amount of pavement being surveyed is much smaller than in a mass inventory survey (i.e., a survey of all segments or sections).

5.2 SIMPLE RANDOM SAMPLE

A simple random sample is the most fundamental sampling technique. A recent WSDOT activity using a simple random sample was to "check" the annual visual pavement condition surveys (the annual condition survey results can be defined as a population and constitutes a "mass inventory"). To achieve this goal, a simple random sample was assumed for each district. Each sample unit was assumed to be a one mile long pavement segment located between two mileposts. Thus, if it is further assumed that only one side of a route is rated during the annual survey and there are about 1,000 centerline miles in each district, then a district population is composed of 1,000 individual sample units.

A straightforward way to examine the issue of how many samples are needed to check the annual survey is to use estimates of the standard error. The standard error of a survey is analogous to the standard deviation for a set of data, and, literally, is the standard deviation of means computed from samples taken from a population of data. The standard error decreases as the sample size increases (not exactly a surprise). The estimate of the standard error for a simple random sample is [5.1, 5.4]:

$$SE = \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} \qquad \text{(Eq. 5.1)}$$

where SE = standard error of a simple random sample,

S = standard deviation of the population,

n = number of one-mile highway segments sampled for a specific sample size,

N = total number of one-mile highway segments in a specific district, and

$\frac{n}{N}$ = sampling fraction.

If S is the standard deviation of all PSC values within a district, then assume, for now, various levels. A summary of possible SE estimates for various levels of S and sample sizes are shown in Table 5.1 (all based on calculations using Equation 5.1).

Based on district and state condition rating statistics from Washington [5.5] and Texas [5.4], a district S is about 20 (i.e., the PSC standard deviation = 20). Using this value in Table 5.1, the associated sample size standard errors are:

Sample Size	Condition Rating Standard Error (SE)
1%	6.3
2%	4.4
3%	3.6
4%	3.1
5%	2.8
10%	1.9

Table 5.1. Standard Error Estimates for a Simple Random Sample

Standard Deviation of District PSC ("S")	Standard Error for Various Sample Sizes[1]					
	1%	2%	3%	4%	5%	10%
S = 10 (low)	3.1	2.2	1.8	1.5	1.4	0.9
S = 20 (low-medium)	6.3	4.4	3.6	3.1	2.8	1.9
S = 30 (medium)	9.4	6.6	5.4	4.6	4.1	2.8
S = 40 (high)	12.6	8.9	7.2	6.2	5.5	3.8
S = 50 (very high)	15.7	11.1	9.0	7.7	6.9	4.7

Note

1. If N = 1000, then a 1 percent sample = 10 sample units (each sample unit one-mile long). Similarly, a 10 percent sample = 100 sample units.

An "optimum" sample size of two to three percent appears reasonable. If a two percent sample is used, then 20 individual sample units must be measured (again, assuming a district population of 1,000 sample units). For the six WSDOT districts, this suggests a total of 120 sample units. Of course, other possibilities exist.

For convenience, Table 5.2 is provided which shows various WSDOT route system mileages.

5.3 SAMPLE SIZE DETERMINATION BY USE OF PRECISION

A method which uses probability considerations can also provide an indication of the required number of samples for a sampling plan. The method is based on the fact that the precision of the data estimates improves as the number of samples increases.

The population mean for a given data type lies within an interval defined by the following probability statement:

$$P(\bar{x} - z_{1-\alpha/2}\,SE \leq \mu \leq \bar{x} + z_{1-\alpha/2}\,SE) = 1 - \alpha \qquad \text{(Eq. 5.2)}$$

where: $\bar{x}$ = sample mean,

$z_{1-\alpha/2}$ = standard normal variable at a specified level of significance,

SE = $S/\sqrt{n}$ = sample error of a randomly obtained number of samples,

S = standard deviation of the population,

μ = population mean, and

α = level of significance.

By use of Equation 5.2 we can specify with a 100 $(1 - \alpha)$ percent confidence level that the population mean will fall within an interval length of $\pm d$ which is equal to $\pm z_{1-\alpha/2}\,S/\sqrt{n}$. The interval length also represents the precision of the estimate (or the amount of deviation from the true value in actual units or percent allowed). By rearranging terms the required number of samples for a given confidence level is:

$$n = \left| \frac{z_{1-\alpha/2}\,S}{d} \right|^2 \qquad \text{(Eq. 5.3)}$$

To calculate the required number of samples by use of Equation 5.3, the population standard deviation must be known or estimated and the data precision and confidence level selected.

Table 5.2. WSDOT Mileages

All Route Systems (lane-miles)

District	BST	ACP	PCCP	Total w/o Ramps	Ramps	Grand Total
1	193	2,536	939	3,668	277	3,945
2	1,558	1,037	6	2,601	24	2,625
3	283	2,383	179	2,845	136	2,981
4	76	1,660	195	1,931	83	2,014
5	983	1,193	721	2,897	145	3,042
6	1,486	1,604	135	3,225	57	3,282
	4,579	10,413	2,175	17,167	722	17,889

Interstate

District	Centerline Miles	Lane-Miles
1	208	1,211
2	54	217
3	55	331
4	96	498
5	252	1,069
6	107	445
	772	3,771

By use of our example in Paragraph 5.2 and Equation 5.3 for the following input values:

S = PSC district level population standard deviation

= 20,

z = 1.960 for a 95 percent confidence level, and

d = 10 PSC points,

then, $$n = \left|\frac{(1.960)(20)}{(10)}\right|^2 = 15.4 \simeq 16 \text{ one mile sample segments}$$

If we could only "tolerate" d = 5 PSC points, then (all other inputs the same)

$$n = \left|\frac{(1.960)(20)}{(5)}\right|^2 = 61.5 \simeq 62 \text{ one mile sample segments}$$

Clearly, the selected level of d is critical in selecting a sample size as is the standard deviation (S). This is further illustrated in Table 5.3 which shows the required 95 percent confidence level sample size for various levels of S and d.

Table 5.3. Sample Size Estimates by the Precision Method for a 95% Confidence Level

Standard Deviation of District PSC ("S")	Sample Size (n) for Various Levels of Precision (PSC)[1]			
	d = 2.5	d = 5.0	d = 7.5	d = 10.0
S = 10 (low)	62	16	7	4
S = 20 (low-medium)	246	62	28	16
S = 30 (medium)	554	139	62	35
S = 40 (high)	984	246	110	62
S = 50 (very high)	1,537	385	171	96

Note

1. Sample sizes rounded up to nearest whole number.

SECTION 5.0
REFERENCES

5.1 Cochran, W.G., Sampling Techniques, John Wiley and Sons, Inc., New York, 1963.

5.2 Mendenhall, W., Ott, L., and Scheaffer, R.L., Elementary Survey Sampling, Duxbury Press, Belmont, California, 1971.

5.3 Yamane, T., Elementary Sampling Theory, Prentice-Hall, Englewood, New Jersey, 1967.

5.4 Mahoney, J., "Measurements of Pavement Condition and Performance for Texas Highways Using Sampling Techniques," Dissertation, Texas A&M University, College Station, Texas, May 1979.

5.5 Mahoney, J., Kay, R.K., and Jackson, N.C., "Pavement Performance Equations," Research Report WA-RD 143.1, Washington State Department of Transportation, Olympia, Washington, March 1988.