C. Tolmie

REFIGURING ANTHROPOLOGY

First Principles Of Probability & Statistics

David Hurst Thomas
American Museum of Natural History

Waveland Press, Inc.
Prospect Heights, Illinois

For information about this book, write or call:

Waveland Press, Inc.
P.O. Box 400
Prospect Heights, Illinois 60070
(312) 634-0081

For permission to use copyrighted material, the author is indebted to the following:

FIG. 2.1. By permission of the Trustees of the British Museum (Natural History).
TABLE 2.3. (p. 24) *From Physical Anthropology: An Introduction* by A. J. Kelso. Reprinted by permission of the publisher, J. B. Lippincott Company. Copyright © 1974. (p. 25) Reproduced by permission of the Society for American Archaeology from *Memoirs of the Society for American Archaeology*, Vol. 11, 1956.
FIG. 3.1. From Hulse, Frederick S. *The Human Species: An Introduction to Physical Anthropology*. Copyright © 1963 by Random House, Inc.

FIG. 3.2. From Dozier, Edward P., *The Pueblo Indians of North America*. Copyright © 1970 by Holt, Rinehart and Winston, Inc. Reproduced by permission of Holt, Rinehart and Winston. [This book reissued 1983 by Waveland Press, Inc.]
FIG. 3.4. Reproduced by permission of the Society for American Archaeology from *American Antiquity*, Vol. 35 (4), 1970.
FIG. 3.5. Reproduced by permission of the American Anthropological Association from the *American Anthropologist*, Vol. 73 (3), 1971.
FIG. 13.14. From *Biometry* by Robert R. Sokal and F. James Rohlf, W. H. Freeman and Company. Copyright © 1969.

Second Printing
The 1976 version of this book was entitled *Figuring Anthropology*.

ISBN 0-88133-223-2

Printed in the United States of America.

10 The Student's *t*-Distribution

● *I ask for more information because I am unable to unscrew the unscrutable*.—S. Ervin

10.1 INTRODUCTION

Now we know how to translate research hypotheses into the language of statistical inference and how to test some of the more elementary propositions. The null hypothesis, you will remember, posits an expected value of some population parameter, and the alternative hypothesis covers the other potential values of that parameter. Once the variability of the sampling distribution is determined from the Central Limit Theorem, the *z*-value is compared with a predetermined level of statistical significance. In this manner decisions can be rendered regarding the credibility of null hypotheses based upon the sample at hand.

But here we encounter yet another snag. It seems that all the examples considered thus far have assumed that σ is known. At the time, of course, I did not explain just how we came to know σ; I just stated an arbitrary value. Unfortunately, such is rarely the case in actual research, and we must now face the practical difficulty of modifying what we have already learned to account for a more realistic application. Specifically, we must assess the impact of substituting S for σ in computing the standardized normal deviate z.

10.2 THE *t*-DISTRIBUTION

Chapter 8 established that the *standard error* is really just the standard deviation of the sampling distribution of sample means:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

So $\sigma_{\bar{x}}$ applies to the theoretically infinite population of sample means. But when σ is unknown, then the standard error must be estimated from the sample data at hand. The best *estimate of the standard error of the mean* is

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} \tag{10.1}$$

$S_{\bar{x}}$ is generally called simply the *standard error of the mean*, but keep in mind that $S_{\bar{x}}$ functions as a statistic whose job is to estimate $\sigma_{\bar{x}}$ (and, by extension, σ). The standard error tells us that any difference between the population mean and the sample mean drawn from the population is an "error" which has been spawned by the vicissitudes of sampling. Were there no errors, then all the $\bar{X}_i$ would be identical, regardless of how many samples had been drawn. In this case, the standard error would drop to zero. But a standard error of zero is impossible for any real run of data because of the omnipresent errors of sampling.

$S_{\bar{x}}$, therefore, estimates $\sigma_{\bar{x}}$ when σ is unknown. We know that the quantity $z = (\bar{X} - \mu)/\sigma_{\bar{x}}$ is a random variable with a mean of zero and a variance of 1. If the X_i are normally distributed, these quantities are exact; otherwise, the result is only approximate. Table A.3 (Appendix) provides the probabilities associated with various areas contained under this z-distribution. What effect does the substitution of $S_{\bar{x}}$ for $\sigma_{\bar{x}}$ have upon the distribution of z?

In the early days of statistical theory, the estimated standard error of the mean, $S_{\bar{x}}$, was simply substituted forthwith into the formula for z, as though no estimation was involved at all. But we now know that substituting $S_{\bar{x}}$ for $\sigma_{\bar{x}}$ produces a different, rather distinctive, mathematical entity called t:

$$t = \frac{\bar{X} - \mu}{S_{\bar{x}}} \tag{10.2}$$

Despite the superficial similarity of z and t, a couple of critical differences distinguish the behavior of each.

The numerator of the familiar z-score depends upon two quantities: The sample mean and the population mean. Sample means are always random variables, but the population mean is a parameter and hence is constant for a given population. Thus, for a particular population, the numerator of z depends strictly upon $\bar{X}$, and μ is constant. The denominator of z is also invariant because $\sigma_{\bar{x}}$ is constant for a sample size n. The specific value of any z depends strictly upon the value of the sample mean.

The distribution of the t-ratio is more complex. As with z, the numerator of t is a random variable, dependent upon $\bar{X}$. But unlike z, the denominator of t is not constant; $S_{\bar{x}}$ is a statistic varying from sample to sample. The value of any particular t depends upon *both* the sample mean and the sample variance. The t-ratio has become a function of sample size, since $S_{\bar{x}} = S/\sqrt{n}$, and herein lies the salient difference between z and t. If the same population were repeatedly sampled, a given value of $\bar{X}_i$ would always produce exactly the same value of z. But any given value of $\bar{X}_i$ can produce widely different values of t because $S_{\bar{x}}$ is computed from the specific sample at hand.

The t-ratio is thus more variable than z, and the extreme variates create longer tails for a t-distribution than for a normal distribution. The probability distribu-

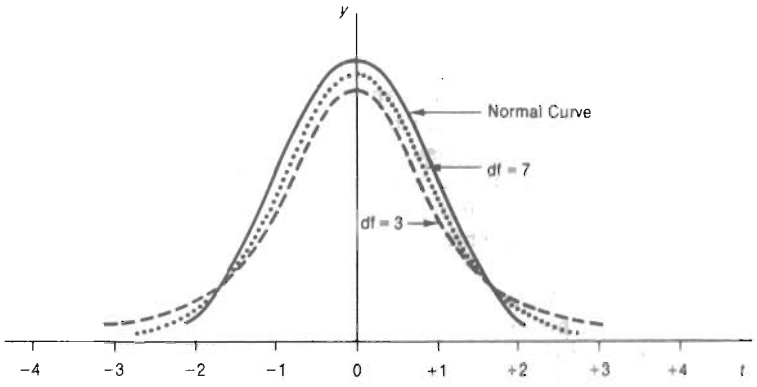

Fig. 10.1 Comparison of two *t*-distributions with the normal curve (after Alder and Roessler 1972: 156).

tion function of *t* becomes flatter than the normal curve, especially when small values of *n* are involved (see Fig. 10.1); the smaller the sample size, the "flatter" becomes the *t*-distribution relative to the normal curve. Conversely, as *n* increases, the distribution of *t* tends toward normality. In fact, for samples of size $n = 30$ and larger, the normal distribution and the *t*-distribution are virtually identical.

The problem of defining a probability distribution for *t* when σ is unknown remained a puzzle throughout the nineteenth century, despite other notable advances in statistical theory. The precise mathematical distribution of *t* was finally established by William S. Gosset, a statistician employed by the Guinness brewery in Dublin. The Guinness people had a strict rule prohibiting their employees from publishing their discoveries, but due to the importance of Gosset's computation, the company granted him the "privilege" of publishing his findings, provided he remain anonymous.

Gosset published his classic paper "The probable error of the mean" in 1908 under the pseudonym of "Student," and many feel that this single article laid the foundation for modern statistical theory. Curiously, the name "Student" has remained permanently affixed to the *t*-distribution even though Gosset's real name was publicly released shortly thereafter. Gosset's mathematical findings are beyond the present scope,[1] but his derivation of the equation for *t* allowed others to tabulate the various probabilities contained under the probability distribution of *t* (see Table A.4). The *t*-tables are quite simple to operate and allow ready computation for practical research problems when σ is unknown.

Assigning a probability figure to *t* requires only two simple quantities: the level of significance and the sample size. Probability values ranging from $\alpha = 0.450$ to $\alpha = 0.005$ are listed across the top of Table A.4. This table has been constructed for testing two-tailed hypotheses, so each probability includes the area under *both* tails of the *t*-distribution. The appropriate significance level for

[1]The equation for the Student's *t*-distribution is given by Mood and Graybill (1963: 233) and discussed in detail by Hays (1973: 392–399).

a one-tailed case is found by consulting the figure listed under $p = 2\alpha$. If a one-tailed hypothesis were to be tested at the 0.01 level, the appropriate column of Table A.4 is $p = 2\alpha = 2(0.01) = 0.02$.

The sample size is also necessary to enter Table A.4, but note that the rows are labelled "df" rather than the familiar n. The abbreviation df stands for *degrees of freedom*, a most important statistical concept. For now, we must settle for a relatively general explanation of this concept. The number of degrees of freedom in a sample is the number of freely varying quantities. Suppose you wished to find four integers which sum to 20:

$$a + b + c + d = 20$$

You could assign any possible value for any three digits; say, a, b, and c. But because the sum must equal 20, the last digit to be selected (d in this case) is *not* free to vary. The value of the final digit is "fixed," predetermined, because there is only a single value which will still produce a sum of 20. Suppose you selected the following integers:

$$a = 42 \qquad b = 26 \qquad c = -96$$

The d can take only one possible value: $d = -48$. The total number of integers involved in this example is $n = 4$. But the *total number of independent choices* is only $(n - 1) = 3$. The number of independent choices is termed the number of *degrees of freedom*. We have lost one degree of freedom by imposing the condition that the numbers must sum to 20. The number of degrees of freedom are given by n minus the number of conditions imposed upon the variates.

Although you may not have realized it, a condition has been imposed on the sample being tested against the t-distribution: The sample variates must have a mean of $\bar{X}$. Every sample has exactly n variates, but only $(n - 1)$ of these variates are free to vary independently of one another. The last value is predetermined by the equation $\bar{X} = \Sigma X_i / n$. *Hence, for the t-distribution, the number of degrees of freedom is always* $(n - 1)$.

10.3 COMPARING A SAMPLE TO A POPULATION WHEN σ IS UNKNOWN

We now possess a distribution which facilitates hypothesis testing, regardless of whether or not σ is unknown. We can now approach realistic data without making unrealistic assumptions. The t-test is not without assumptions, of course, and these are discussed in Section 10.9.

One common application of the t-distribution often involves comparing a sample mean with some parametric mean. For two-tailed testing, the statistical hypotheses are

$$H_o: \ \mu = A \qquad H_1: \ \mu \neq A$$

where A is some hypothetical value. The directional versions of these hypotheses are

$$H_o: \ \mu \geq A \qquad (\text{or } \mu \leq A)$$
$$H_1: \ \mu < A \qquad (\text{or } \mu > A)$$

The constant A is zero in many cases, but A is free to assume any a priori value.

The equation for t has already been introduced as Formula (10.2)

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} \quad \text{with df} = (n - 1)$$

where $S_{\bar{X}} = S/\sqrt{n}$. A couple of simple examples should suffice to illustrate this application of the t-distribution.

Example 10.1

Consider the following generalization: Hunter-gatherers tend to have an average population density of about 10 square miles per person. Test this hypothesis upon the following data taken from Steward (1938: 48–49) for the Northern Paiute of the western United States.

Owens Valley	2.1 square miles per person
Deep Springs	10.7
Fish Lake Valley	9.9
Saline Valley	16.6
Death Valley	30.0

Let μ be the population mean for the continuous random variable "population density." The statistical hypotheses are

H_o: $\mu = 10$ square miles per person

H_1: $\mu \neq 10$ square miles per person

Because σ, the standard deviation of the random variable X, is unknown, the t-test must be used instead of the familiar model of the normal distribution. This is a two-tailed test, with $\alpha = 0.05$. With df $= 5 - 1 = 4$, the boundary of the critical region for t is $t_{0.05} = 2.776$. The critical region itself is actually a set of t-values such that $|t| > 2.776$.

The sample size is $n = 5$ and the descriptive statistics are

$$\bar{X} = \frac{69.3}{5} = 13.86 \text{ square miles per person}$$

$$S = \sqrt{\frac{431.97}{4}} = \sqrt{107.99}$$

The standard error of the mean is estimated by

$$S_{\bar{X}} = \frac{\sqrt{107.97}}{\sqrt{5}} = 4.65 \text{ square miles per person}$$

The t-ratio in this case is

$$t = \frac{13.86 - 10.0}{4.65} = 0.83$$

Since this computed t-statistic does not fall within the critical region, H_o is not rejected. We conclude that the Northern Paiute data are consistent with the generalization that hunter-gatherers have an average population density of 10 square miles per person.

Suppose that one erroneously applied the normal distribution model instead. The sample standard deviation must be taken to estimate σ, despite the small sample size of $n = 5$:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{13.86 - 10.0}{10.39/\sqrt{5}} = 0.83$$

We find the associated probability to be $p = 0.4066$. Since we know that the true probability ($t = 0.83$) is about 0.44, the incorrect application of the normal distribution model would cause us to underestimate the true probability.

Example 10.2

Clovis projectile points tend, on the average, to be about 7.5 cm long (Wormington 1957: 263). The Lehner Ranch site in southern Arizona yielded 13 projectile points associated with the butchered remains of mammoth, horse, bison, and tapir. The excavators (Haury, Sayles, and Wasley 1959: table 1) list the following length measurements for these artifacts (numbers in parentheses are estimates):

Point	Length, cm	Point	Length, cm
1	(8.7)	8	4.7
2	7.9	9	5.6
3	8.3	10	3.1
4	7.4	11	7.8
5	(3.6)	12	9.7
6	6.2	13	5.2
7	8.1		

Since the average length of these 13 points is less than 6.7 cm, are these points significantly shorter than most Clovis points (at the 0.05 level)?

Let μ be the population mean of the random variable "total length." σ is unknown.

Statistical hypotheses:

$$H_o: \quad \mu \geq 7.5 \text{ cm}$$
$$H_1: \quad \mu < 7.5 \text{ cm}$$

Critical region: With $df = (13 - 1) = 12$ and a one-tailed test:

$$t_{2\alpha} = t_{0.10} = 1.782$$

Sample statistics:

$$\bar{X} = \frac{86.3}{13} = 6.64 \text{ cm}$$

$$S = 2.06 \text{ cm}$$

$$S_{\bar{X}} = \frac{2.06}{\sqrt{13}} = 0.57 \text{ cm}$$

t-ratio:

$$t = \frac{6.64 - 7.5}{0.57} = -1.51$$

Statistical decision: Since the computed value does not fall within the region of rejection, H_o is not rejected.

Research decision: The sample of 13 projectile points from the Lehner Ranch are not significantly shorter than typical Clovis points.

As an aside, let us examine the computation of the sample standard deviation. The above value of S was computed by Formula (4.13):

$$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{50.88}{12}} = 2.06 \text{ cm}$$

But suppose that the formula with divisor of n had erroneously been applied. Then

$$\hat{S} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}} = \sqrt{\frac{50.88}{13}} = 1.98 \text{ cm}$$

This value of S can readily be corrected by using the correction factor presented earlier:

$$S = \sqrt{\frac{n}{n-1}}\hat{S}$$

$$= \sqrt{\frac{13}{12}}(1.98) = 2.06 \text{ cm}$$

This correction is often necessary when dealing with standard deviations computed on a computer.

10.4 CONFIDENCE INTERVALS FOR μ WHEN σ IS UNKNOWN

The earlier discussion of confidence intervals for the population mean (Section 8.5) assumed that σ was known. Substituting $S_{\bar{X}}$ for $\sigma_{\bar{X}}$ vitiates use of the normal distribution, upon which Expression (8.6) was based.

The *t*-distribution allows us to derive a new expression which is applicable even though σ is unknown. Solving Expression (10.2) for μ, we find the

confidence limits for μ when σ is unknown to be

$$\mu = \bar{X} \pm tS_{\bar{X}} \tag{10.3}$$

The *t*-distribution is symmetrical, so the confidence limits fall equidistant from $\bar{X}$. Confidence intervals are computed in a manner identical to those of the normal distribution except that the tabled value of t is substituted for z and $S_{\bar{X}}$ is involved rather than $\sigma_{\bar{X}}$.

Example 10.3

Find the 99 percent confidence limits for the 13 Clovis points from the Lehner Ranch site (Example 10.2).

The appropriate value of t with 12 degrees of freedom is $t_{0.01} = 3.055$. Substituting into Expression (10.3):

$$\text{Confidence limits} = 6.64 \pm 3.055\left(\frac{2.06}{\sqrt{13}}\right) \text{cm}$$

$$= 6.64 \pm 1.74 \text{ cm}$$

We conclude with 99 percent confidence that the true parametric length of these Clovis points lies between 4.90 and 8.38 cm.

Example 10.4

Five skulls were excavated from a large Pleistocene cave in mainland China:

Skull	Cranial Capacity, cc
1	1225
2	1135
3	1055
4	1225
5	1030

Find the 95 percent confidence interval for this population.

We compute $\bar{X} = 1134.0$ cc with $S = 91.68$ cc. The appropriate value of t with four degrees of freedom is found from Table A.4 to be $t_{0.05} = 2.776$. Substituting into Equation (10.3):

$$\text{Confidence limits} = 1134.0 \pm 2.776\left(\frac{91.68}{\sqrt{5}}\right) \text{cc}$$

$$= 1134.0 \pm 113.63 \text{ cc}$$

We can conclude that the probability is 0.95 that the true population mean lies between 1020.4 and 1247.7 cc.

10.5 COMPARING TWO SAMPLE MEANS WHEN σ IS UNKNOWN

Section 10.3 presented a method for evaluating statistical hypotheses about the mean of a single population. Some value for μ (which we called A) was compared to the sample value. But sometimes it is impossible to frame a hypothesis specific enough to predict exact values for μ. Anthropologists are often interested in *comparisons* which evaluate differences between two sample means without reference to specific values for μ. Not only are comparative studies important in themselves in anthropology, but relational hypotheses (as opposed to absolute hypotheses) also sidestep the task of predicting exact values for μ. It is almost impossible, for example, to predict an absolute value for the cranial capacity of a sample of *Australopithecus* skulls. But it is a relatively easy matter to predict that *Australopithecus* skulls should have a smaller cranial capacity than a sample of Neanderthal skulls. Similarly, we can guess that hunters such as the Eskimo will have a higher per capita protein intake than a largely plant-gathering group such as the Western Shoshoni, even though the precise value of the protein intake for either group is unknown.

Only rarely will a sample mean ever exactly equal the population mean. This is because of sampling error. Similarly, two populations with identical means ($\mu_1 = \mu_2$) will almost always yield samples with different means ($\bar{X}_1 \neq \bar{X}_2$), once again because of sampling error. So the question must arise when comparing two sample means as to whether the difference is due to a real difference between the populations or whether the disparity between the samples should be attributed to chance alone.

Consider the archaeologist attempting to infer prehistoric population dynamics from a settlement pattern survey. He might suspect that one plant community supported a denser population than did the adjacent biotic community, even though he cannot accurately predict the population densities of two areas. Valley floor biota might, for instance, be expected to support a greater population density than the neighboring hilly mountain flanks. Accepting the "number of rooms per building" in an archaeological site as an operational indicator of population density, the hypotheses would appear

$$H_0: \quad \mu_X \leq \mu_Y \qquad H_1: \quad \mu_X > \mu_Y$$

We are predicting an ordinal relationship ("greater than") rather than a metric hypothesis ("*how much* greater than"). Suppose that a sample of nine contemporary sites were excavated to test this proposition:

Valley Floor, Site	No. of Rooms
1	9
2	10
3	7
4	10

$\bar{X} = 9.0$ rooms per site; $S_x = 1.41$ rooms per site.

Mountain Flanks, Site	No. of Rooms
1	6
2	5
3	7
4	4
5	5

$\bar{Y} = 5.4$ rooms per site; $S_y = 1.14$ rooms per site.

The descriptive statistics tell us that the valley floor sites tend to have more rooms than sites on the mountain flanks ($\bar{X} = 9.0 > \bar{Y} = 5.4$). The research hypothesis would appear to be correct (at least the direction is right). But the size difference is not overwhelming and might well be due to mere sampling error rather than a true difference in site size.

That is, we must consider the *standard error* of this difference because the larger the standard error, the less chance there is of a true population difference. But if the standard error is relatively small, the *population* of valley floor sites probably has more rooms than the mountain sites, as suggested in the research hypothesis.

This situation is analogous to that encountered in Chapter 8, when two samples were compared. If the population variances were known, then Expression (8.5) could have been used to determine the standardized normal deviate. But as in so many problems of this sort, we must deal with the results at hand.

The first difficulty is to estimate these unknown population variances. We must assume that the two populations have identical variances. By so doing, we can argue that any discrepancy between the samples relates only to differences in central tendency rather than differences in *shape* of the distribution of variates about the mean.

All statistical estimates improve as n increases, so the best possible estimate of either population variance will include the relevant variates. There are two distinct samples involved here, but because we assume the population variances to be equal, we can combine the deviations about the respective sample standard deviation. The individual variances are *pooled* into one single, best estimate of population variance:

$$S_p = \sqrt{\frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2}{n_x + n_y - 2}} \tag{10.4}$$

This new expression is called the *pooled estimate of the standard deviation*. S_p combines the total amount of deviation about $\bar{X}$ in the first sample with the amount of deviation about $\bar{Y}$ in the second sample and then averages this by dividing by the combined number of degrees of freedom. Two degrees of freedom are lost because two independent means were computed. S_p is an unbiased estimator of σ only as long as the individual population variances are assumed to be equal.

Let us see how the pooled estimate of σ works on the archaeological data at hand. Each sample has a known standard deviation: S_x estimates the variability

in the number of rooms per site on the valley floor (σ_X) and S_Y estimates the variability in the mountain sites (σ_Y). By assuming that $\sigma_X = \sigma_Y$, S_X and S_Y are combined (*pooled*) into a single estimator, S_p.

$$S_p = \sqrt{\frac{6.00 + 5.20}{4 + 5 - 2}} = \sqrt{1.60} = 1.26$$

Thus, S_p estimates that the population standard deviation of rooms per site on the valley floor—and by assumption, also on the mountain slopes—is $\sigma_X = \sigma_Y = 1.26$ rooms per site.

With this new estimate of total variability firmly in hand, it becomes possible to define an appropriate expression of the *t*-ratio to test for a difference between two samples:

$$t = \frac{(\bar{X} - \bar{Y}) - \mu_{\bar{X}-\bar{Y}}}{S_{\bar{X}-\bar{Y}}} \tag{10.5}$$

where $df = n_X + n_Y - 2$. In this expression,

$$\mu_{\bar{X}-\bar{Y}} = \mu_X - \mu_Y$$

and

$$S_{\bar{X}-\bar{Y}} = \sqrt{\frac{S_p^2}{n_X} + \frac{S_p^2}{n_Y}}$$

Note that S_p^2/n_X corresponds to $S_{\bar{X}}^2$. The general configuration of the *t*-ratio remains as before. A parametric mean (in this case $\mu_{\bar{X}-\bar{Y}}$) is subtracted from the sample estimate of this mean, $\bar{X} - \bar{Y}$, and is then divided by an estimate of the standard error of the difference between the sample means ($S_{\bar{X}-\bar{Y}}$).

We can now statistically assess the difference between two small sample means. $S_{\bar{X}-\bar{Y}}$ is found in the archaeological example to be

$$S_{\bar{X}} = \sqrt{\frac{1.60}{4}} = \sqrt{0.40}; \qquad S_{\bar{Y}} = \sqrt{\frac{1.60}{5}} = \sqrt{0.32}$$

$$S_{\bar{X}-\bar{Y}} = \sqrt{0.40 + 0.32} = \sqrt{0.72} = 0.85 \text{ room per site}$$

Note there is no need to take the square root when computing $S_{\bar{X}}$ and $S_{\bar{Y}}$. The radicals will automatically cancel when $S_{\bar{X}}$ and $S_{\bar{Y}}$ are substituted into $S_{\bar{X}-\bar{Y}}$.

The value of *t* in the example is

$$t = \frac{3.6 - 0}{0.85} = 4.24$$

with $df = 4 + 5 - 2 = 7$. This observed *t* is highly signficant since $t_{0.02} = 2.998$ with 7 degrees of freedom. Hence, the archaeological samples allow rejection of H_0, and we may justifiably conclude that valley sites tend to have more rooms than do the mountain sites. (Whether the index of "rooms per archaeological site" is a relevant indicator of prehistoric population density, of course, remains an archaeological rather than a statistical matter.)

An understandable degree of confusion can arise from the several variance

estimates involved in comparing two samples. To summarize:

S_X = standard deviation of sample X (estimates σ_X).
S_Y = standard deviation of sample Y (estimates σ_Y).
S_P = pooled standard deviation of both sample X and Y (best estimate of both σ_X and σ_Y, which are assumed to be equal).
$S_{\bar{X}} = S_P/\sqrt{n_X}$ is the standard error of sample X.
$S_{\bar{Y}} = S_P/\sqrt{n_Y}$ is the standard error of sample Y.
$S_{\bar{X}-\bar{Y}}$ = standard error of the difference between the two sample means (estimates $\sigma_{\bar{X}-\bar{Y}}$).

Example 10.5

A paradox in the evolution of culture is how consistently man's technological advances seem to backfire; Marvin Harris (1971: 216) has called such advances the "labor-saving devices that increase work." It can be said, for example, that advanced agricultural techniques have *increased* (rather than decreased) the per capita amount of work required for survival. To test this hypothesis, fieldwork was carried out among the Bushmen (a hunter-gatherer people) and a group of West African subsistence farmers. This sample of 26 Bushmen indicated that each works an average of 805 hours per year, with $S = 10.3$. The 16 West Africans in the sample spent an average of 820 hours per year, with $S = 12.9$ hours. Do these results support the hypothesis that hunter-gatherer groups tend to work less than agriculturalists (at the 0.01 level)?

Let us term the Bushmen as group X and the farmers as group Y.

Statistical hypotheses:

$$H_o: \ \mu_X \geq \mu_Y \qquad H_1: \ \mu_X < \mu_Y$$

Region of rejection: For a one-tailed test with $(26 + 16 - 2) = 40$ degrees of freedom, $t_{0.02} = 2.423$.

We know the two sample standard deviations, $S_X = 10.3$ and $S_Y = 12.9$, so it is necessary to work back to find the sum of the squared deviations:

$$S_X = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$$

$$\Sigma(X_i - \bar{X})^2 = S_X^2(n_X - 1) = (10.3)^2(25) = 2652.25$$
$$\Sigma(Y_i - \bar{Y})^2 = S_Y^2(n_Y - 1) = (12.9)^2(15) = 2496.15$$

The pooled estimate of the standard deviation is

$$S_P = \sqrt{\frac{2652.25 + 2496.15}{40}} = 11.35$$

The standard error of the difference is

$$S_{\bar{X}-\bar{Y}} = \sqrt{\frac{128.71}{26} + \frac{128.71}{16}} = \sqrt{4.95 + 8.04}$$

$$= \sqrt{12.99} = 3.60$$

The *t*-ratio is found to be

$$t = \frac{(805 - 820) - 0}{3.60} = -4.17$$

Since $|t| = 4.17 > t_{\alpha=.02} = 2.423$, the results are judged to be statistically significant and H_0 is rejected. These two samples lead us to conclude that Bushmen seem to work significantly less than West African agriculturalists. Further generalization—to all hunter-gatherers and agriculturalists—becomes an anthropological rather than a statistical matter.

Example 10.6

In Example 10.1, a sample of five Northern Paiute bands were found to have an average population density of $\bar{X} = 13.86$ square miles per person, with $S_X = 10.39$. The following sample of 11 Western Shoshoni bands (the Northern Paiute and Western Shoshoni are neighbors in the Great Basin) shows an average population density of $\bar{Y} = 7.91$ square miles per person. Can the Western Shoshoni be said to have a higher population density than the Northern Paiute at the 0.05 level (data from Steward 1938: 48–49)?

Band	Population Density, square miles per person
Reese River	3.6
Railroad Valley	9.0
Antelope Valley	11.0
Gosiute	12.5
Diamond Valley	3.8
Ruby Valley	2.8
Palisade	3.3
Halleck	4.0
Battle Mountain	2.5
Kawich	17.0
Little Smoky Valley	17.5

Statistical hypotheses:

$$H_0: \ \mu_X \leq \mu_Y \qquad H_1: \ \mu_X > \mu_Y$$

Region of rejection: For a one-tailed test with $(5 + 11 - 2) = 14$ degrees of freedom, $t_{0.10} = 1.761$.

To find $S_{\bar{x}-\bar{y}}$, we must first find S_p, the pooled estimate:

$$S_p = \sqrt{\frac{431.8 - 334.1}{5 + 11 - 2}} = \sqrt{54.71}$$

$$S_{\bar{x}-\bar{y}} = \sqrt{\frac{54.71}{5} + \frac{54.71}{11}} = \sqrt{15.91} = 3.99$$

Note that the sample standard deviation per se (S_x and S_y) is not needed in finding $S_{\bar{x}-\bar{y}}$, since $\Sigma(X_i - \bar{X})^2$ and $\Sigma(Y_i - \bar{Y})^2$ are the appropriate terms for finding S_p, which in turn is substituted into $S_{\bar{x}-\bar{y}}$.

The t-ratio is found to be

$$t = \frac{(13.86 - 7.91)}{3.99} = 1.49$$

The computed value of t far exceeds the critical region, so H_o is not rejected. On the basis of the two samples at hand, Western Shoshoni can not be said to have a higher population density than Northern Paiute. Another way of stating this conclusion is that Northern Paiute and Western Shoshoni samples appear to have been selected from the same statistical population.

10.6 COMPARING A SINGLE VARIATE TO A SAMPLE

The following formula can be used to determine the probability that a single isolated variate belongs to the same population as a given sample:

$$t = \frac{(\bar{X} - X_i)\sqrt{n/(n+1)}}{S_x} \tag{10.6}$$

where S_x is the standard deviation of the sample. The number of degrees of freedom are equal to $df = (n - 1)$. This formula is derived from a simplification of Expression (10.5) which compared the means of two independent samples: One "sample" in this case consists of a single variate. Note that had two "samples," each containing only a single variate, been compared, then $df = n_x + n_y - 2 = 0$. Two isolated variates cannot be compared.

Example 10.7

Paleoanthropologist Bryan Patterson found a fragment of human mandible at Kangatotha, west of Lake Rudolf, Kenya. A radiocarbon analysis determined a probable age of 2835 B.C. ± 100. The crown area of M_1 on the Kangatotha mandible is 139.2 mm^2 (data from Coon 1971b: table 2). By contrast, Shaw measured a series of 73 South African Bantu informants and found the crown area of M_1 to be only 115.5 mm^2 (assume S = 11.0 mm). Is the Kangatotha molar too large to be Bantu at the 0.01 level of significance?

Statistical hypotheses:

H_o: $\mu \leq 115.5$ mm H_1: $\mu > 115.5$ mm

Region of rejection: For a one-tailed test with $\alpha = 0.01$, and df $= 73 - 1 = 72$, $t_{0.02} = 2.390$. (The tabled value for 60 df is sufficient in this case.)

The *t*-ratio is computed to be

$$t = \frac{(115.5 - 139.2)\sqrt{73/74}}{11.0} = -2.14$$

This value of t is less than the critical value, so H_o is not rejected. Thus, the crown area of the Kangatotha molar is not significantly different from the Bantu sample. They could represent the same statistical population. You should note, however, that this conclusion does *not* have taxonomic implications.

10.7 SPECIAL CASE: STATISTICAL INFERENCE IN RADIOCARBON DATING

Radiocarbon dates are the final product of a fascinating collaboration between nuclear physicists, statisticians, and archaeologists. Radioactive decay is a random process. Beta emissions are produced as C^{14} atoms decay to N^{14} (nitrogen), and these emissions can be detected by sensitive Geiger counters. The underlying principle of this complex technique is simple—the fewer emissions, the older the carbon. Although the average number of emissions can be predicted over a given span of time, nobody can ever predict precisely which atoms will decay at any particular time. The radiocarbon laboratory employs Geiger counters to measure the number of beta particles emitted over a 1000-minute interval. Because radioactive decay is a random process, the sample variability must be taken into account, and samples are always counted twice. If the counts from the two runs are in "statistical agreement," further counting is unnecessary.

The "radiocarbon date" itself consists of two parts, a mean and a standard deviation: $\bar{X} \pm S$. For example, an archaeologist might receive the following radiocarbon determination from the laboratory:

$$950 \pm 40 \text{ radiocarbon years B.P. (before present)}$$

In this case, $\bar{X} = 950$, which estimates the true age of the sample (μ). The degree of variability between counting runs is expressed by S, the sample standard deviation. The population standard deviation is unknown and estimated by S. The larger the S, the more variability was observed between counting runs, and the less reliable is $\bar{X}$ in estimating μ. From what we already know about the nature of the normal curve, this means that there is a 68.26 percent chance that the true age falls within the range of $\bar{X} \pm S$, that is, between 910 and 990 radiocarbon years ago. The average age of any sample is only an *estimate* of the true age, so the plus-minus factor should never be omitted from radiocarbon determinations.

An example should clarify these elementary statistical aspects of radiocarbon

dating. During the 600-year Classic period, the Maya erected carved stone monuments (*stelae*) bearing "Long Count" dates. The dates seem often to denote the date of dedication of a temple or other ceremonial structure, although the exact meaning of the inscriptions is still unknown. Mayan epigraphers have struggled for decades trying to correlate the Mayan Long Count system with the Christian calendar.[2] The search was finally narrowed to a series of discrete choices.

Any given *katun* (Maya period of 20 years, each consisting of 366 days) can recur in the Maya system only once every 260 years. As a result, scholars have correlated given Maya dates to several intervals along the Christian calendar, depending upon the zero point chosen for the Maya system. The Maya date 9.15.10.0.0 3 Ahau 8 Mol, for example, dedicated Temple IV at the Classic Maya site of Tikal, Guatemala. George Spinden correlated this date to August 29, A.D. 481. A second reckoning, the Goodman–Thompson–Martinez correlation, sets this same Maya dedicatory date exactly 13 *katuns* (260 years) later, at June 28, A.D. 741. A solid case was made for both correlations, based upon historic records of the Maya calendar, and a stalemate resulted. Fortunately, some of the inscriptions at Temple IV were upon wooden lintels, so the radiocarbon laboratory at the University of Pennsylvania ran a series of tests upon the lintel itself in an effort to resolve the correlation problem. The hope was that the C^{14} dates would correspond to one of the two likely correlations, setting the dispute to rest.

Although dozens of radiocarbon determinations were processed on the Tikal beams, consider for the moment the implications of a single date (from Satterthwaite and Ralph 1960: table 1).

Laboratory Number	Beam Number	Age, years B.P.	Age, years A.D.
P-236	Room 2, VB2	1262 ± 38	697 ± 38

Each "radiocarbon date" is assigned a laboratory number. If subsequent runs were made on the same sample, a new number would be assigned to keep the independent determinations separate. Date P-236 (the 236th determination run by the Pennsylvania Laboratory) has an average age of $\bar{X} = 697$ radiocarbon years,[3] with a sample standard deviation of $S = 38$ years. Remembering that $\bar{X}$ is only an *estimate* of the true age of the Tikal lintel, the standard deviation can be used to compute the same limits of confidence for the true age (μ). There is, for example, a 0.6826 probability that the true age lies between A.D. 659 and A.D. 735 (see Fig. 10.2). There is also a probability of 95 percent that the true age falls between A.D. 622 and A.D. 771 ($\bar{X} \pm 1.96S$). While these reliability estimates place the true age of the sample within a known error factor, the data do not directly tell us about the correlation problem.

By inspection, we can see that the sample mean of P-236 is 44 years younger

[2]See the discussion earlier in Section 2.4.3 for a consideration of this problem in terms of levels of measurement.

[3]By convention, all C^{14} dates are computed as years before 1950.

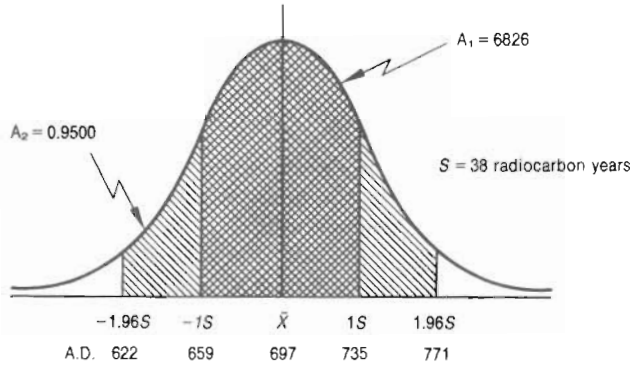

Fig. 10.2

than the Goodman–Thompson–Martinez correlation, but is 216 years too old for the date predicted from the Spinden correlation? Can we therefore say that date P-236 supports the Goodman–Thompson–Martinez correlation? Is the date close enough, or is the 54-year discrepancy too large a difference? Could the error of dating be so great as to support *both* correlations? Or does P-236 suggest that both correlations are wrong? Because radiocarbon dating is a random process, and because of the error introduced in the counting process itself, all radiocarbon dates involve such variability. The solution to the correlation problem will not be absolutely clear-cut. The answer must be expressed in terms of probability.

Figure 10.3 includes both correlation dates for the Maya calendar. We are now dealing with sample statistics (rather than population parameters), so the expressions on the normal curve are denoted by $\bar{X}$ rather than μ, as before. The point X_1 is A.D. 741, the date predicted by the Goodman–Thompson–Martinez correlation. The probability that the true age of sample P-236 is A.D. 741 or older corresponds to

$$z_1 = \frac{X_1 - \bar{X}}{S} = \frac{741 - 697}{38} = 1.16$$

$$p(X \geq 741) > 0.12$$

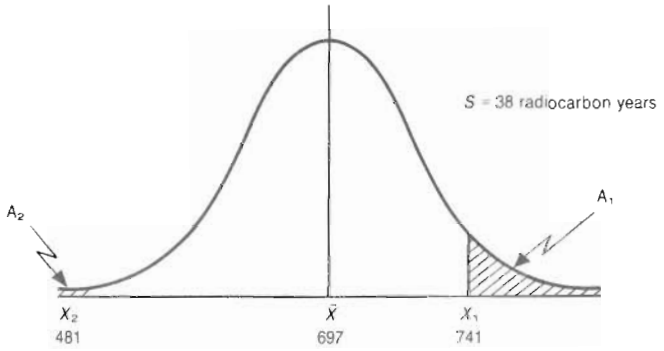

Fig. 10.3

The probability that sample P-236 dates the event to A.D. 481 or younger is

$$z_2 = \frac{X_2 - \bar{X}}{S} = \frac{481 - 697}{38} = -5.68$$

$$p(X \leq 481) < 0.0001$$

These results tell us that while the probability that P-236 actually dates the Spinden correlation is virtually nil, the chances of this single sample corresponding to the Goodman-Thompson-Martinez correlation are more than 12 percent. Taking only the results from P-236, the Spinden correlation seems to be eliminated. There remains a good chance that the Goodman–Thompson-Martinez correlation is correct.

But because of the randomness and uncertainty involved in C^{14} dating, archaeologists have learned never to trust a single radiocarbon determination. The large series of dates run for the Tikal lintels, for example, eventually confirmed the Goodman–Thompson–Martinez correlation by an overwhelming margin. The methods for comparing C^{14} dates to see whether they date a single episode will be considered later in this chapter.

● *What we seek in any realm of human thought is not absolute certainty, for that is denied us as men, but rather the more modest path of those who find dependable ways of discerning different degrees of probability.*—E. Trueblood

10.7.1 Computing the Radiocarbon Estimates

The radiocarbon age estimate—really a sample mean—is merely the adjusted mean of Geiger counts on ancient charcoal-bearing samples. But the statistical deviation is a more complex statistic, reflecting three major sources of variability: variation in

1. The ancient sample
2. Environmental radiation striking the Geiger counter
3. The known-age calibration sample.

Let us examine how these independent sources of variation are integrated into a single estimate of standard deviation. This discussion not only provides added insight into the workings of the radiocarbon method, but also furnishes an excellent opportunity to review the mechanics of computing (and combining) standard errors.

We begin from scratch by following an actual sample through the various manipulations involved in the radiocarbon process. The following data were obtained from vault beam 2, room 3 in Temple IV at Tikal (Satterthwaite and Ralph 1960: table 1):

(P-243) 1223 ± 46 radiocarbon years B.P. (before 1950)

P-243 can also be expressed as A.D. 727 ± 46 radiocarbon years. But this final age estimate results only after a series of laboratory and statistical manipulations.

Once the beam was removed from the temple, a small sample of zapote wood was cleaned manually to remove termite remains and then soaked in hydrochloric acid to dissolve inorganic carbon compounds. The sample was then placed in a combustion tube with pure oxygen gas and the mixture was ignited to convert the ancient solid carbon into carbon dioxide gas. This gas was filtered to remove contaminants, and then piped into a Geiger counter. We know that a beta particle is emitted each time a C^{14} atom decays back into N^{14}. The actual radiocarbon analysis counts the number of beta emissions—and by extension, the number of C^{14} decays—with a Geiger counter. The length of the counting interval depends both upon the material being dated and also the age of the sample; most laboratories count their samples overnight for a standard interval of 1000 minutes, and every sample is counted at least twice.

The laboratory worksheet for the Tikal date P-243 appears as follows:

Date Counted	Total Count, X_t
3/15/59	37,069
3/16/59	36,918

The total count, X_t, is the exact number of beta emissions recorded in a single 1000-minute counting run. P-243 was counted on both 15 and 16 March. As long as these two net counts are found to be in statistical agreement—by a chi-square test (discussed in Chapter 11)—no further counting runs are necessary. The average of the two total counts is

$$\bar{X}_t = \frac{37{,}069 + 36{,}918}{2} = 36{,}993 \text{ counts}$$

But the University of Pennsylvania radiocarbon laboratory, like every other place on this planet, is subject to atmospheric radioactivity which registers on laboratory Geiger counters along with the ancient sample emissions. The amount of this background radiation, called b, must be determined for each radiocarbon laboratory and then periodically rechecked for fluctuation. During March 1959, the University of Pennsylvania radiocarbon laboratory was bombarded by an average of $b = 9416$ radioactive emissions per 1000-minute counting interval. The net number of emissions, $\bar{X}_n$, from sample P-243 is thus found by subtracting b from each of the total counts. The average net count for both counting runs is

$$\bar{X}_n = \frac{(37{,}069 - 9416) + (36{,}918 - 9416)}{2}$$

$$= \frac{(27{,}653 + 27{,}502)}{2} = 27{,}578 \text{ counts}$$

So the average net count, $\bar{X}_n$, is actually a sample mean. More precisely, $\bar{X}_n$ is the mean number of beta emissions per 1000-minute counting interval. The standard error (the standard deviation of the mean) is given by

$$S_{\bar{x}} = \frac{\sqrt{(X_t + b)n}}{n}$$

where n is the number of counting runs.[4] This rather unusual expression is different from the previous standard errors we have encountered because radiocarbon emissions follow the *Poisson distribution*, a variant of the binomial distribution.

The standard error of the net counting rate for sample P-243 is

$$S_{\bar{x}} = \frac{\sqrt{(36{,}993 + 9416)2}}{2} = 152.3 \text{ counts per 1000 minutes}$$

The *net rate per minute*, called I, is then found by dividing by the standard counting interval, 1000 minutes:

$$I = \frac{27{,}578}{1000} = 27.578 \text{ counts per minute}$$

with a standard error of

$$S_I = \frac{S_{\bar{x}}}{1000 \text{ minutes}} = 0.152 \text{ counts per minute}$$

The net counting rate is then converted to an age estimate by comparing the amount of decay in the ancient sample relative to a modern sample. To find this relative amount of decay, it is necessary to know the existing radioactivity of modern samples. The University of Pennsylvania laboratory measured the beta emissions in a number of recent oak tree samples and found the average zero-age counting rate, I_0, to be

$$I_0 = 32.146 \pm 0.040 \text{ counts per minute}$$

The standard error of the difference between the average zero-age rate of emission (I_0) and the emission rate of the ancient Tikal sample (I) is found, as before, as the square root of the sum of the squared individual standard errors:

$$S_t = \sqrt{(0.152)^2 + (0.040)^2} = 0.157 \text{ counts per minute}$$

The value of S_t thus reflects the total combined variability due to fluctuations in (1) the ancient sample, (2) the background, and (3) the zero-age sample. The results of the radiocarbon analysis are hence summarized as

$$\bar{X}_n \pm S_t \text{ counts per minute}$$

Translating the figures from "counts per minute" to "radiocarbon years ago" is accomplished by substitution into the routine formula for age computation (based upon a half-life of 5568 years):

$$\text{absolute time} = \log (I_0/I) \times 18.5 \times 10^3$$

S_t is added to and subtracted from $\bar{X}_n$ to compute the range of one standard error from the mean. These "minimum" and "maximum" ages (±1 standard error) are substituted into the conversion formula:

[4]The symbolism employed here departs somewhat from that generally used by radiocarbon laboratories (for example, Ralph 1971), to remain consistent with the present discussion.

Maximum age:

$$\bar{X}_n - S_t = 27.578 - 0.157$$
$$= 27.421 \text{ counts per minute}$$

Maximum time:

$$\log\left(\frac{27.421}{32.146}\right) \times 18.5 \times 10^3 = \log(0.85301) \times 18.5 \times 10^3$$
$$= -1277 \text{ radiocarbon years}$$

Minimum age:

$$\bar{X}_n + S_t = 27.578 + 0.157 = 27.735$$

Minimum time:

$$\log\left(\frac{27.735}{32.146}\right) \times 18.5 \times 10^3 = -1186 \text{ radiocarbon years}$$

The average of the minimum and maximum ages of this sample provides the best estimate of the true age of the sample: $(1277 + 1186)/2 = 1232$ radiocarbon years ago. The standard error (expressed in years ago rather than in counts per minute) is found as simply half the difference between the "minimum" and "maximum" ages: $(1277 - 1186)/2 = 46$ radiocarbon years ago.

All that remains is to convert the date to "years before 1950." Since the counting runs took place in 1959, the date is converted to $1232 - 9 = 1223$. The final report from the radiocarbon laboratory is

(P-243) 1223 ± 46 radiocarbon years B.P.

Thus, the plus-minus factor appended to radiocarbon dates is really a standard error (the standard deviation of the sample mean).

But a couple of critical assumptions are necessary before the procedures of statistical inference can be applied to radiocarbon dates. We must initially assume that the large number of counts recorded on each run renders the distribution of means (or the distribution of the difference between sample means, if two dates are being compared) practically indistinguishable from that expected for a normally distributed population (Spaulding 1958). That is, the *t*-distribution with an infinitely large number of degrees of freedom is assumed to hold for radiocarbon determinations. We also assume that the rounding of published standard errors does not introduce any significant inaccuracy.

In addition, we are using the standard error derived from averaging the "maximum" and "minimum" ages when it is actually known that the true standard error (expressed in years) always has a plus error somewhat greater than the minus error. But for dates of moderate age, this discrepancy is not marked. For these reasons, comparing radiocarbon ages using the *t*-distribution is only an approximation which becomes less accurate as the age of the sample increases. When greater accuracy is required, it will be necessary to work with the actual counting runs rather than with the dates as expressed in absolute years (see Satterthwaite and Ralph 1960, for a more detailed discussion of these points).

10.7.2 Comparing a Radiocarbon Date to a Fixed Age

Let us return to the Maya Long Count problem. The *t*-distribution is useful in determining which, if any, of the standard correlations is consistent with the Tikal radiocarbon dates. Although only two of the correlations were mentioned earlier, the Temple IV dates at Tikal were actually tested against five different Maya-Christian correlations (Satterthwaite and Ralph 1960: tables 15 and 17).

Correlation	Estimated Age of Temple IV, Tikal
Spinden	A.D. 481
Dinsmoor	A.D. 504
Goodman–Thompson–Martinez	A.D. 741
Kreichgauer	A.D. 858
Escalona Ramon	A.D. 1001

We can see from inspection that P-243 (A.D. 727) is remarkably close to the Goodman–Thompson–Martinez (GTM) reckoning, but a test of statistical significance will show us just how close the GTM date and P-243 really are.

Statistical hypotheses[5]:

$$H_o: \quad \mu = \text{A.D. } 741 \qquad H_1: \quad \mu \neq \text{A.D. } 741$$

Region of rejection: For a two-tailed test at $\alpha = 0.05$ with df $= \infty$, $t_{0.05} = 1.96$.

Observed t-ratio:

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} = \frac{727 - 741}{46} = -0.30$$

The null hypothesis cannot be rejected in this case because $t = 0.30 < t_{0.05} = 1.96$. We conclude that Tikal sample P-243 is consistent with the Goodman–Thompson–Martinez hypothesis.

Note that this conclusion in no way *confirms* the GTM correlation because other correlations might also account for a C^{14} date of A.D. 727 at Tikal. Each of the other population correlations can be tested against P-243 in precisely the same manner:

$$H_o: \quad \mu = \text{A.D. } 481 \qquad t_{\text{Spinden}} = \frac{727 - 481}{46} = 5.35$$

$$H_o: \quad \mu = \text{A.D. } 504 \qquad t_{\text{Dinsmoor}} = \frac{727 - 504}{46} = 4.85$$

$$H_o: \quad \mu = \text{A.D. } 858 \qquad t_{\text{Kreichgauer}} = \frac{727 - 858}{46} = -2.85$$

$$H_o: \quad \mu = \text{A.D. } 1001 \qquad t_{\text{Escalona Ramon}} = \frac{727 - 1001}{46} = -5.96$$

[5]It matters little whether radiocarbon samples are expressed in years A.D., B.C., or years ago. Only the *difference* between the dates appears in the numerator of the *t*-ratio.

Every observed t falls well within the critical region and the null hypothesis for each of the four correlations must be rejected. We reject the Spinden, Dinsmoor, Kreichgauer, and Escalona Ramon correlations as untenable, in light of date P-243 from Tikal.[6]

Note how carefully both the statistical findings and the substantive implications have been expressed. Scientific theories such as these are never actually proved *correct*; practical research is directed only toward proving the competing theories *wrong*. Radiocarbon evidence from Tikal allows rejection of the four prevalent hypotheses competing with the Goodman–Thompson–Martinez correlation. But the GTM correlation has by no means been *proved correct*, since there could always be additional hypotheses which are likewise consistent with the C^{14} evidence. After a thorough and well-designed attempt at refutation such as this has failed, a theory can only *tentatively* be presumed to be correct. It can never be proved so (see Naroll and Cohen 1970: 26).

● *No study, whether a true experiment or not, ever proves a theory; it merely probes it.*—R. Winch and D. Campbell

Confidence Limits of a Radiocarbon Date Because there might be other hypotheses to explain the Tikal dates, a further step can be taken toward a final solution to the correlation problem by computing the limits within which other acceptable hypotheses must fall. The 95 percent confidence interval for date P-243 is

$$\mu = \bar{X} \pm t_{0.05} S_{\bar{X}}$$

$$\mu = \text{A.D. } 727 \pm 1.96\,(46)$$

$$\mu = \text{A.D. } 727 \pm 90.2 \text{ radiocarbon years}$$

Thus, at a 0.95 level of probability, any acceptable correlation must place the dedicatory date of Temple IV at Tikal no earlier than A.D. 637 and no later than A.D. 817. None of the seriously proposed correlations fall within this interval, so we are still left with a provisional acceptance of the Goodman–Thompson–Martinez reckoning. Note that computing the confidence interval is a superior method (in this case) of decision making.

Comparing Two Radiocarbon Dates Sometimes one needs to apply statistical logic inference when two radiocarbon dates are compared. Consider the dating of the Lehner Ranch site in southern Arizona, where Paleo-Indian artifacts were found in clear-cut association with the remains of nine butchered mammoths. A firehearth was discovered nearby and charcoal samples were submitted to the University of Arizona radiocarbon laboratory, with the following results:

(A-40a)	10,900 ± 450 years ago
(A-40b)	12,000 ± 450 years ago

[6] Of course no right-thinking archaeologist would rely upon a single radiocarbon date for so bold a conclusion; Satterthwaite and Ralph ran a total of ten C^{14} dates on beams and lintels from Temple IV alone.

The means of these two samples differ by some 1100 years, even though the charcoal came from a single firehearth. Does this 1100-year gap represent a true difference or can this discrepancy more readily be accounted for by statistical error?

Statistical hypotheses:

$$H_o: \quad \mu_{A\text{-}40a} = \mu_{A\text{-}40b}$$
$$H_1: \quad \mu_{A\text{-}40a} \neq \mu_{A\text{-}40b}$$

Region of rejection: For a two-tailed test at $\alpha = 0.05$, and with infinite degrees of freedom, $t_{0.05} = 1.96$.

The standard error of the difference between sample means is found as before:

$$S_{\bar{X}-\bar{Y}} = \sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2} = \sqrt{450^2 + 450^2} = 636 \text{ years}$$

The t-ratio is

$$t = \frac{(\bar{X} - \bar{Y}) - \mu_{\bar{X}-\bar{Y}}}{S_{\bar{X}-\bar{Y}}} = \frac{(10{,}900 - 12{,}000) - 0}{636} = -1.73$$

Since $t = 1.73 < t_{0.05} = 1.96$, H_o is not rejected, and we conclude that the difference between dates A-40a and A-40b is not significant. The two dates could well date a contemporary event at the Lehner Ranch site.

What should we conclude when a significant difference emerges, indicating that two dates are really "different?" Statistically, this decision tells us that the two radioactive samples have probably been selected from different statistical populations, but the archaeological ramifications are more difficult to assess. Archaeologists generally assume that, all else being equal, a difference in radiocarbon dates results from a true age difference between the samples. But this remains only an assumption because several other factors could cause contemporaneous samples to "date" differently: impure CO_2, radon in counter, electronic circuit breakdowns, Geiger counter failure, cosmic ray showers, even atmospheric fallout. In the Tikal study alone, Satterthwaite and Ralph rejected over 40 percent (34 of 83) of their counting runs as spurious. There is also the danger of contaminating the sample itself by sloppy excavation, by percolating groundwater, by rodent burrowing, by rootlets, or even by insects.

It is always possible to introduce significant error into the samples and hence create a spurious radiocarbon date. There are even cases when several dates on the same log have produced widely different age determinations, although the samples must be of exactly identical age. There seems to be many a "slip 'twixt the cup and the lip" in radiocarbon dating, and statistical inference establishes whether or not a significant discrepancy exists between dates. Only nonstatistical considerations can *explain* that discrepancy.

Problems may also arise when structuring the research hypotheses into statistical hypotheses. If the Lehner Ranch null hypothesis had been directional (one-tailed), the region of rejection would have been $t_{0.10} = 1.65$, and the observed difference between the dates would have been declared "significant." The two-tailed alternative was selected in this case because no prior hypotheses existed to suggest *which* sample should be older than the other. It simply turned

out that the determination for A-40b was older than that for A-40a. But had there been some specific reason to suggest that A-40b would be older *before the actual results were known*, then a one-tailed test would have been in order.

● *Ours is the age which is proud of machines that think, and suspicious of men who try to.*—H. Jones

10.8 THE CASE OF PAIRED VARIATES

The data considered thus far were purposely selected so that each variate was totally uneffected by the other sample variates. The assumption of independent variates follows from our earlier definition of random sampling. But there are some hypotheses of interest involving data which are not independent of one another; the variates are "paired" with each other. Consider the following hypotheses:

Right-handed individuals tend to have larger right arms than left arms.
First-born individuals are usually stronger than their second-born siblings.
Students are rarely smarter than their professors.
Wives tend to be more motivated than their husbands.

These variates are linked into naturally occurring dyads (right-left, male-female, older-younger), and such linkage vitiates any usage of the *t*-test discussed so far.

Pairing of variates has an importance far greater than simple convenience because pairing is a tactic in the general strategy of efficient research design. The idea behind a purposeful pairing of variates is to increase the basis of comparison on a desired effect. Extraneous factors ("noise") can sometimes produce a significant difference even when there is no difference resulting from the phenomenon under study. Conversely, these same extraneous factors can sometimes mask a true difference, resulting in an incorrect acceptance of the null hypothesis. Errors of this sort can never be totally eliminated, but cautious design of experiments can purge a great deal of noise from the data.

A basic rule in designing experiments is to control what can be controlled and to randomize the uncontrollable. Pairing controls extraneous factors by grouping variates which are alike in all respects save the condition under study. In learning experiments, for instance, subjects are often paired by IQ scores so that variable degrees of intelligence will not mask the actual rates of learning or retension. Pairs are also commonly constructed to control for bias by sex, age, generation, socioeconomic background, motivation, or achievement. Acculturation studies often involve the natural pairings produced in "before-after" observations. Such variates are termed *self-pairing* when a single variable is measured on two occasions under different conditions.

But the use of paired variates destroys the assumption of statistical independence and necessitates an alteration in *t*-testing methods. The following example illustrates this simple modification.

A controversial topic in anthropology has been the so-called nature-nurture problem: To what extent is behavior conditioned by environmental as opposed

The means of these two samples differ by some 1100 years, even though the charcoal came from a single firehearth. Does this 1100-year gap represent a true difference or can this discrepancy more readily be accounted for by statistical error?

Statistical hypotheses:

$$H_o: \quad \mu_{A\text{-}40a} = \mu_{A\text{-}40b}$$
$$H_1: \quad \mu_{A\text{-}40a} \neq \mu_{A\text{-}40b}$$

Region of rejection: For a two-tailed test at $\alpha = 0.05$, and with infinite degrees of freedom, $t_{0.05} = 1.96$.

The standard error of the difference between sample means is found as before:

$$S_{\bar{x}-\bar{y}} = \sqrt{S_{\bar{x}}^2 + S_{\bar{y}}^2} = \sqrt{450^2 + 450^2} = 636 \text{ years}$$

The t-ratio is

$$t = \frac{(\bar{X} - \bar{Y}) - \mu_{\bar{x}-\bar{y}}}{S_{\bar{x}-\bar{y}}} = \frac{(10{,}900 - 12{,}000) - 0}{636} = -1.73$$

Since $t = 1.73 < t_{0.05} = 1.96$, H_o is not rejected, and we conclude that the difference between dates A-40a and A-40b is not significant. The two dates could well date a contemporary event at the Lehner Ranch site.

What should we conclude when a significant difference emerges, indicating that two dates are really "different?" Statistically, this decision tells us that the two radioactive samples have probably been selected from different statistical populations, but the archaeological ramifications are more difficult to assess. Archaeologists generally assume that, all else being equal, a difference in radiocarbon dates results from a true age difference between the samples. But this remains only an assumption because several other factors could cause contemporaneous samples to "date" differently: impure CO_2, radon in counter, electronic circuit breakdowns, Geiger counter failure, cosmic ray showers, even atmospheric fallout. In the Tikal study alone, Satterthwaite and Ralph rejected over 40 percent (34 of 83) of their counting runs as spurious. There is also the danger of contaminating the sample itself by sloppy excavation, by percolating groundwater, by rodent burrowing, by rootlets, or even by insects.

It is always possible to introduce significant error into the samples and hence create a spurious radiocarbon date. There are even cases when several dates on the same log have produced widely different age determinations, although the samples must be of exactly identical age. There seems to be many a "slip 'twixt the cup and the lip" in radiocarbon dating, and statistical inference establishes whether or not a significant discrepancy exists between dates. Only nonstatistical considerations can *explain* that discrepancy.

Problems may also arise when structuring the research hypotheses into statistical hypotheses. If the Lehner Ranch null hypothesis had been directional (one-tailed), the region of rejection would have been $t_{0.10} = 1.65$, and the observed difference between the dates would have been declared "significant." The two-tailed alternative was selected in this case because no prior hypotheses existed to suggest *which* sample should be older than the other. It simply turned

out that the determination for A-40b was older than that for A-40a. But had there been some specific reason to suggest that A-40b would be older *before the actual results were known*, then a one-tailed test would have been in order.

● *Ours is the age which is proud of machines that think, and suspicious of men who try to.*—H. Jones

10.8 THE CASE OF PAIRED VARIATES

The data considered thus far were purposely selected so that each variate was totally uneffected by the other sample variates. The assumption of independent variates follows from our earlier definition of random sampling. But there are some hypotheses of interest involving data which are not independent of one another; the variates are "paired" with each other. Consider the following hypotheses:

Right-handed individuals tend to have larger right arms than left arms.
First-born individuals are usually stronger than their second-born siblings.
Students are rarely smarter than their professors.
Wives tend to be more motivated than their husbands.

These variates are linked into naturally occurring dyads (right-left, male-female, older-younger), and such linkage vitiates any usage of the *t*-test discussed so far.

Pairing of variates has an importance far greater than simple convenience because pairing is a tactic in the general strategy of efficient research design. The idea behind a purposeful pairing of variates is to increase the basis of comparison on a desired effect. Extraneous factors ("noise") can sometimes produce a significant difference even when there is no difference resulting from the phenomenon under study. Conversely, these same extraneous factors can sometimes mask a true difference, resulting in an incorrect acceptance of the null hypothesis. Errors of this sort can never be totally eliminated, but cautious design of experiments can purge a great deal of noise from the data.

A basic rule in designing experiments is to control what can be controlled and to randomize the uncontrollable. Pairing controls extraneous factors by grouping variates which are alike in all respects save the condition under study. In learning experiments, for instance, subjects are often paired by IQ scores so that variable degrees of intelligence will not mask the actual rates of learning or retension. Pairs are also commonly constructed to control for bias by sex, age, generation, socioeconomic background, motivation, or achievement. Acculturation studies often involve the natural pairings produced in "before-after" observations. Such variates are termed *self-pairing* when a single variable is measured on two occasions under different conditions.

But the use of paired variates destroys the assumption of statistical independence and necessitates an alteration in *t*-testing methods. The following example illustrates this simple modification.

A controversial topic in anthropology has been the so-called nature-nurture problem: To what extent is behavior conditioned by environmental as opposed

to genetic factors? Identical (monozygotic) twins are a common tool in this dialogue, especially when an investigator can study pairs of twins who have been raised separately, under different environmental conditions. If the performance of the twins varies, this difference is probably due to environmental factors, since the twins have inherited identical genetical material. Below are the actual performance scores of 11 pairs of identical twins. Each twin was rated on the quality of his or her educational background, and then each was tested on the Stanford–Binet IQ test (data from Newman, Freeman, and Holzinger 1937: chapter 10). Does a superior educational background produce a highly significant difference in IQ?

Pair	Superior Education	Inferior Education
I	97	85
II	78	66
III	101	99
IV	106	89
V	93	89
IX	102	96
X	127	122
XI	116	92
XII	109	116
XVII	115	105
XVIII	96	77

By inspection we see that in nearly all cases (10 of 11), the individual from the superior educational background also exhibits a higher score on the IQ test. But we also know that such results might occur by chance alone. Let us find just how likely (or unlikely) these findings really are.

The population standard deviation ($\sigma_{\bar{x}-\bar{y}}$) is unknown and the sample size is too small to use the sample standard deviation ($S_{\bar{x}-\bar{y}}$) to estimate that parameter. Because two distinct samples are involved, one might be tempted to apply the t-test to compare the two samples (Section 10.5). The hypotheses would be

$$H_0: \quad \mu_X \leq \mu_Y \qquad H_1: \quad \mu_X > \mu_Y$$

where μ_X represents the average IQ score of the twin raised in the superior educational environment.

But such a test would be incorrect because a basic assumption has been violated. Not only must the population variances be equal and both populations follow a normal distribution, but the two populations must also be *statistically independent of one another.* The standard t-test requires that the selection of variates in the first sample be logically independent from selection of the second sample. But since each individual in the first sample has a corresponding individual (its twin) in the second sample, neither samples nor populations are independent.

We must introduce a new variable in order to test for differences in paired

data:

$$D = (X_i - Y_i)$$

The paired scores are subtracted and their difference produces a new variable, called D ("the pair differences"). In effect, D recasts the pairs into a single sample. Sample statistics can then be found in the conventional manner, except that the values of D_i are substituted for the X_i:

$$\bar{D} = \frac{\Sigma D_i}{n}$$

$$S_d = \sqrt{\frac{\Sigma (D_i - \bar{D})^2}{n-1}}$$

$$S_{\bar{d}} = \frac{S_d}{\sqrt{n}}$$

where n is the number of pairs. To determine the sampling distribution, t is computed as

$$t = \frac{\bar{D} - \mu_{\bar{d}}}{S_{\bar{d}}} \tag{10.7}$$

where $\mu_{\bar{d}}$ is the population value of the mean difference. The number of degrees of freedom are $df = (n - 1)$.

The data in the example are analyzed as follows:

Statistical hypotheses:

$$H_o: \quad \mu_{\bar{d}} = 0 \qquad H_1: \quad \mu_{\bar{d}} \neq 0$$

Region of rejection: For a significance level of 0.01 in a two-tailed test with $df = (11 - 1) = 10$, $t_{0.01} = 3.169$.

The t-ratio is most easily found by using the following table.

Pair	Superior Ed., X	Inferior Ed., Y	D	$(D - \bar{D})$	$(D - \bar{D})^2$
I	97	85	+12	+ 2.5	6.25
II	78	66	+12	+ 2.5	6.25
III	101	99	+ 2	- 7.5	56.25
IV	106	89	+17	+ 7.5	56.25
V	93	89	+ 4	- 5.5	30.25
IX	102	96	+ 6	- 3.5	12.25
X	127	122	+ 5	- 4.5	20.25
XI	116	92	+24	+14.5	210.25
XII	109	116	- 7	-16.5	272.25
XVII	115	105	+10	+ 0.5	0.25
XVIII	96	77	+19	+ 9.5	90.25
			+104		760.75

$$\bar{D} = \frac{104}{11} = 9.5$$

$$S_d = \sqrt{\frac{760.75}{10}} = 8.72$$

$$S_{\bar{d}} = \frac{8.72}{\sqrt{11}} = 2.63$$

Substituting into Expression (10.7) to find t with 10 degrees of freedom,

$$t = \frac{9.5 - 0}{2.63} = 3.61$$

The computed value of t is sufficiently large to fall within the region of rejection. We conclude that a superior educational environment does seem to influence IQ scores when hereditary factors are held constant.

Example 10.8

Early twentieth century anthropology attempted to combat the prevalent racist theories of the time by demonstrating how environmental factors often overshadow the influence of heredity (that is, race). Franz Boas, himself a member of an immigrant minority, argued that the better nutritional and health care available in the United States caused far-reaching physical effects on the offspring of recent immigrants. Boas collected an incredible volume of data on physical changes occurring in immigrants and their children so that he could monitor the relationship between environmental and hereditary factors. The data in the following table are stature measurements for American-born and foreign-born Bohemian males (data from Boas 1912: table 1, appendix). Informants are paired to eliminate age effects. Do the American-born Bohemians appear to be larger than their foreign-born counterparts, as Boas suggested?

These data cannot be compared by the simple t-test for the difference between sample means because the informants have been purposely paired into age grades. But we can test the hypothesis that the average difference between the American-born and foreign-born informants is significantly different from zero. That is

$$H_o: \ \mu_d \leq 0 \qquad H_1: \ \mu_d > 0$$

Age	American-born Males, cm	Foreign-born Males, cm	D	$(D - \bar{D})$	$(D - \bar{D})^2$
4	99.4	98.0	+1.4	−0.5	0.25
5	105.7	101.0	+4.7	+2.8	7.84
6	110.7	110.6	+0.1	−1.8	3.24
7	116.0	111.7	+4.3	+2.4	5.76
8	122.5	118.2	+4.3	+2.4	5.76
9	128.5	128.1	+0.4	−1.5	2.25

Age	American-born Males, cm	Foreign-born Males, cm	D	$(D-\bar{D})$	$(D-\bar{D})^2$
10	132.7	135.1	−2.4	−4.3	18.49
11	137.7	134.7	+3.0	+1.1	1.21
12	141.1	140.0	+1.1	−0.8	0.64
13	147.9	148.1	−0.2	−2.1	4.41
14	152.3	150.4	+1.9	0.0	0.0
15	155.5	155.2	+0.3	−1.6	2.56
16	162.7	160.7	+2.0	+0.1	0.01
17	167.6	165.0	+2.6	+0.7	0.49
18	175.0	167.7	+7.3	+5.4	29.16
19	171.2	167.0	+4.2	+2.3	5.29
20	168.6	171.0	−2.4	−4.3	18.49
			+32.6		105.85

$\bar{D} = 32.6/17 = 1.9$ cm; $S_d = \sqrt{105.85/16} = 2.57$; $S_{\bar{d}} = 2.57/\sqrt{17} = 0.62$.

These values are substituted into the t-ratio:

$$t = \frac{1.9 - 0}{0.62} = 3.06$$

The critical region in this case is defined for a one-tailed test with df $= (17 - 1) = 16$ and a significance level of 0.01. From Table A.4 we find $t_{0.02} = 2.583$. The computed t-ratio exceeds this value, so we reject H_o and conclude that the sample of American-born Bohemians are significantly taller than those of foreign birth. Note again how a careful pairing of the data permits us to control for age in this experiment.

10.9 ASSUMPTIONS OF THE *t*-TEST

Once statistical hypotheses are formulated, more than one statistical test method is often available to test the propositions. Exactly which test is appropriate depends upon the underlying models and assumptions. There is a real danger in applying tests to data which violate critical assumptions, since false assumptions lead to the rejection of H_o just as surely as can the legitimate properties of the data. The null hypothesis of a particular test might be concerned with comparing two sample means. For instance: Should the underlying assumptions of the test model not be met, the results can appear "significant" whether or not there is any true difference between the two means. As long as there is doubt about the validity of the assumptions, one cannot be certain that H_o has been properly rejected or whether the rejection results from a spurious assumption.

Four explicit assumptions accompany the application of Student's t-test: interval scale of measurement, independent errors, normally distributed popula-

tion, and homogeneity of variance. Let us consider each of the prerequisites in more detail.

1. The variable is measurable on an interval scale. Level of measurement is really more a procedural matter than an assumption of the *t*-test. The sample mean and variance appear in the *t*-ratio, and these statistics can be computed only upon interval (or ratio) level variates. Nonparametric alternatives to the *t*-test are readily available whenever the level of measurement fails to reach an interval scale (see Chapter 12).

2. The variables must exhibit independent errors (except for paired variates). This second assumption requires that the selection of any single variate in no way influences the probability of selection of any other variate from the population. This requirement rarely poses a problem in disciplines such as psychology, where research usually centers about closely controlled experiments. The psychologist usually establishes purposeful pairing, control groups, repetitive testing, or some other technique to maintain the independent errors of observation. But too little attention has been paid in anthropology to the problems of research design, especially by archaeologists and paleoanthropologists. Sampling in anthropology is a distressingly complex subject and will be considered in more detail in Chapter 15.

3. The sample variates are randomly selected from a normally distributed population. It was necessary to assume that the basic population distribution was originally normal in order to find the exact probability distribution of the *t*-ratio. This is due to a theorem of mathematical statistics which states that, given random and independent observation, the sample mean and variance are independent of one another *if and only if* the population distribution is normal (see Mood and Graybill 1963: 228–231). Only for normal distributions can we be certain that the random variables necessary for the *t*-ratio (the sample mean and standard deviation) are statistically independent.

Unfortunately, we can seldom justify this assumption in practical application. Faced with the problem of analyzing obviously nonnormal data, one could attempt to transform the data into a form which does meet this assumption (by methods discussed in Chapter 14) or look elsewhere for another statistical test. The distribution-free (nonparametric) family of statistics are particularly useful in this regard (Chapters 11 and 12). But even nonparametric statistics exact a price because we lose some available information in exchange for freedom from restrictive assumptions.

There is, fortunately, another alternative. Now that the assumption of normality has been clearly stated and justified on mathematical grounds, it becomes my pleasure to inform you that normality can be ignored in most applications of the *t*-test. Mechanical sampling experiments by investigators in the early 1930s and recent computer simulations have shown that nonnormality has only a slight effect on the *t*-test as long as (1) the sample sizes are fairly large and (2) the test is not directional. The only error introduced into two-tailed testing is a slight modification in the true level of probability. If, for example, one operates within a tabled significance level of 0.05, the actual probability of a nonnormal population will really lie somewhere between 0.04 and 0.07, depending upon the degree of skewness. Thus, the overall effect of ignoring the normal assumption is that the table value of *t* will lead us to report slightly too many significant

findings (Cochran 1947). With this in mind, one should attempt to use larger samples when the underlying normality of the variates is in question.

More serious errors result from one-tailed testing because highly skewed distributions can seriously alter the tabled values of *t*, seriously over- or underestimating the true probability figures. A larger sample should be taken when one suspects a departure from normality in a directional hypothesis. Some techniques for detecting such departures from normality are discussed in Chapters 11 and 14.

4. When comparing two samples, the two parent populations must have homogeneous variances. Although the *t*-test does not directly involve the population variances, σ_X^2 and σ_Y^2, these two parameters must still be assumed to be equal. This is necessary so that observed differences between samples can be ascribed strictly to differences in the *central tendencies* rather than to differing *shapes* of the distributions about the mean. This important assumption, sometimes termed *homoscedasticity*, is a concept we will encounter again in the discussion of correlation.

Note that the assumption about homogeneity of variances applies only when two small samples are being compared. There is no assumption about σ when testing a single sample because S is obtained empirically and substituted directly into the *t*-ratio.

But what if this assumption is violated? Although the assumption of homogeneity of variance is more critical than that of normality, sampling experiments also indicate that (1) as long as the sample sizes are roughly equal and (2) the parent populations have distributions of approximately the same shape, the two population variances can deviate substantially from one another without introducing undue error into the level of probability. As long as these conditions are met (no matter what the variances may be) samples as small as $n = 5$ will produce acceptable results. The only difficulty is that a tabled probability value of 0.05 will only be within ± 0.03 of the true level. For samples larger than 15, the true probability will most likely be within ± 0.01 of the true value. When one has strong reason to suspect that the variances are truly unequal and the distributions are also of different shapes, then one should explore the possibility of applying the *Behrens test*, described in Bliss (1967: 215–218).

● *Sanity is not statistical.*—G. Orwell

SUGGESTIONS FOR FURTHER READING

Statistical Aspects of Radiocarbon Dating

Long and Rippeteau (1974)

Ralph (1971). A beginner's introduction to the laboratory and statistical methods involved in radiocarbon dating; Ralph takes a single charcoal sample through the dating process at the University of Pennsylvania laboratory.

Spaulding (1958)

EXERCISES

10.1 The average pithecanthropine cranial capacity is generally estimated to be about 1000 cc. Based upon cranial capacity alone, could the skulls discussed in Example 10.4 be pithecanthropine ($\alpha = 0.01$)?

10.2 A group of ten male skeletons has just arrived at a large eastern museum. Unfortunately, they have been improperly catalogued, and their place of origin is uncertain. Based upon the inadequate records available with the collection, the museum staff has guessed that these are North American Indian skeletons. The physical anthropologist in charge computes that the average stature of the ten specimens is 161.3 cm with $S = 10$ cm. Judging strictly from stature, is there sufficient reason to doubt that these skeletons are American Indian? (Kelso 1970: 235, gives 163.7 cm as the average Amerind stature.)

10.3 The following two radiocarbon dates were obtained for level DI at Danger Cave, Utah (Jennings 1957: table 11):

$10{,}270 \pm 650$ (M-204)
$11{,}151 \pm 570$ (C-610)

(a) What are the two-thirds limits of confidence for the Michigan date?
(b) What are the 95 percent limits for the true age of the Chicago sample?
(c) What is the probability that the true age of M-204 is actually older than 10,800 years?
(d) What is the probability that the true age of M-204 lies between 10,000 and 11,000 years old?
(e) What is the probability that C-610 is actually 10,800 years or younger?
(f) Suppose that the true age of both samples was known to be 10,715 years. Which sample came closer (in terms of probability) to estimating the true age? (*Hint*: Be certain to consider the relative standard deviations.)

10.4 In a study designed to determine the relationship between climate and facial structure, Koertvelyessy (1972) obtained the following figures for frontal sinus surface area in Eskimo males:

	Mean, cm^2	Standard Deviation, cm^2	n
Colder habitat	2.076	1.974	33
Warmer habitat	3.794	2.866	29

(a) Do the Eskimo from the colder environment tend to have significantly smaller frontal sinuses?
(b) Would an Eskimo with a frontal sinus area of 5.0 cm^2 be considered "aberrant" in the colder habitat?

10.5 A team of investigators measured the root length of the first mandibular premolar in a sample of American Whites and American Blacks (data from Moss, Chase, and Howes 1967: table 4):

	American Whites, mm	American Blacks, mm
Mean	14.8	14.4
Standard deviation	0.97	1.97
n	15	7

(a) Is there a significant difference in root length?
(b) Could the American White population average a root length longer than 15 mm?
(c) Could the American Black population average less than 12 mm?

10.6 Two kinds of rooms are often found in the pueblos of the American Southwest: large rooms, probably involved in day-to-day living, and smaller rooms, most probably used for storage (Hill 1970). One useful indicator of the prehistoric function of these rooms involves the kinds of pottery sherds they contain. Since modern pueblo families generally take their meals in the habitation rooms, we can expect to find more pottery from food plates and bowls in the habitation areas than in the storage rooms. Similarly, large storage jars should be more common in the storage rooms. Unfortunately, several other variables—such as family size, methods of food preparation and storage, differential hygiene (some families sweeping their floors cleaner than others), and time of occupation—also enter into the recovery of pottery sherds, hence obscuring room function.

In order to minimize the effects of these extraneous factors and concentrate strictly upon room function, an experiment was designed to test for differences in pottery frequency. In a particular pueblo, it became apparent that each large room was directly connected by a doorway to a smaller room. The inference is that a single family used both rooms, one for storage and the other for habitation. By pairing the large and small rooms on the basis of a shared doorway, many of the extraneous variables, such as family size, sanitary practices, and so forth, can be controlled. After excavation, the density of cooking sherds was computed as follows:

	Sherds per Cubic Meter	
Doorway	Large Room	Small Room
A	23	11
B	42	36
C	12	10
D	15	17
E	62	49
F	39	28

Assuming that these sherds reflect only food-preparation vessels and not food storage, can we conclude at the 0.05 level that more cooking took place in the large rooms?

*10.7 The Grasshopper site is a rather large masonry ruin located in Arizona. In an attempt to infer changes in prehistoric social organization, excavators have carefully recorded the dimensions of each room, and also of the fire hearths associated with rooms (data from Ciolek-Torello and Reid 1974: table 1):

Room	Room Size, m^2	Firehearth Size, cm^2
Later rooms		
3	18.6	838
5	13.7	589
6	15.8	1860
7	21.3	1440
11	12.9	1456
13	14.4	800
205	17.7	1004
216	17.5	761
218	22.4	1435
319	23.4	1444
349	16.1	1386
359	25.9	1140
371	15.3	1013
398	12.4	870
425	12.6	1534
Earlier rooms		
1	17.3	1864
2	16.4	1350
18	22.0	2937
28	18.1	1564
146	15.7	1665

(a) Is there a significant difference in room size between early and late rooms?

(b) Do these data support the hypothesis that the earlier rooms had larger firehearths?

(c) Is the firehearth size more variable in the earlier rooms?