INTRODUCTION TO STATISTICAL METHODS

1. Introduction

Statistics is the branch of scientific method that is concerned with collecting, arranging, and using numerical observations or data that arise from natural phenomena or experiment. For our purposes, statistical methods will be useful in one or more of the following areas:

1. *Reduction of data*. Much numerical information may often be "condensed" into a simple relationship, together with a statement as to the confidence we may place in the relationship.

2. *Estimates and significance tests*. From experimental data, certain population parameters (such as a mean) can be estimated. It is usually possible to determine whether or not these estimates differ significantly from certain preconceived values.

3. *Reliability of inferences depending on one or more variables*. For example, the total impurity content of a bulk shipment is predicted from samples taken throughout the shipment. We then ask, what reliability can be attached to the prediction of total impurity content? This leads to uncertainty analysis}.

4. *Relationships between two or more variables*. Suppose that some measurable quantity depends on one or more separate factors. Then, if it is possible by *experimental design* to control the separate factors at a series of fixed levels and to observe the measurable quantity at each level, the technique known as *analysis of variance* (ANOVA) is used to evaluate the dependency.

2. Definitions and Notation

The following definitions will be important in later sections:

- *Statistic*-—an item of information deduced from the application of statistical methods.
- *Population*-—a collection of objects (or measurements) having in common some observable or measurable characteristic known as a *variate*.
- *Individual*-—a single member of a population.
- *Sample*-—a group of individuals drawn from a population, usually at random, so that each individual is equally likely to be selected.
- *Random variable* —a numerical quantity associated with the variate. The value of the random variable for a given individual is determined by the value or nature of the variate for that individual. For example, if the variate were the color of an object, values 1, 2, 3, etc. might be assigned to the colors red, green, blue, etc.
- *Continuous variable*-—one that can assume any value within some continuous range, such as the temperature of a reactor.
- *Discontinuous or discrete variable*-—one that can assume only certain discrete values, such as the number of days in a month.

- *Probability*-—*P,* that a random variable *x*, belonging to an individual drawn from a population, shall have some particular value *A*, equals the fraction of individuals in the population having that value *A* associated with them.

Example

A deck of 52 cards contains 13 cards in each of the spade (♠), heart (♥), diamond (♦), and club (♣) suits. What is the probability that two cards dealt randomly from the deck will both} be spades?

Solution

By definition, the probability that the *first* card dealt will be a spade equals the fraction of cards in the deck that are spades, namely, 13/52 = 1/4. Likewise, the probability that the *second* card dealt will be a spade is 12/51, since there are now only 12 spades in the deck, whose overall number has been reduced from 52 to 51.

The combined probability that *both* events will occur is obtained by multiplying the individual probabilities, and is

$$\mathrm{P} = \frac{13}{52} \times \frac{12}{51} = \frac{3}{51} = \frac{1}{17} = 0.0588$$

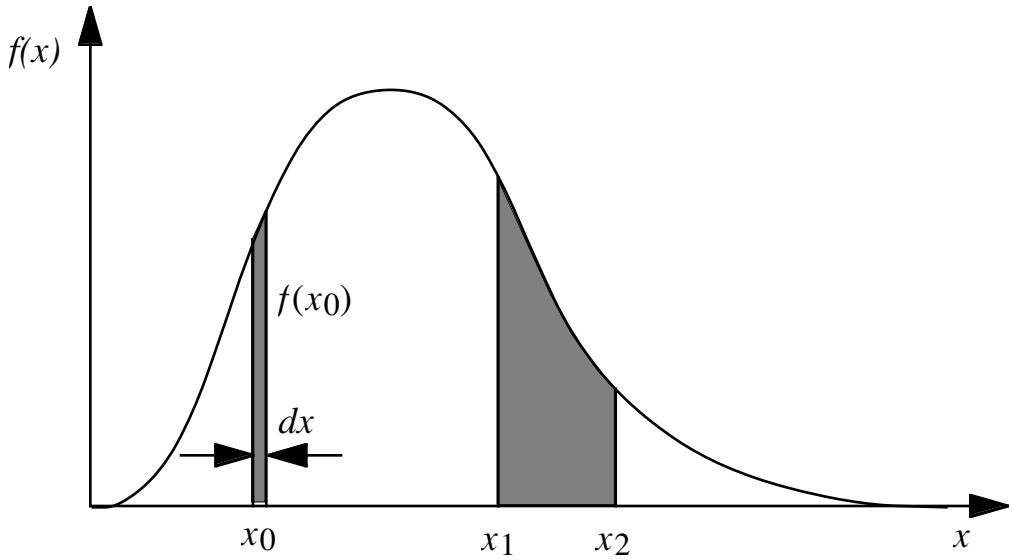

Fig. 1 Frequency function for continuous distribution.

Frequency function — continuous distribution. In general, the probability that a *continuous* random variable *x* will have a specified value is infinitesimally small, since an infinite range of values is possible. However, we can define the probability that the random variable will lie in the very narrow *range* between x_0 and $x_0 + dx$, as shown by the tall thin shaded area at the left of Fig. 1. By definition of the frequency function *f(x)* (also known as the probability density function), this probability is:

$$P(x_1 \le x \le x_0 + dx) = f(x_0)dx \tag{2.1}$$

The probability that the random variable will lie in a *finite* range is obtained by integration under the frequency-function curve, also illustrated by the wider shaded area at the right of Fig. 1:

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} f(x)dx. \tag{2.2}$$

A special case of Eqn. (2.2) is

$$P(-\infty \le x \le \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1, \tag{2.3}$$

since the probability that x must lie somewhere is one. In Eqn. (2.3), the extreme limits of infinity have been used to account for all possibilities. In practice, however, finite limits can often be used with the same effect.

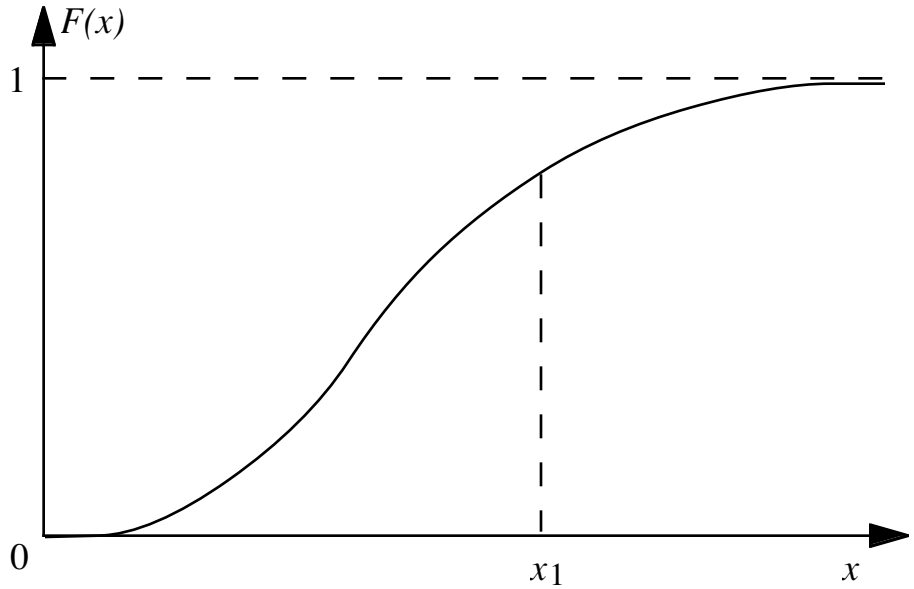

Fig. 2 Cumulative frequency function for continuous distribution.

The cumulative frequency function, $F(x)$, illustrated in Fig. 2, is defined as the probability that the random variable x has a value less than or equal to a specified value such as x_1

$$F(x_1) = P(-\infty \le x \le x_1) = \int_{-\infty}^{x_1} f(x)\,dx. \tag{2.4}$$

Clearly, $F(\infty) = 1$..

It is also possible to consider *functions* of the random variable, such as $y(x)$. For example, if x were the diameter of a sphere, then y could be its volume, $\pi x^3/6$. In this event, the *expected* or *arithmetic average value* of $y(x)$ is defined for a continuous distribution by weighting the function with $f(x)$ and then integrating over the whole spectrum:

$$E(y)[= \text{ave}(y)] = \int_{-\infty}^{\infty} y(x) f(x)\, dx, \tag{2.5}$$

The term "expected" is actually a misnomer, since the probability of any one value occurring is infinitesimally small, as already explained. However, the terminology has stuck, so we shall use it.

For the special case of y being the random variable x itself, the result is the *mean* μ of the distribution:

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{2.6}$$

In addition to the mean, there are two further statistics that give an idea of the "center of gravity" of the distribution:

1. The *median* x_m is defined so that half the values of x lie above it, and half below:

$$F(x_m) = \frac{1}{2}. \tag{2.7}$$

2. The *mode* is the most frequently occurring value of x, and (for a singly peaked distribution, such as in Fig. 1), is the value of x corresponding to the peak of the distribution.

For symmetrical distributions, the above statistics coincide. For the distribution of Fig. 1, which is "skewed" to the left, the mean and median would both lie to the right of the mode.

The *variance* is a measure of the "spread" of a particular distribution and is defined for a continuous distribution as

$$\sigma^2 = \text{var}(x) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \tag{2.8}$$

The standard deviation σ is the square root of the variance:

$$\sigma = \sqrt{\text{var}(x)}. \tag{2.9}$$

A useful formula for $E(x^2)$ is obtained by first noting that

$$\begin{aligned} \sigma^2 &= E[(x-\mu)^2] = E(x^2 - 2x\mu + \mu^2) \\ &= E(x^2) - 2\mu E(x) + E(\mu^2) = E(x^2) - \mu^2, \end{aligned} \tag{2.10}$$

from which it follows that

$$E(x^2) = \mu^2 + \sigma^2. \tag{2.11}$$

Frequency function—discontinuous random variable. With some modifications, the above definitions also carry through for a *discontinuous* random variable. The probability of observing a *specified* value of the random variable, such as x_0, is

$$P(x = x_0) = f(x_0). \tag{2.12}$$

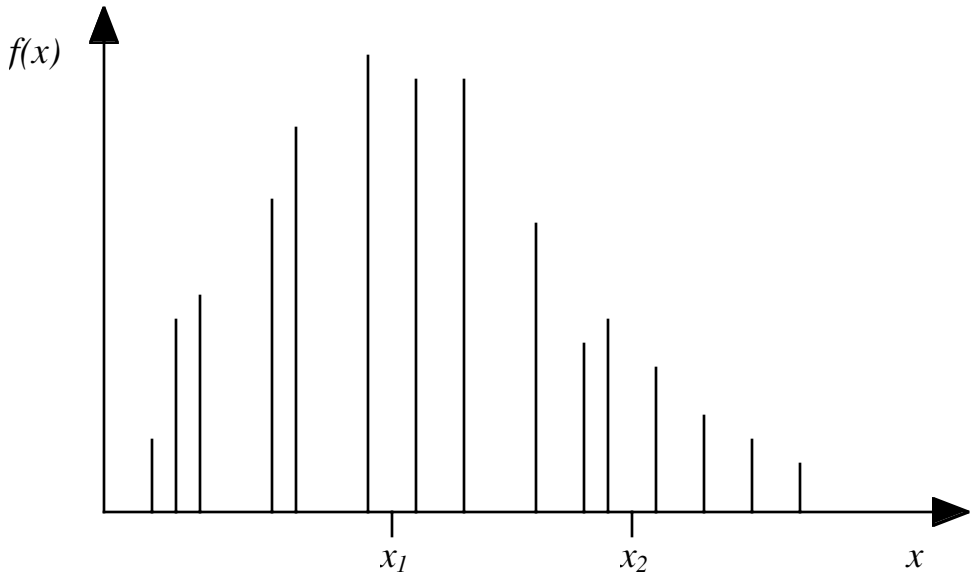

Fig. 3 Frequency function for discontinuous distribution.

The probability that x lies within a range is now obtained by summation rather than integration:

$$P(x_1 \le x \le x_2) = \sum_{x_1}^{x_2} f(x). \tag{2.13}$$

The arithmetic average for a function $y(x)$ is defined as

$$E(y) = \sum_{-\infty}^{\infty} y(x) f(x), \tag{2.14}$$

and the mean and variance of the distribution are similar to Eqns. (2.6) and (2.8), except that summation replaces integration. Finally, for discontinuous distributions, the mean and median may lie between two possible values of x.

Linear combination. Later in these notes, we shall need to examine a *linear combination* or sum of n independent random variables, $x_1, x_2, \ldots, x_n$, each weighted by a corresponding coefficient:

$$y = \sum_{i=1}^{n} a_i x_i = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n. \tag{2.15}$$

The *mean* μ_y of the linear combination is the sum of the individual means μ_i, each weighted by the coefficient a_i:

$$\mu_y = E(y) = E\sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} a_i E(x_i) = \sum_{i=1}^{n} a_i \mu_i . \tag{2.16}$$

The variance σ_y^2 of the linear combination is given by

$$\sigma_y = E(y-\mu_y)^2 = E\left[\sum_{i=1}^{n} a_i (x_i - \mu_i)\right]^2$$

$$= E\left[\sum_{i=1}^{n} a_i^2 (x_i - \mu_i)^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_i a_j (x_i - \mu_i)(x_j - \mu_{j)}\right]. \tag{2.17}$$

But, for x_i and x_j independent,

$$E(x_i - \mu_i)(x_j - \mu_j) = E(x_i - \mu_i)E(x_j - \mu_j) = 0. \tag{2.18}$$

Hence, Eqn. (2.17) yields

$$\sigma_y^2 = \sum_{i=1}^{n} a_i^2 E(x_i - \mu_i)^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 , \tag{2.19}$$

so the variance of the linear combination is the sum of the individual variances, each weighted by the *square* of the corresponding coefficient a_i.

3. Sample Statistics

Consider a sample that comprises n independent observations $x_1,x_2,...,x_n$ on the random variable x. (For example, x could represent the populations of the states in which members of the ChE 360 class normally reside.) Within the limitations of the sample size (it is necessarily limited to a finite number of observations), we would like to *estimate* as accurately as possible some statistics about the population from which x is drawn, notably its mean and variance.

We start by defining the *sample mean* as the arithmetic average of the values $x_1,x_2,...,x_n$ in the sample:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{1}{n} x_i . \tag{3.1}$$

Thus, $\bar{x}$ is a *linear combination* of the x_i , each of which has the same mean μ and variance σ^2, and each of which is weighted by the same coefficient, $a_i = 1/n$. It immediately follows from Eqns. (2.16) and (2.19) that the mean $\mu_{\bar{x}}$ and variance $\sigma_{\bar{x}}^2$ of the sample mean are:

$$\mu_{\bar{x}} = \sum_{i=1}^{n} \frac{1}{n}\mu = \mu . \tag{3.2}$$

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n} . \tag{3.3}$$

Thus, the sample mean is an *unbiased estimate* of the population mean. That is, if a large number of samples is taken, their mean will accurately reflect the population mean. Also, the variance of the sample mean becomes smaller as the sample size is increased ("there is safety in numbers"); however, the decrease is proportional only to $1/n$, so that a quadrupling of sample size would be needed to halve the variance of $\bar{x}$.

The *sample variance* is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ not } \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.4}$$

The *mean* or *expected* value of the sample variance is

$$\mu_{s^2} = E\left(s^2\right) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n-1}E\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right). \tag{3.5}$$

But, from Eqn. (2.11), noting that the variances of x and $\bar{x}$ are σ^2 and [from Eqn. (3.3)] σ^2/n, respectively:

$$E\left(x_i^2\right) = \mu^2 + \sigma^2, \tag{3.6}$$

and

$$E\left(\bar{x}\right)^2 = \mu^2 + \frac{\sigma^2}{n}. \tag{3.7}$$

Hence

$$\mu_{s^2} = \frac{1}{n-1}\left[n\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] = \sigma^2. \tag{3.8}$$

Thus the sample variance as defined by the *first* expression in Eqn. (3.4) is an unbiased estimate of the population variance. If the *second* expression had been used, an underestimate would have resulted.

4. The Normal Distribution

The general concept of the normal distribution is that, superimposed on the *true* value μ for any random variable x, there is a very large number n of very small errors δ. Each δ has an equal probability of being positive or negative. The corresponding frequency function can be shown to be

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{4.1}$$

in which $\sigma^2 = 2n\delta^2$ as $n \to \infty,\ \delta \to 0$.

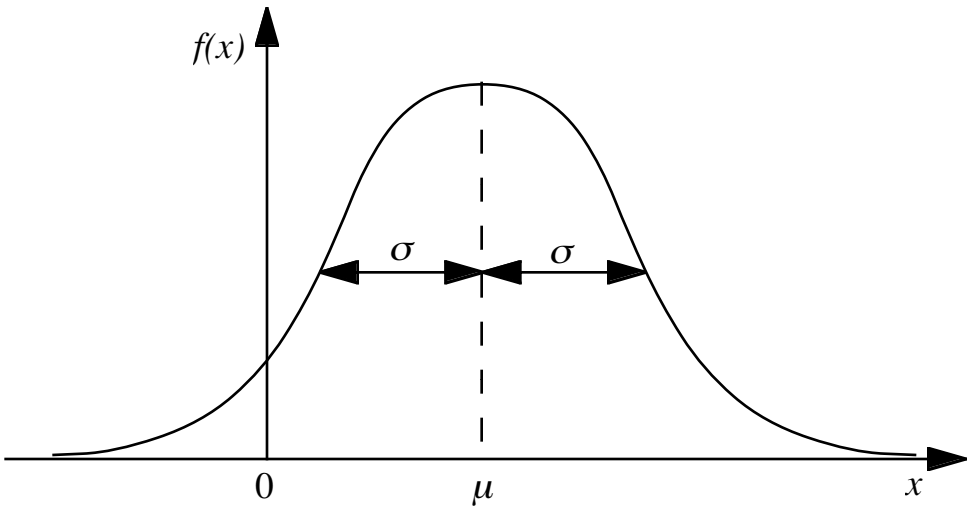

Fig. 4 Normal distribution frequency function.

Examples of variables likely to conform to the normal distribution are:

1. A particular experimental measurement subject to several random errors.
2. The time taken to travel to work along a given route. Here, the variability of traffic lights would afford the "errors"---some positive, some negative---that are superimposed on an average driving time.
3. The heights of women belonging to a certain race. (If "people" were substituted for "women," would you still expect the normal distribution to be obeyed?)
4. The logarithm of the diameter of particles in a powder is an example of a case in which one variable (diameter) does not conform to the normal distribution, but a transformation of that variable (the logarithm) does.

In Eqn. (4.1), $1/\sigma\sqrt{2\pi}$ can be viewed as a normalizing factor that automatically insures that

$$\int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{4.2}$$

By integration—not particularly easy---μ and σ^2 may indeed be shown to be the mean and variance of the distribution:

$$\text{Mean:} \quad E(x) = \int_{-\infty}^{\infty} xf(x)\,dx = \mu. \tag{4.3}$$

$$\text{Variance:} \quad E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \sigma^2. \tag{4.4}$$

A random variable that is normally distributed with mean μ and standard deviation σ is said to be an $\mathrm{N}(\mu,\sigma)$ variable.

Standardized normal distribution. To accommodate normal distributions with different means and variances, we can define a *standardized* random variable ξ by

$$\xi = \frac{x - \mu}{\sigma}, \tag{4.5}$$

which is the deviation of x from its mean μ, as a fraction of its standard deviation σ. The reader should check from Eqn. (4.1) that the corresponding standardized normal frequency function $\phi(\xi)$ is given by

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2}, \tag{4.6}$$

which is also illustrated in Fig. 5. Note that ξ is an N(0,1) variable.

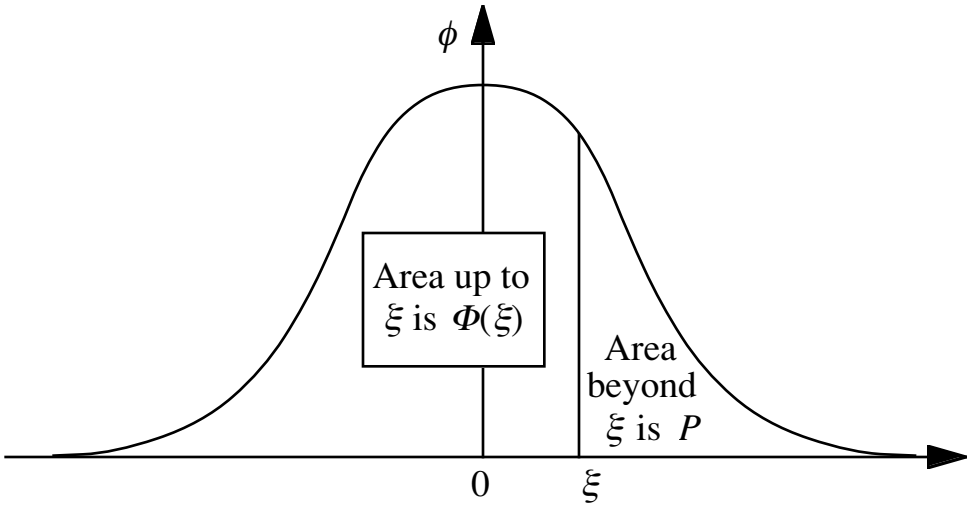

Fig. 5 Standard normal distribution.

The *cumulative frequency function* $\Phi(\xi)$ for the standardized normal variable is the probability that the variable has a value less than or equal to ξ:

$$\Phi(\xi) = \int_{-\infty}^{\xi} \phi(\xi)\, d\xi. \tag{4.7}$$

Table 1 tabulates the frequency function ϕ and the cumulative frequency function Φ for zero and positive values of the standardized normal variable ξ. By referring to Fig. 5, the reader should verify that negative values $-\xi$ are accommodated by the formulas:

$$\phi(-\xi) = \phi(\xi), \tag{4.8}$$

$$\Phi(-\xi) = 1 - \Phi(\xi). \tag{4.9}$$

Table 1 Values of the Standardized Normal Distribution

ξ	ϕ	Φ	ξ	ϕ	Φ
0.0	0.3989	0.5000	2.0	0.0540	0.9772
0.1	0.3970	0.5398	2.1	0.0440	0.9821

0.2	0.3910	0.5793	2.2	0.0355	0.9861
0.3	0.3814	0.6179	2.3	0.0283	0.9893
0.4	0.3683	0.6554	2.4	0.0224	0.9918
0.5	0.3521	0.6915	2.5	0.0175	0.9938
0.6	0.3332	0.7258	2.6	0.0136	0.9953
0.7	0.3122	0.7580	2.7	0.0104	0.9965
0.8	0.2897	0.7881	2.8	0.0079	0.9974
0.9	0.2661	0.8159	2.9	0.0060	0.9981
1.0	0.2420	0.8413	3.0	0.0044	0.9986
1.1	0.2178	0.8643	3.1	0.0033	0.9990
1.2	0.1942	0.8849	3.2	0.0024	0.9993
1.3	0.1714	0.9032	3.3	0.0017	0.9995
1.4	0.1497	0.9192	3.4	0.0012	0.9997
1.5	0.1295	0.9332	3.5	0.0009	0.9998
1.6	0.1109	0.9452	3.6	0.0006	0.9998
1.7	0.0940	0.9554	3.7	0.0004	0.9999
1.8	0.0790	0.9641	3.8	0.0003	0.9999
1.9	0.0656	0.9713	3.9	0.0002	1.0000
2.0	0.0540	0.9772	4.0	0.0001	1.0000

Table 2 Tail Areas of the Standardized Normal Distribution}

P	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
ξ_P	0.843	1.282	1.647	2.056	2.329	2.578	2.880	3.092

Example

Referring to Table 1, what value of ξ has a 20% probability of being exceeded? Check your answer against Table 2.

Solution

We seek a value for ξ below which 80% of the population lies. Thus, the appropriate value of Φ is 0.8. Table 1 indicates that $\Phi(0.8) = 0.7881$ and $\Phi(0.9) = 0.8159$. By linear interpolation, the required value for ξ is approximately

$$0.8 + (0.9 - 0.8) \times \frac{0.8000 - 0.7881}{0.8159 - 0.7881} = 0.8428,$$

which is essentially the same as the first entry in Table 2, namely, $\xi_{P=0.2} = 0.843$.

Example

The times taken to travel from your home to class on n=7 different days are:

30, 33, 26, 23, 30, 35, 27 min.

(a) If you are willing to take a 1\% chance of being late for class, how long before the class starts should you set out?

(b) In this event, what is the probability of your being able to buy a cup of coffee and a donut (which take 12 minutes to consume) before class starts?

Solution

Since the exact population parameters are unknown, we must estimate them from the sample mean and variance, which are:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{30+33+26+23+30+35+27}{7} = 29.14 \text{ min.} \tag{4.10}$$

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

$$= \frac{\left(30^2 + 33^2 + \ldots + 27^2\right) - 7 \times 29.14^2}{7-1} = 17.34 \text{ min}^2. \tag{4.11}$$

The corresponding standard deviation is $s = \sqrt{17.34} = 4.14$ min.

Based on the above, and the knowledge that the travelling time is subject to many small uncertainties, assume that it is *normally* distributed, with

$$\mu = 29.14 \text{ min}, \quad \sigma = 4.14 \text{ min.} \tag{4.12}$$

It is convenient to work in terms of the standardized normal variable,

$$\xi = \frac{x-\mu}{\sigma} = \frac{x-29.14}{4.14}, \tag{4.13}$$

as shown in Fig. 6. For a *tail* area of 1% = 0.01, observe from Table 2 that the value of the standardized normal variable having has a 0.01 probability of being exceeded is $\xi = 2.329$. Therefore, to incur this risk, you start at

$$x = \mu + \xi\sigma = 29.14 + 2.329 \times 4.14 = 38.78 \text{ min} \tag{4.14}$$

before the scheduled class time.

For part (b), which asks for the probability of arriving at least 12 minutes before class, i.e., at $x = 38.78 - 12 = 26.78$ (or earlier), the corresponding value of ξ is

$$\xi = \frac{x-\mu}{\sigma} = \frac{26.78-29.14}{4.14} = -0.57, \tag{4.15}$$

for which [from Table 1, also invoking Eqn. (4.9) because ξ is negative] the fractional area in the tail is 0.285.

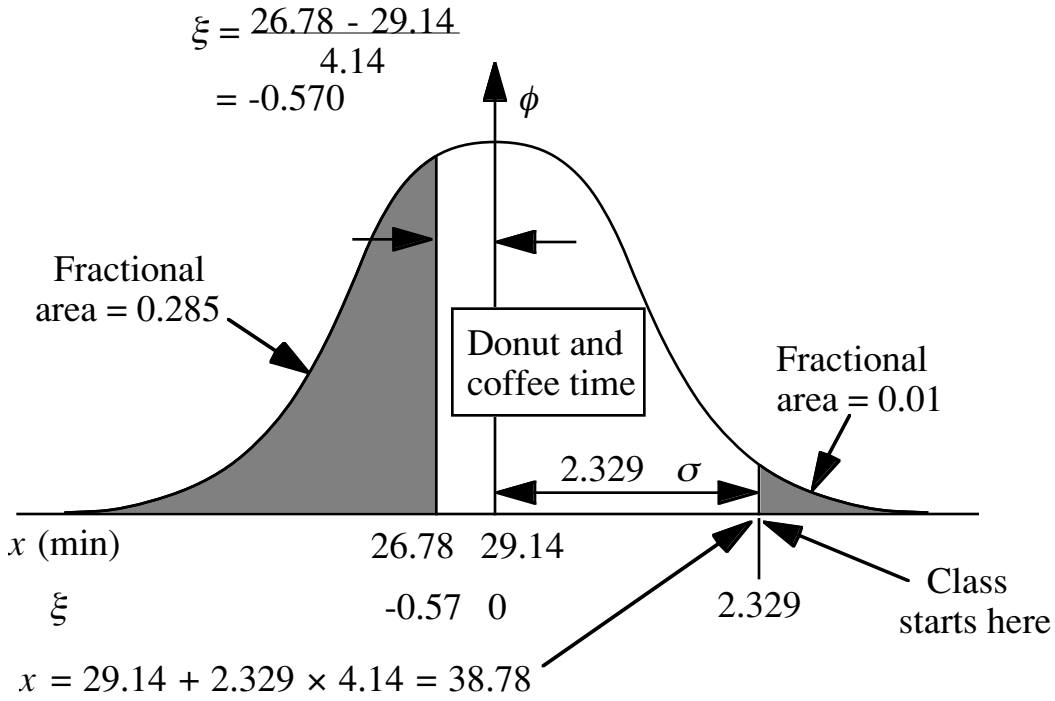

Fig. 6 Arrival-time frequency function.

5. Student *t* Distribution

In statistics, we often wish to test the hypothesis that a sample could conceivably have been taken from a certain population. One way of testing is to see if the sample mean $\bar{x}$ is realistically close (in terms of the standard deviation) to the population mean μ. Two different situations arise, depending on whether or not the population standard deviation σ is known *a priori*, or has to be estimated from the sample standard deviation *s*.

Hypothesis A: "A sample, whose mean is $\bar{x}$, comes from a normal distribution N(μ, σ) with mean μ and *known* variance σ2."

In such an event, we first note from Eqn. (3.3) that the sample mean has a variance σ2/*n* and hence a standard deviation $\sigma/\sqrt{n}$. The hypothesis is then *tested* by computing the N(0,1) variable

$$\xi = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{\sigma} \tag{5.1}$$

and checking it against tabulated values. The hypothesis is rejected if ξ is unduly large.

Hypothesis B: "A sample , whose mean is $\bar{x}$, comes from a normal distribution N(μ, s) with mean μ and variance σ2, the latter of which is *estimated* as being the same as the sample variance *s*2."

Table 3 Percentile Values (t_p) for Student's t Distribution for Various Degrees of Freedom

ν	$t_{.55}$	$t_{.60}$	$t_{.70}$	$t_{.75}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.158	.325	.727	1.000	1.376	3.08	6.31	12.71	31.82	63.66
2	.142	.289	.617	.816	1.061	1.89	2.92	4.30	6.96	9.92
3	.137	.277	.584	.765	.978	1.64	2.35	8.18	4.54	5.84
4	.134	.271	.569	.741	.941	1.53	2.13	2.78	3.75	4.60
5	.132	.267	.559	.727	.920	1.48	2.12	2.57	3.36	4.03
6	.131	.265	.553	.718	.906	1.44	1.94	2.45	3.14	3.71
7	.130	.263	.549	.711	.896	1.42	1.90	2.36	3.00	3.50
8	.130	.262	.546	.706	.889	1.40	1.86	2.31	2.90	3.36
9	.129	.261	.543	.703	.883	1.38	1.83	2.26	2.82	3.25
10	.129	.260	.542	.700	.879	1.37	1.81	2.23	2.76	3.17
11	.129	.260	.540	.697	.876	1.36	1.80	2.20	2.72	3.11
12	.128	.259	.539	.695	.873	1.36	1.78	2.18	2.68	3.06
13	.128	.259	.538	.694	.870	1.35	1.77	2.16	2.65	3.01
14	.128	.258	.537	.692	.868	1.34	1.76	2.14	2.62	2.98
15	.128	.258	.536	.691	.866	1.34	1.75	2.13	2.60	2.95
16	.128	.258	.535	.690	.865	1.34	1.75	2.12	2.58	2.92
17	.128	.257	.534	.689	.863	1.33	1.74	2.11	2.57	2.90
18	.127	.257	.534	.688	.862	1.33	1.73	2.10	2.55	2.88
19	.127	.257	.533	.688	.861	1.33	1.73	2.09	2.54	2.86
20	.127	.257	.533	.687	.860	1.32	1.72	2.09	2.53	2.84
21	.127	.257	.532	.686	.859	1.32	1.72	2.08	2.52	2.83
22	.127	.256	.532	.696	.858	1.32	1.72	2.07	2.51	2.82
23	.127	.256	.532	.685	.858	1.32	1.71	2.07	2.50	2.81
24	.127	.256	.531	.685	.857	1.32	1.71	2.06	2.49	2.80
25	.127	.256	.531	.684	.856	1.32	1.71	2.06	2.48	2.79
26	.127	.256	.531	.684	.856	1.32	1.71	2.06	2.48	2.78
27	.127	.256	.531	.684	.855	1.31	1.70	2.05	2.47	2.77
28	.127	.256	.530	.683	.855	1.31	1.70	2.05	2.47	2.76
29	.127	.256	.530	.683	.854	1.31	1.70	2.04	2.46	2.76
30	.127	.256	.530	.683	.854	1.31	1.70	2.04	2.46	2.75
40	.126	.255	.529	.681	.851	1.30	1.68	2.02	2.42	2.70
60	.126	.254	.527	.679	.848	1.30	1.67	2.00	2.39	2.66
120	.126	.254	.526	.677	.845	1.29	1.66	1.98	2.36	2.62
∞	.126	.253	.524	.674	.842	1.28	1.665	1.96	2.33	2.58

In this case, the reader is asked to accept that we can now compute a new variable, called "Student's *t*'":

$$t = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{s}, \tag{5.2}$$

and check it against tabulated values of the *t* distribution with ν = n - 1 degrees of freedom, for which the frequency function is

$$f(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \tag{5.2}$$

in which Γ is the gamma function. Note that the number of degrees of freedom ν is one less than the number n in the sample, essentially because the standard deviation has had to be estimated from the sample. It may be shown that t, illustrated in Fig. 7, is very similar to the N(0,1) variable ξ for all but small values of ν.

The probability points of Student's t distribution are shown in Table 3, where t has a probability P of falling below the tabulated t_P.

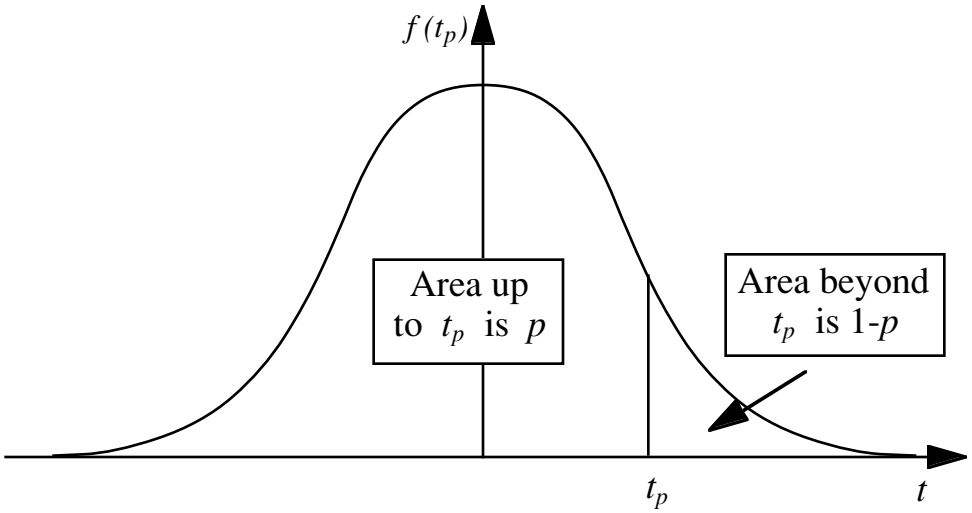

Fig. 7 Student's t distribution.

The following pressures (psig) were found in a random sample of $n = 9$ tires taken from a fleet of delivery vehicles:

32, 26, 32, 33, 25, 27, 31, 28, and 34

The pressures are normally distributed, and the person in charge asserts that they have a mean value of 32. Check this assertion.

Solution

For the sample of nine, the mean and standard deviation are, by calculator:

$$n = 9, \quad \bar{x} = 29.78, \quad s = 3.31. \tag{5.3}$$

Hence,

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s} = \frac{(29.78 - 32)\sqrt{9}}{3.31} = -2.01. \tag{5.4}$$

Note that Table 3 is only for *positive* values of *t*, from which we first find by interpolation that *t* based on eight degrees of freedom has about a P = 0.958 probability of falling below +2.01, and hence a 4.2% probability of falling above +2.01. Since the distribution is symmetric, *t* also has a 4.2% probability of being as negative as -2.01 (or more so). Thus, the observed value of *t* could have occurred about once in every 24 such samples, so we are naturally suspicious about the assertion that the tire pressures are being properly maintained at 32 psig, and would want to institute some sort of quality-control program.

Corresponding to various probability levels, Table 4 gives a verbal equivalent that could be useful when interpreting statistical evidence to the layman. In the case of the tires, the evidence is "probably significant."

Table 4 interpretation of Probabilities

P	*Interpretation*
0.05	Probably significant
0.01	Significant
0.001	Highly significant

6. Uncertainty Analysis

Suppose we have an explicit equation for the dependency of y , which is a function of n independent variables $\$x_i$, i = 1, 2, …

$$y = y\left(x_1, x_2, ..., x_n\right) \tag{6.1}$$

For example, y could be the overall heat-transfer coefficient in the ChE 360 heat-exchanger experiment, and the individual x's could represent the two water flow rates and four temperatures from which it is computed. Then, if the "uncertainties" in the x_i are known or can be estimated (as standard deviations σ_i, for example), *uncertainty analysis* then predicts the corresponding uncertainty in y (σ_y, for example). Thus, a knowledge of the uncertainties in the water flow rates and temperatures can be translated into the corresponding uncertainty for the overall heat-transfer coefficient.

Each of the variables x_i will have the following statistical parameters:

$$\begin{aligned} &\text{Mean:} && E\left(x_i\right) = \mu_i, \\ &\text{Variance:} && E\left(x_i + \mu_i\right)^2 = \sigma_i^2. \end{aligned} \tag{6.2}$$

It is important to note that the x_i are no longer individual observations of the same random variable, but now represent *different* variables.

Taylor's expansion of y about its value $y\left(\mu_1, \mu_2, ..., \mu_n\right)$ —when the x_i assume their mean values—gives

$$y = y\left(\mu_1, \mu_2, ..., \mu_n\right) + \left(x_1 - \mu_1\right)\frac{\partial y}{\partial x_1}$$

$$+\left(x_2-\mu_2\right)\frac{\partial y}{\partial x_2}+\ldots+\left(x_n-\mu_n\right)\frac{\partial y}{\partial x_n}. \tag{6.3}$$

Equation (6.3) says that y deviates from its central value by amounts that are proportional to the product of the deviations of the individual x_i from their means and the rate at which y changes with each individual x_i. That is,

$$y = y\left(\mu_1,\mu_2,\ldots,\mu_n\right)+\sum_{i=1}^{n}\left(x_i-\mu_i\right)\frac{\partial y}{\partial x_i}, \tag{6.4}$$

in which all the derivatives are evaluated at $\left(\mu_1,\mu_2,\ldots,\mu_n\right)$. Noting that these derivatives are constant, Eqn. (6.4) is a linear combination of the variables $x_i-\mu_i$, with the extra constant $y\left(\mu_1,\mu_2,\ldots,\mu_n\right)$ up front. Therefore, according to Eqn. (2.16), since the mean value of each of the variables $x_i-\mu_i$ is zero, the mean of y is

$$\mu_y = E(y) = y\left(\mu_1,\mu_2,\ldots,\mu_n\right). \tag{6.5}$$

The variance of y is also obtained by first noting that

$$\sigma_y^2 = E\left(y-\mu_y\right)^2 = E\left[\sum_{i=1}^{n}\left(x_i-\mu_i\right)\frac{\partial y}{\partial x_i}\right]^2. \tag{6.6}$$

But $\partial y/\partial x_i$ in Eqn. (6.6) is equivalent to a_i in Eqn. (2.17). Hence, Eqn. (2.19) immediately gives the variance of y as

$$\sigma_y^2 = \sum_{i=1}^{n}\sigma_i^2\left(\frac{\partial y}{\partial x_i}\right)^2. \tag{6.7}$$

Thus, the *variances* of the x_i are *additive*, and are weighted by $(\partial y/\partial x_i)^2$ in each case in order to give the variance of y.

Eqn. (6.7) may be recast in fractional form by dividing by y and rearranging:

$$\left(\frac{\sigma_y}{y}\right)^2 = \sum_{i=1}^{n}\left(\underbrace{\frac{x_i}{y}\frac{\partial y}{\partial x_i}}_{m_i}\right)^2\frac{\sigma_i^2}{x_i^2} = \sum_{i=1}^{n} m_i^2\left(\frac{\sigma_i}{x_i}\right)^2, \tag{6.8a}$$

in which

$$m_i = \frac{x_i}{y}\frac{\partial y}{\partial x_i}. \tag{6.8b}$$

Or, since the uncertainty in x_i is $\Delta x_i = a\sigma_i$ (some arbitrary multiple of its standard deviation, as explained in Section 7), and $\Delta y = a\sigma_y$ (note, all multiples must be the same):

$$\left(\frac{\Delta y}{y}\right)^2 = \sum_{i=1}^{n} m_i^2 \left(\frac{\Delta x_i}{x_i}\right)^2 .$$

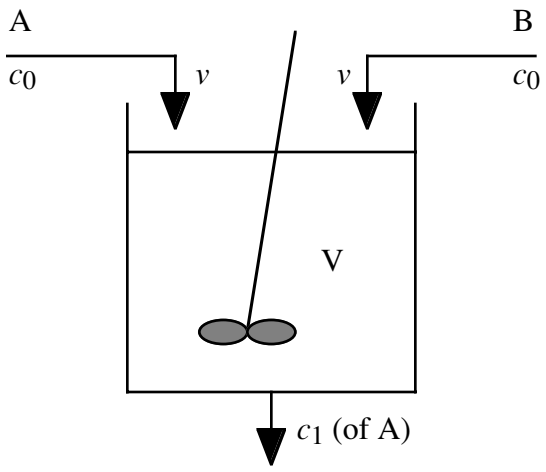

Fig.8 Continuous stirred tank reactor.

Suppose the following reaction is conducted in the CSTR (continuous stirred-tank reactor) shown in Fig. 8:

$$A + B \longrightarrow C \tag{6.10}$$

A molal rate balance for A gives

$$kVc_1^2 = \upsilon\left(c_0 - c_1\right). \tag{6.11}$$

Thus, the reaction rate constant is

$$k = \frac{\upsilon\left(c_0 - c_1\right)}{Vc_1^2} \tag{6.12}$$

whose form should be compared with Eqn. (6.1).

The experimental values are shown in Table 4, together with the estimated absolute and fractional uncertainties. In two cases, for c_0 and c_1, the right-pointing arrow indicates that the absolute uncertainty Δx_i is specified, and that the fractional uncertainty $\Delta x_i/x_i$ has been derived from it. The left-pointing arrows for v and V indicate the reverse.

Variable		*Value*	Units	Δx_i		$\Delta x_i/x_i$
x_1	υ	30	liter/min	1.5	←	0.050
x_2	c_0	1.5	g mole/liter	0.02	→	0.0133
x_3	c_1	0.5	g mole/liter	0.02	→	0.0400
x_4	V	1,500	liter	30	←	0.020
y	k	0.08	liter /g mole min	?		?

The values of the error multipliers,

$$m_i = \frac{x_i}{y}\frac{\partial y}{\partial x_i}, \tag{6.13}$$

are next determined. We have, after appropriate differentiation and insertion of values from Table 4:

$x_1 = \upsilon$:

$$m_1 = \upsilon \frac{Vc_1^2}{\upsilon\left(c_0 - c_1\right)} \times \frac{\left(c_0 - c_1\right)}{Vc_1^2} = 1. \tag{6.14}$$

The value $m_1 = 1$ indicates that the fractional change in k is directly proportional to the fractional change in v.

$x_2 = c_0$:

$$m_2 = c_0 \frac{Vc_1^2}{\upsilon\left(c_0 - c_1\right)} \times \frac{\upsilon}{Vc_1^2} = \frac{1}{1 - \dfrac{c_1}{c_0}} = 1.5. \tag{6.15}$$

The situation is now somewhat more complicated, since there is no longer a direct proportionality with c_0 because is appears in the difference c_0 - c_1.

$x_3 = c_1$:

$$m_3 = c_1 \frac{Vc_1^2}{\upsilon\left(c_0 - c_1\right)}\left(-\frac{2c_0}{c_1^3} + \frac{1}{c_1^2}\right)\frac{\upsilon}{V} = \frac{-2c_0 + c_1}{c_0 - c_1} = -2.5. \tag{6.16}$$

This case is even more complicated, since c_1 appears in both the numerator and denominator of Eqn. (6.12).

$x_4 = V$:

$$m_4 = -V \frac{Vc_1^2}{\upsilon\left(c_0 - c_1\right)} \times \frac{\upsilon\left(c_0 - c_1\right)}{V^2 c_1^2} = -1. \tag{6.17}$$

m_4 = -1 signifies that k is directly inversely proportional to V.

We can now complete the uncertainty analysis table, shown in Table 5.

Table 5 Uncertainty Analysis Table

Source	x_i	m_i	Δx_i		$\frac{\Delta x_i}{x_i}$	$10^4\left(m_i \frac{\Delta x_i}{x_i}\right)^2$	% Contr.
ν liter/min	30.0	1.0	1.50	←	0.0500	25.0	18.8
c_0 g mole/liter	105	1.5	0.02	→	0.0133	4.0	3.0
c_1 g mole/liter	0.5	-2.5	0.02	→	0.0400	100.0	75.2
V liter	1,500.0	-1.0	30.00	←	0.0200	4.0	3.0
k liter/g mole min	0.0800					133.0	100.0

It follows from Table 5 that

$$\left(\frac{\Delta k}{k}\right)^2 = 0.0133, \qquad \left(\frac{\Delta k}{k}\right) = 0.115 \tag{6.18}$$

Thus, since $k = 0.0800$, its uncertainty is

$$\Delta k = 0.115 \times 0.0800 = 0.0092, \tag{6.19}$$

and we can finally say that $k = 0.0800 \pm 0.0092$.

The idea is continued with several short example:

Example 1

$$\begin{aligned} y &= x_1 + x_2, \\ \sigma_y^2 &= \sigma_1^2 + \sigma_2^2, \\ (\Delta y)^2 &= (\Delta x_1)^2 + (\Delta x_2)^2. \end{aligned} \tag{6.20}$$

Example 2

$$\begin{aligned} y &= x_1 - x_2, \\ \sigma_y^2 &= \sigma_1^2 + \sigma_2^2, \\ (\Delta y)^2 &= (\Delta x_1)^2 + (\Delta x_2)^2. \end{aligned} \tag{6.21}$$

Example 3

$$\begin{aligned} y &= x_1^a x_2^b, \\ \ln y &= a \ln x_1 + b \ln x_2 \\ \frac{dy}{d} &= a\frac{dx_1}{x_1} + b\frac{dx_2}{x_2} \\ \left(\frac{\Delta y}{y}\right)^2 &= a^2\left(\frac{\Delta x_1}{x_1}\right)^2 + b^2\left(\frac{\Delta x_2}{x_2}\right)^2. \end{aligned} \tag{6.22}$$

Examples 4 and 5

From Example 3, with $a = 1$ and $b = \pm 1$,

$$y = x_1 x_2 \quad or \quad y = \frac{x_1}{x_2} \tag{6.23}$$

both yield

$$\left(\frac{\Delta y}{y}\right)^2 = \left(\frac{\Delta x_1}{x_1}\right)^2 + \left(\frac{\Delta x_2}{x_2}\right)^2 . \tag{6.24}$$

Case of implicit functional representation. Sometimes, we will *not* have an *explicit* formula for *y*. Instead, it will be incorporated *implicitly* into a functional relation as follows:

$$f(y, x_1, x_2, \ldots, x_n) = 0. \tag{6.25}$$

An example is given by van der Waals's equation,

$$f(V,P,T,a,b,R) = \left(P + \frac{a}{V^2}\right)(V - b) - RT = 0 \tag{6.26}$$

in which we can identify y = V, and the other five variables as x_1,..., x_5. Note that even though *a*, *b*, and *R* are "constants," they are still subject to uncertainties, even though these may be small in comparison with the uncertainties for the "physical" variables *P* and *T*.

To proceed, note that differential changes in y and the x_i gives rise to a corresponding change in f;

$$df = \frac{\partial f}{\partial y} dy + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \partial x_i = 0, \tag{6.27}$$

Hence, when all the *x*'s except x_i are constant,

$$df = \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial x_i} \partial x_i = 0, \tag{6.28}$$

so that

$$\frac{\partial f}{\partial x_i} = -\frac{\partial f / \partial x_i}{\partial f / \partial y}, \tag{6.28}$$

The analysis then proceeds as for the explicit case, but will usually be more involved, both algebraically and numerically, especially in computing *y*.

7. Methods for Estimating Uncertainty

The uncertainties, such as the Δx_i used in Section 6, may be estimated by one or more of the following ways:

1. By approximate guesses of "maximum probable uncertainty," for example:

   (a) When reading an instrument, it may be appropriate to take the uncertainty as a certain fraction (1%, for example) of the full scale. The value of the fraction is sometimes supplied with the instrument.

   (b) If an instrument is fluctuating somewhat during the course of an experiment, even though steady-state conditions are intended, then the uncertainty can be based on the range of the fluctuations.

2. By taking multiple measurements ("replicated data") of the same variable, under presumably unchanged conditions, in order to get the imprecision and hence the uncertainty.

3. By calculating the variances (and standard deviations) of distributed quantities, such as pellet lengths of particles in a packed bed.

4. By observing the "least count" or smallest division on a scale as the lower limit on readability.

5. By regression estimates of slopes and intercepts (see Section 8). We might also make "intelligent guesses" of the uncertainties in these quantities. For example, Fig. 9 shows confidence limits for the slope of a straight line that is used to fit a set of data points.

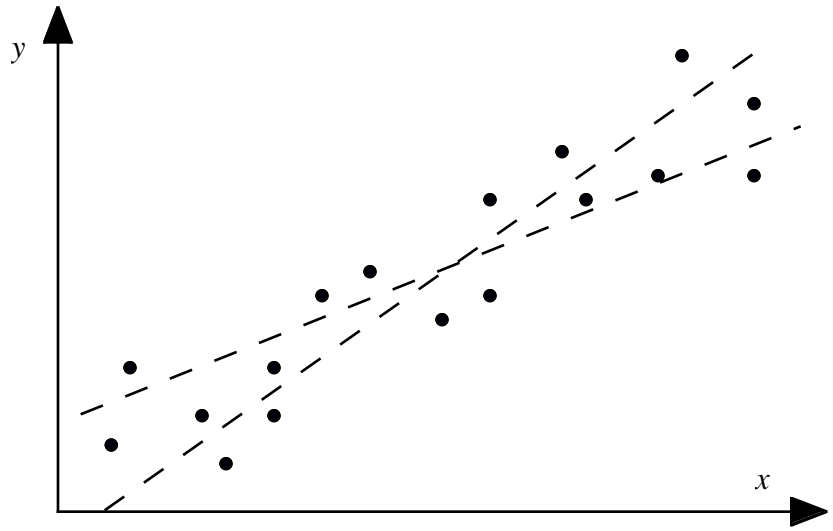

Fig. 9 Confidence interval for slopes.

We conclude this section by distinguishing between *accuracy*} and *precision*.

Accuracy is the degree to which an instrument registers reality in the absence of any errors in reading or using the instrument, and is often specified as either:

(a) Absolute accuracy, e.g., ± 0.1 gpm.
(b) Percent accuracy, e.g., 2% of full scale.

Precision} refers to the *reproducibility* of a measurement, which may be affected by factors such as:

(a) Parallax errors in reading a scale (such as that on the barometer in the ChE 360 laboratory) from slightly different viewpoints.

(b) Hysteresis effects such as might occur in mechanical linkages in a pressure gauge, when the same indicated value could correspond to slightly different values, depending on whether the pressure is increasing or decreasing.

(c) Disturbances or "noise."

(d) Least count, referring to the fineness of the scale of the instrument.

8. Linear Regression and the Method of Least Squares

In experimental work, we frequently plot one variable, such as y, against another, such as x, and then wish to develop an equation that expresses y as a function of x. The general situation is shown in Fig. 10, in which there are m sets of data points (x_i, y_i), i = 1, 2,..., m.. Our goal in this section is to represent the points by drawing the straight line $y=a + bx$ (also known as the *regression* line) "through" them.

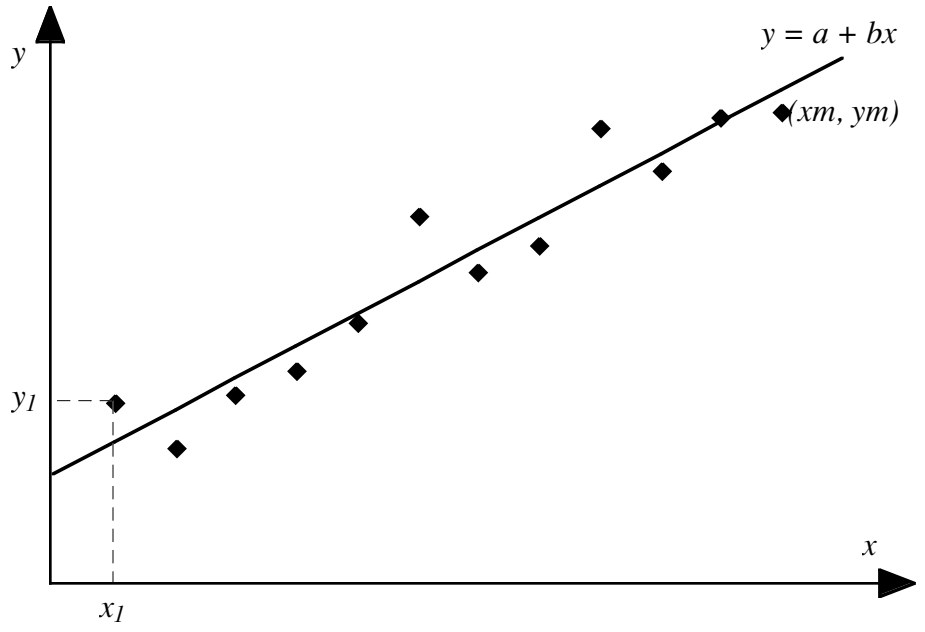

Fig. 10 Linear regression.

Thus, we wish to determine suitable values for the intercept a and the slope b, together with a value for the variance σ^2 that indicates how much scatter there is of the y values about the regression line.

The following theory is based on a model, shown in Fig. 11, in which we assume that the x_i are known precisely, whereas the y_i are distributed as

$$N(\underbrace{\alpha + \beta x_i}_{mean}, \sigma^2) \tag{8.1}$$

That is, the y values are normally distributed about the regressions line $y = \alpha + \beta x$ with (constant) variance σ^2. In practice, there will probably be imprecision in *both* the x and y values, in which case we choose x to be the more precisely known quantity. In a reaction for example, x might be the mass of catalyst (which can be weighed fairly precisely), and y could be the product yield (dependent on a chemical analysis, subject to random errors).

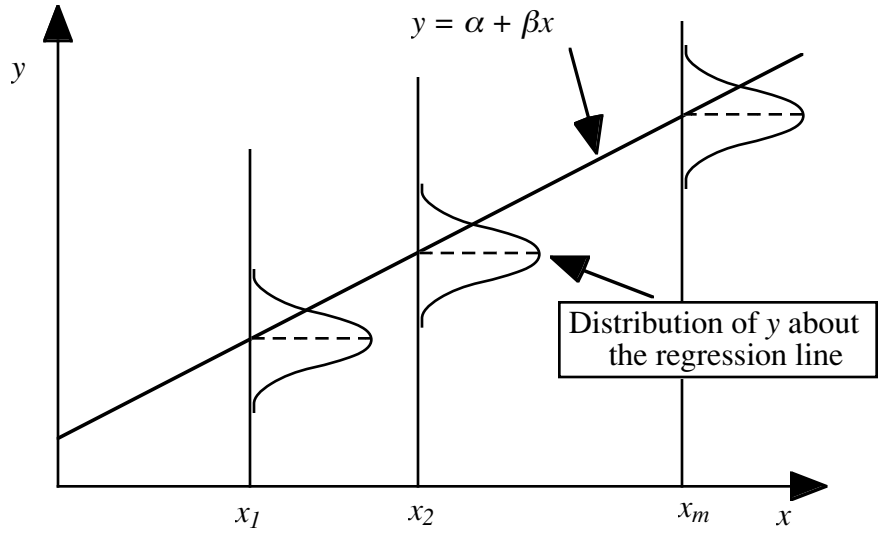

Fig. 11 Model for linear regression

Assuming that the measurement of y locates it within a small interval of extent Δy, the probability of observing the value y (an uncertainty interval), is, from the normal distribution frequency function:

$$P(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{(y_1 - \alpha - \beta x_1)^2 / 2\sigma^2} \Delta y = k e^{(y_1 - \alpha - \beta x_1)^2 / 2\sigma^2}, \tag{8.2}$$

where $k = \Delta y / \sigma\sqrt{2\pi}$ is a constant. Similar expressions hold for $P(y_2)$, ... , $P(y_m)$.

The probability of all these values occurring simultaneously is obtained by multiplying the individual probabilities:

$$P = P(y_1)P(y_2)....P(y_m) = k^m e^{-\sum_{i=1}^{m} (y_1 - \alpha - \beta x_1)^2 / 2\sigma^2}, \tag{8.3}$$

in which α and β are, so far, unknown. The regression line is then considered to be the best representation of the data points if α and β are chosen so that P is maximized.

P is clearly *maximized* when α and β are chosen so as to *minimize* the sum of squares of the deviations from the regression line:

$$S = \sum_{i=1}^{m} \left(y_i - \alpha - \beta x_i \right)^2 . \tag{8.4}$$

The sum *S* of squares of the deviations is minimized when its derivatives with respect to α and β are each equated to zero:

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{m} \left(y_i - \alpha - \beta x_i \right) = 0 ,$$
$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{m} x_i \left(y_i - \alpha - \beta x_i \right) = 0 . \tag{8.5}$$

Replacing the model parameters α and β by their *estimates a* and *b*, and rearranging, we obtain the simultaneous *normal equations* in *a* and *b*:

$$ma + b \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i ' \tag{8.6}$$
$$a \sum_{i=1}^{m} x_i + b \sum_{i=1}^{m} x_i^{\ 2} = \sum_{i=1}^{m} x_i y_i ' \tag{8.7}$$

Solution of the simultaneous normal equations, (8.6) and (8.7), gives:

Slope:

$$b = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m \sum_{i=1}^{m} x_i^{\ 2} - \left(\sum_{i=1}^{m} x_i \right)^2} = \frac{\sum_{i=1}^{m} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{\sum_{i=1}^{m} \left(x_i - \bar{x} \right)^2} = \frac{s_{xy}^2}{s_x^2} . \tag{8.8}$$

Here, s_{xy}^2 is the *covariance*} of *x* and *y*, and will be either +1 or -1 for a perfect correlation between *y* and *x*, zero for no correlation, and somewhere in between these extremes for most situations. The reader should check the alternative forms given within Eqn. (8.8).

Intercept:

$$a = \frac{1}{m} \left(\sum_{i=1}^{m} y_i - b \sum_{i=1}^{m} x_i \right) = \bar{y} - b\bar{x} . \tag{8.9}$$

It can also be shown that the estimate of the variance σ^2 of the *y* values about the regression line is:

$$s^2 = \frac{1}{m-2} \sum_{i=1}^{m} \left(y_i - a - b x_i \right)^2$$
$$= \frac{1}{m-2} \left(\sum_{i=1}^{m} y_i - a \sum_{i=1}^{m} y_i - b \sum_{i=1}^{m} x_i y_i \right) . \tag{8.10}$$

Note that the denominator contains m - 2 (and not m or m - 1) because *both a and b* have been estimated from data. (Accept this!)

We now recall formulas relating to the linear combination

$$b = a_1 y_1 + a_2 y_2 + \ldots + a_m y_m = \sum_{i=1}^{m} a_i y_i \,, \tag{8.11}$$

which is normally distributed if the y_i are normally distributed. From Eqns. (2.16) and (2.19), assuming the y_i are independent, the mean and variance of b are given by

$$\mu = \sum_{i=1}^{m} a_i \mu_i, \tag{8.12}$$

$$\sigma^2 = \sum_{i=1}^{m} a_i^2 \sigma_i^2 . \tag{8.13}$$

in which μ_i and σ_i^2 are the means and variances of the individual y_i.
Equations (8.12) and (8.13) will now be used to find the means and variances of the slope b and intercept a of the regression

Slope of regression line. From Eqn. (8.8), the slope can be reexpressed as

$$\begin{aligned} b &= \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right) y_i}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} \\ &= \frac{x_i - \bar{x}}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} y_1 + \ldots + \frac{x_m - \bar{x}}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} y_m, \end{aligned} \tag{8.14}$$

which is a linear combination of they_i with coefficients

$$a_i = \frac{x_m - \bar{x}}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}, \tag{8.15}$$

so that

$$\begin{aligned} E(b) &= \sum_{i=1}^{m} \frac{x_i - \bar{x}}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}\left(\alpha + \beta x_i\right) \\ &= \alpha \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} + \beta \frac{\sum_{i=1}^{m} x_i\left(x_i - \bar{x}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}. \end{aligned} \tag{8.16}$$

Note in developing Eqn. (8.16) that $\mu_i = E\left(y_i\right) = \alpha + \beta x_i$. The reader should also check that the coefficient of α is zero and that the coefficient of β is one, so that

$$E(b) = \beta \,, \tag{8.17}$$

thus verifying that the slope b given by Eqn. (8.8) is an *unbiased* estimate of the model slope β.

Similarly, since each y_i has variance σ^2,

$$\sigma_b^2 \,\mathrm{var}(b) = \sum_{i=1}^{m} \frac{\left(x_i - \bar{x}\right)^2}{\left[\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2\right]^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}. \tag{8.18}$$

With these results, it is easy to show that the random variable

$$t = \frac{b - \beta}{s}\sqrt{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2} \tag{8.19}$$

has the Student's t distribution with m - 2 degrees of freedom, which may then be used either:

(a) As a *test of significance* to see if β is likely to have some preconceived value.

(b) To establish a symmetric (or other) *confidence interval* for β.

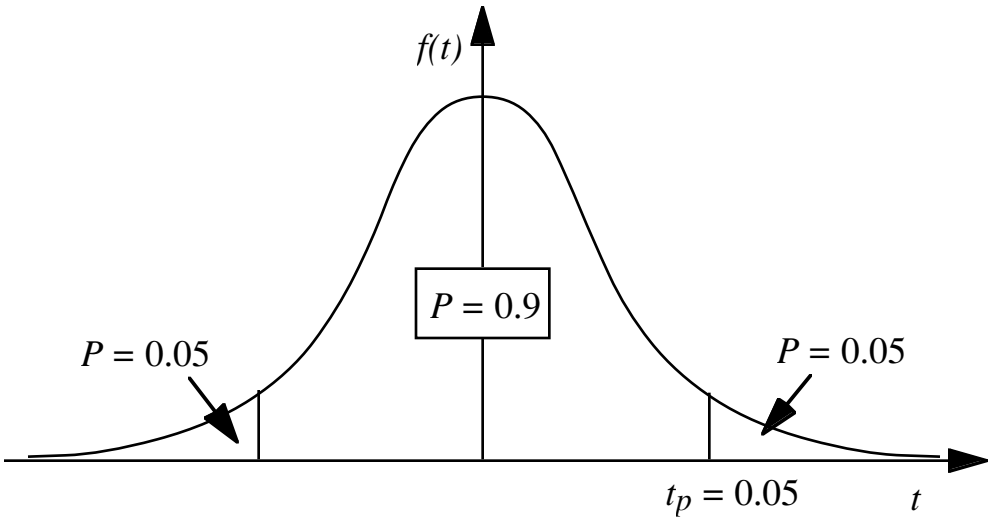

Fig. 12 P = 90% symmetric confidence interval.

For example, if we wished to establish a symmetric confidence interval for β, within which it would have a probability P of lying, we would use the formula

$$\beta = b \pm t_{(1+P)/2}\frac{s}{\sqrt{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}}. \tag{8.20}$$

The reader should check, by examining Figs. 7 and 12, and Table 3, why the subscript on t in Eqn. (8.20) is (1 + P)/2.

Intercept. It may likewise be that the mean and variance of the intercept a are:

$$E(a) = \alpha, \tag{8.21}$$

$$\sigma_a^2 = \text{var}(a) = \frac{\sigma^2}{m} + \bar{x}^2 \,\text{var}(b) = \frac{\sigma^2 \sum_{i=1}^{m} x_i^2}{m \sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2}. \tag{8.22}$$

Thus, a given by Eqn. (8.9) is again an unbiased estimate of the model parameter α.

Correlation Coefficient The following parameter, r^2, also called the square of the correlation coefficient, is often generated when curve-fitting (by CricketGraph, for example):

$$r^2 = \frac{\sum_{i=1}^{m} \left(a = bx_i - \bar{y}\right)^2}{\sum_{i=1}^{m} \left(y_i - \bar{y}\right)^2} \tag{8.23}$$

$$= \frac{\text{"Explained" variation}}{\text{Total variation}} = 1 - \frac{(m-2)s^2}{\sum_{i=1}^{m} \left(y_i - \bar{y}\right)^2} \tag{8.24}$$

For a perfect fit, depending on whether the slope was positive or negative, we would have:

$$r^2 = 1, \qquad r = \pm 1. \tag{8.25}$$

In practice, a seemingly high value such as $r^2 = 0.9$ does not indicate a very good fit, and values in the range 0.95-0.99 are much more desirable.

There also exists the possibility of fitting polynomials to values of y and x, as will probably already be familiar to users of Cricket Graph.

STATISTICS PROBLEMS

Note: Unless stated otherwise, you should work from first principles, not using the built-in statistical functions on your calculator.

1. Try and find the mean and standard deviation of the following on your calculator:

(a) The numbers 1, 2, and 3.
(b) The single number 1.

What happened in (b)? Why?

2. The frequency function for the distribution of residence times t in a CSTR with volume v and volumetric flow rate v is:

$$f(t) = \frac{v}{V} e^{-vt/V}. \tag{1}$$

Starting with equation (1), derive expressions for: (a) the mean residence time μ, and (b) the variance σ^2, as functions of V and v. If needed, you may assume the following formulas, which relate to integration by parts and the derivative of an exponential:

$$\int_a^b u \frac{dv}{dx} dx = uv\Big|_a^b - \int_a^b v \frac{du}{dx} dx, \qquad \frac{d}{dx} e^{-x} = -e^{-x}. \tag{2}$$

3. Evaluate the sample mean, variance, and standard deviation for the following values x_i, which are the diameters in cm of a random sample of $n = 8$ rods: 2.25, 2.31, 2.16, 2.29, 2.32, 2.19, 2.26, and 2.29.

Perform the calculations from first principles, making sure you work to four decimal places; then check your results against those given by the special-purpose mean- and standard-deviation functions on your calculator.

4. Based on the first harmonic and a period of 60 sec, the following sample of four values was reported by Baker and Cook (1990) for the thermal diffusivity (in^2/s) of a metal bar: 0.1878, 0.1809, 0.1824, and 0.1711.

Compute the mean $\bar{x}$ and standard deviation s for this sample. *Note:* you may have to use *eight* decimal places for adequate accuracy in the calculations! Are these values unbiased estimates of the corresponding population mean μ and standard deviation σ? In the long run, if further measurements are made, and are normally distributed, within what symmetric interval would you expect the values of the thermal diffusivity to lie 90% of the time?

5. Through extensive testing, a certain brand of light bulbs is known to have a life x in hours that is normally distributed with mean $\mu = 450$ hours and standard deviation $\sigma = 50$ hours. That is, x is an N(450, 50) random variable. Now please answer the following:

(a) What is the probability that any *one* bulb will last for at least 400 hours?
(b) What is the probability that any *five* bulbs will *all* last for at least 400 hours?
(c) What proportion of all the light bulbs can be expected to fail between 400 and 401 hours, inclusive. *Hint:* there is an easy way of answering this one—the interval between 400 and 401 can be treated as if it were differentially small!
(d) What bulb life will probably be exceeded by 10% of the bulbs?

6. Through extensive testing, a certain brand of light bulbs is known to have a life x in hours that is normally distributed with population mean $\mu = 450$ hours and variance $\sigma^2 = 2{,}500$ hours.

Four light bulbs are selected at random. What value will the mean life of the bulbs (taken as a group) have a 95% probability of exceeding?

7. The times taken to travel from your home to the ChE 360 class on $n = 7$ different days are: 30, 33, 26, 23, 30, 35, and 27 min.

(a) If you are willing to take a 1% chance of being late for class, how long before the class starts should you set out?

(b) In this event, what is the probability of your being able to buy a cup of coffee and a donut, which take 12 minutes to consume, before class starts?

8. Ten samples of waste water were tested for their pH value, and showed a sample mean of $\bar{x} = 5.64$ and a sample variance of $s^2 = 0.124$. The quality-control manager asserts that the waste water is being controlled at a pH value of 6.0. How do you feel about his assertion?

9. During the batch saponification between equimolar amounts of sodium hydroxide and ethyl acetate, the concentration c (gm moles/liter) varies with time t (min) according to the equation:

$$\frac{1}{c} = \frac{1}{c_0} + kt,$$

where c_0 is the initial concentration, and k (liter/gm mole min) is the reaction rate constant. The following results were obtained in the laboratory by Wilkes (1965), at a temperature of 77° F:

t:	1	2	3	4	5	7	10	12	15	20	25
1/c:	24.7	32.4	38.4	45.0	52.3	65.6	87.6	102	135	154	192

Now, obtain least-squares estimates of:

(a) The reaction-rate constant, k.
(b) The initial concentration, c_0.
(c) The standard deviation of k.
(d) A symmetric confidence interval for k, within which it has a 90% probability of lying.

(e) The correlation coefficient, r.

10. Buehler and Olsen (1989) obtained the following $m = 7$ data values relating to the phase angle θ (radians) as a function of distance x (in.) from the end of a rod in which transient heat conduction is occurring:

x:	0.0	0.5	1.0	2.0	4.0	6.0	10.0
θ:	2.657	2.471	2.182	1.686	0.644	-0.830	-2.931

In the absence of random errors, the data obey the equation

$$\theta = \gamma - \beta x . \tag{3}$$

Now, obtain least-squares estimates of:

(a) The slope $-\beta$ of the regression line, and hence of β.
(b) The standard deviation of β assuming the above data lead to an estimated value of 2.753 for γ.
(c) A symmetric confidence interval for β, within which it has an 80% probability of lying.

Based on the above data, the following summations may be assumed:

$\sum x_i$	$\sum x_i^2$	$\sum \theta_i$	$\sum \theta_i^2$	$\sum x_i \theta_i$	$\sum \left(x_i - \bar{x}\right)^2$
23.5	157.25	5.879	30.464	-24.925	78.36

11. The following data were among those data obtained by Jelic, Martin, and Thome (1988), for the diffusivity of toluene in air:

$z\,[t/(L - L_0 0,\ \text{sec/cm}]$	3,500	3,657	3,772	3,950	4,092
$x\,[(L - L_0 0,\ \text{cm}]$:	0.5	0.7	0.9	1.1	1.3

Perform a linear regression analysis to determine:

(a) An estimate b of the slope of the regression line (from which the diffusivity can be determined).
(b) The intercept a of the regression line at $x = 0$ (from which the initial length L_0 can be determined).
(c) Test the hypothesis that the actual slope is $\beta = 720$.

Hint; You may wish to make life a little easier by working with $y = z$ - 3,500 for most of the time. Under these circumstances, you may assume the following summations without calculation:

$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i^2$
4.5	1,471	4.45	1,619.3	651,597

12. The pressure drop Δp in a horizontal pipeline is given by

$$\Delta p = 2 f_F \rho u_m^2 \frac{L}{D},$$

and the volumetric flow rate is $Q = (\pi D^2 / 4) u_m$. The following values have been found experimentally (with uncertainties in parentheses): D = 1.049 in. (0.008); L = 30 ft (1.0); ρ = 49.5 lb_m/ft^3 (0.25); Q = 35.5 gpm (2.5), and Δp = 12.7 psi (0.3). (Of course, conversion factors would be needed when substituting these values into the above equation, but do you need to take these into account?) Calculate the corresponding value of the friction factor, and perform a complete uncertainty analysis.

13. The head increase Δh (in. of water) across a centrifugal pump is related to the volumetric flow rate Q (gpm) by the equation

$$\Delta h = \alpha - \beta Q^2,$$

in which α and β are constants in appropriate units.

If α = 111.2 (with fractional uncertainty 0.02), β = 2.24 (fractional uncertainty 0.04), and Q = 5.0 (*absolute* uncertainty 0.25), what is your estimate of Δh ? Construct a *complete* uncertainty analysis table to find both the fractional and absolute uncertainties in Δh .

14. An article on "The Short, Unhappy Life of Academic Presidents" appeared in the 25 July, 1990 edition of *The New York Times*. A 1986 survey to which 2,100 institutions responded reported that "the American Council on Education found that the average president had served for nearly seven years, but more than half had been in office for five years or less."

Sketch the general appearance of a plausible frequency function for the tenure of university presidents.

J.O. Wilkes, 1 January, 1995