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.

12 Nonparametric Statistics: Ordinal Scales

● *It is better to be ignorant than to know what ain't so.*—S. Ervin

12.1 RANK-ORDER STATISTICS

Chapter 11 considered some statistics relevant to nominal scale data. These statistics were called *nonparametric* because no minimal level of measurement was stipulated—and nominal is as low as one can go—or because no assumptions were necessary regarding the population distribution. This chapter presents further nonparametric methods by considering statistics appropriate to the ordinal level of measurement. These techniques are sometimes called *rank-order statistics* because variates are usually arrayed along an ordered scale rather than being actually measured.

12.2 THE WILCOXON TWO-SAMPLE TEST

The Wilcoxon test examines two samples to see whether their respective populations have different central locations. The *t*-test did this by looking at the sample means. After considering the variances, the *t*-test assessed whether the two samples represented the "same" population mean, or "different" population means, given alpha. The Wilcoxon test also serves this function, but on a different sort of data. The *t*-test required an interval scale of measurement, while the Wilcoxon test is designed for ordinal-level data. As discussed in Chapter 2, ordinal-level samples have no "measurements" in the strict sense. Because variates are simply placed into a relevant order (or *ranking*), ordinal samples cannot be characterized by means. The *median* must suffice for ordinal variates. Thus, the *t*-test examines for a difference between population means, and the Wilcoxon test looks for different population medians.

Suppose that two prehistoric cemeteries were excavated. The stature of each individual could be estimated by measuring the relevant bones, and a *t*-test could tell whether the first population was taller than the second. But suppose that the cemeteries had been disturbed by pothunters, and too few bones were available for the physical anthropologist to make reliable stature estimates. In this case, the burials could be ordered only in a sequence from relatively short to relatively tall, based upon relative robusticity of the bones. The *t*-test would be useless here because no sample means can be computed. The best we could do, given the ranked nature of the skeletal information, would be to find the median (or halfway point) in each skeletal series. The Wilcoxon test could then be used to look for stature differences between the cemeteries.

The initial step in all ordinal testing is to place the variates in a numbered sequence (called a *rank order*). The skeletons from cemetery A would be lined up by increasing stature next to those from cemetery B. Find the shortest skeleton in either collection, and give this specimen the rank 1. The second shortest gets a 2 and so forth until all skeletons have been numbered. Now sum the ranks for the first cemetery and call this sum W_1. The sum of the rankings for the second sample is called W_2. The Wilcoxon test provides a method to tell whether the first samples tend to rank higher overall than the second sample. This would mean that the individuals in the first cemetery tended to be taller. The null hypothesis of the Wilcoxon test holds that W_1 should be greater than W_2 only about half the time (assuming the samples to have the same number of variates). If the sum of ranks for the two samples is roughly the same, there is no reason to doubt H_0. That is, there is no reason to suspect a difference between the medians. But the larger the difference between W_1 and W_2, the less likely it becomes that the samples will be random samples from populations with identical medians. The directional alternative hypothesis suggests either that W_1 in fact exceeds W_2, or perhaps vice versa. The Wilcoxon statistic enables us to see whether significant differences exist between W_1 and W_2.

Another example will illustrate these computations. Anthropologists often assume that there is some advantage for hunting societies to keep the related males within the same residential group throughout their life. Not only do hunters cooperate and share more readily with kinsmen, but they are also more effective when hunting in familiar home territories. So, it is hypothesized that hunting groups should tend to be patrilocal, and this hypothesis can be tested against a random sample of North American societies selected from the *Ethnographic Atlas*. Societies in the *Atlas* can be characterized as either *patrilocal* or *matrilocal*, and can also be rated on the relative importance of hunting in the overall economy. If the above hypothesis is correct, then patrilocal societies should tend to be more dependent upon hunting than are matrilocal groups. The following eight societies were randomly selected:

	Atlas Code	Relative Dependence upon Hunting, %
Matrilocal		
Huron	1	6-15
S. Ute	6	56-65

	Atlas Code	Relative Dependence upon Hunting, %
Matrilocal (contd)		
W. Apache	4	36–45
Antarianunts	3	26–35
Patrilocal		
Slave	5	46–55
Gros Ventre	8	76–85
Santee	7	66–75
Kiowa	9	86–100

Had the relative importance of hunting been expressed in precise percentages, conventional parametric methods could have been used to test for a significant difference. But since the relative importance of hunting is estimated in only rather gross intervals, the exact methods of the *t*-test are inapplicable. To repeat, the *t*-test requires at least an interval scale, but these data are only ordinal.

These subsistence data can easily be rank-ordered according to the relative dependence upon hunting and the numerical rankings assigned to each society. The matrilocal societies and their associated ranks have been underlined.

Scale of hunting importance, %.

Slight Dependence ←						→ Strong Dependence	
Huron	Antarianunts	W. Apache	Slave	S. Ute	Santee	Gros Ventre	Kiowa
0–5	26–35	36–45	46–55	56–65	66–75	76–85	86–100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

The sum of the ranks for the matrilocal societies is

$$W_1 = 1 + 2 + 3 + 5 = 11$$

The sum of the ranks for the patrilocal societies is

$$W_2 = 4 + 6 + 7 + 8 = 25$$

The grand total for all ranks is

$$1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = W_1 + W_2 = 36$$

Note that the sum $W_1 + W_2$ is a constant for all situations containing exactly eight outcomes, regardless of the specifics of the samples.

The research hypothesis suggests that the matrilocal societies (sample 1) should have had less dependence upon hunting than had the patrilocal sample. Thus, the sum of ranks for the patrilocal societies is greater than the sum of ranks for the matrilocal societies: $W_2 > W_1$. Is the result $(W_2 - W_1) = 14$ a sufficient deviation for the result to be considered statistically significant?

In probabilistic terms, there are eight independent trials in this experiment, so

the probability must be found that a sample of $n_1 = 4$ such outcomes will have ranks which sum to less than or equal to the observed value of $W_1 = 11$. That is, the situation requires the number of combinations of eight items taken four at a time:

$$C_{8,4} = \frac{8!}{4!\,4!} = 70$$

By trial and error it is found that only two possible ways exist to obtain a sum of ranks less than or equal to $W_1 = 11$:

$$W_1 = 1 + 2 + 3 + 4 = 10$$

$$W_2 = 1 + 2 + 3 + 5 = 11$$

There are no other possibilities which sum to 11 or less. Thus the total probability of obtaining $W_1 \leq 11$ if H_o is true is

$$p = \frac{2}{70} = 0.0286$$

The statistical hypotheses were one-tailed, so the results are significant beyond $\alpha = 0.05$, and H_o is rejected. This sample is thus consistent with the notion that hunting societies in North America tend to be patrilocal.

To clarify just how the Wilcoxon test is used for statistical inference, this example can be recast into the six steps of hypothesis testing.

Step I. *Statistical hypotheses:* The research hypotheses are as follows:

H_o: Postmarital residence is independent of dependence upon hunting (or matrilocal societies tend to hunt more than patrilocal societies).

H_1: Patrilocal societies tend to hunt more than matrilocal societies.

These statements now must be translated into specific statistical hypotheses. If W_1 is the sum of ranks for matrilocal societies,

$$H_o: \quad p(W_1 = W_2) \geq 1/2$$

$$H_1: \quad p(W_1 = W_2) < 1/2$$

The two propositions actually reflect the relationships between the respective population medians.

Step II. *Statistical model:* The distribution of the Wilcoxon statistic provides a statistical model against which to judge the specific sample values. The Wilcoxon two-sample test assumes (1) both samples are randomly selected, (2) the samples are independent, (3) at least ordinal measurement, and (4) both samples are variates of continuous random variables (see Section 12.7).

Step III. *Significance level:* Let $\alpha = 0.05$ for a directional test.

Step IV. *Region of rejection:* The Wilcoxon test is called an *exact* test because the result is a specific *point* rather than an *area*. The "region of rejection" for this case is defined directly from the level of significance. Any probability for the

TABLE 12.1 Maximum cranial length measurements (in millimeters) from two series of fossil men (data from Coon 1971a: table 37).

Homo erectus		Neanderthals and Skhūl	
Pithecanthropus 4	199?	La Ferrassie	209
Pithecanthropus 1	183?	Neanderthal	199
Solo 11	200	Spy 1	201
Sinanthropus 3 (*)	188	Circeo 1	204
Sinanthropus 10	199	Le Moustier Y	196
Sinanthropus 12	185.5	Tabun 5	206
Saldanha	200	Skhūl 5	192
Broken Hill	208	Skhūl 9	213

(*) Young, subadult; all others are adult.

length measurements which would seem to indicate that *Homo erectus* had a shorter head than Neanderthal. But is this rather small difference in head length statistically significant?

Upon initial inspection of the data, one might be tempted to use a simple *t*-test, but a closer look indicates that the measurements lack the accuracy implied by the *t*-test. The measurements for both Pithecanthropus skulls, for instance, are little more than guesses, while Sinanthropus 12 was accurately measured to 0.1 mm. In addition, the Sinanthropus 3 skull is not a mature individual, so the cranial length is probably somewhat less than that of the adult form. When dealing with specimens as rare as complete fossil crania, one simply cannot control the errors of measurement with much precision; often, inconsistent measurements such as these must suffice. To avoid implying spurious accuracy, the length measurements in Table 12.1 have been reduced to ordinal relationships; the relative rank ordering is thus maintained without implying true interval accuracy.

Considering the *Homo erectus* specimens as sample 1, the following rank ordering is achieved (sample 1 underlined).

Original Data	Rank Number	Original Data	Rank Number
183	1	200	9.5
185.5	2	200	9.5
188	3	201	11
192	4	204	12
196	5	206	13
199	7	208	14
199	7	209	15
199	7	213	16

Note that in two cases (199 and 200 mm), the variates were tied. The rank number is assigned in such instances by using the *average ranking* for that

score; it is thus necessary to sum the tied ranks and divide by the number of ties involved.

$$\text{Rank}_{199} = \frac{6+7+8}{3} = 7$$

$$\text{Rank}_{200} = \frac{9+10}{2} = 9.5$$

The sum of the ranks is found to be

$$\underline{W_1} = \underline{1} + \underline{2} + \underline{3} + \underline{7} + \underline{7} + \underline{9.5} + \underline{9.5} + \underline{14} = 53$$

$$W_2 = 4+5+7+11+12+14+15+16 = 84$$

The value of U is found in the usual manner:

$$U = 53 - \frac{1}{2}8(8+1) = 17$$

Table A.6 indicates that $C_{16,8} = 12{,}870$ and the corresponding value of $U = 17$ is 879, so the two-tailed probability in this case is

$$p = \frac{2(879)}{12{,}870} = 0.1366$$

Because of the relatively large probability figure, we conclude that this sample provides insufficient evidence to reject H_0, assuming $\alpha \leq 0.1366$. We can demonstrate no significant difference between cranial lengths of *Homo erectus* and Neanderthal.

This section has tacitly introduced a slightly different mode of statistical inference. Because the probability values computed by the Wilcoxon test are *exact*, they can be directly used for statistical inference without bothering with a region of rejection. Thus, the value of $p = 0.1366$ will be insufficient to reject H_0 for any alpha level *less than or equal to* 0.1366. This interval includes most common significance levels (that is, 0.05, 0.01, 0.001), and we can safely assume that almost all investigators would retain H_0. The real advantage of exact tests is that the alpha level need not be specified. A current trend in social science applications of statistics is to forego the actual hypothesis-testing procedure and simply state exact levels of probability. This leaves the decision "reject or not reject" to the reader. This trend is perfectly healthy, as long as we understand the procedures of statistical inference *when we finally do wish to make a decision.* Sometimes specifying the six steps helps insure that the statistical model and its assumptions have actually been met.

Example 12.2

The Midland site in west Texas yielded one of the oldest human skulls in the Americas. The artifact inventory included some rather conventional Folsom projectile points and also an artifact called a *Midland* point. The Midland points are identical in every way to the Folsom finds except that Midland points lack the diagnostic channel flute. Since their discovery, Midland points have been found in a number of localities in the American Southwest, but archaeologists are still hard-pressed to explain the curious absence of the channel flute. Some suggest that the Midland points were

originally manufactured on a very thin flake, and therefore the blank was too thin to channel. This argument makes a certain amount of sense if the channel flake was executed to thin the artifact; then artifacts already quite thin would not need to be fluted. The measurements below are for artifacts found at the original Midland site.

Are the Folsom points significantly thicker than the Midland points?

For demonstration purposes, these thickness measurements will be reduced to rank orderings, and the Wilcoxon two-sample test will be used to compare the two samples statistically.

Thickness measurements for artifacts from the Midland site (data from Wendorf, Krieger, Albritton and Stewart 1955).

Folsom		Midland (unfluted Folsom)	
Catalog No.	Thickness, inches	Catalog No.	Thickness, inches
16	0.14	19	0.13
17	0.08	24	0.12
18	0.14	25	0.11
55	0.19	27	0.19
74	0.19	29	0.10
		30	0.09
		31	0.11
		32	0.10

The variates must first be rank-ordered:

Folsom	.08								.14	.14		.19	.19
Midland		.09	.10	.10	.11	.11	.12	.13			.19		
Ranking	1	2	3.5	3.5	5.5	5.5	7	8	9.5	9.5	12	12	12

The sum of ranks can be computed next:

$$n_1 = 5 \qquad W_1 = 1 + 9.5 + 9.5 + 12 + 12 = 44$$

$$n_2 = 8 \qquad W_2 = 2 + 3.5 + 3.5 + 5.5 + 5.5 + 7 + 8 + 12 = 47$$

The Wilcoxon statistic can now be computed:

$$U = 44 - \frac{5(5+1)}{2} = 44 - 15 = 29$$

$$C_{13,5} = \frac{13!}{5!8!} = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{5 \cdot 4 \cdot 3 \cdot 2} = 1287$$

This value of U exceeds the tabled frequencies, so H_o is not rejected. We can conclude that these data do not support the suggestion that Folsom points tend to be thicker than Midland points.

A two-sample t-test of these same data produces a test statistic of $t = 1.465$ with 11 degrees of freedom. This value of t is not significant at even the $\alpha = 0.1$ level. In general, the t-test and the Wilcoxon two-sample test will produce almost identical results.

12.2.2 Normal Approximation to the Wilcoxon Two-Sample Test

As long as neither n_1 nor n_2 exceeds 8, Table A.6 can be used to find the appropriate probability values associated with the Wilcoxon statistic. But ordinal scales are so widespread in the social sciences that problems commonly arise which are appropriate to the Wilcoxon test, but which exceed the tabled values of n_1 and n_2. As an alternative to computing additional—and more cumbersome—probability tables to accommodate these larger samples, it has been shown that as long as the samples are large enough, the distribution of W_s approaches a normal distribution with the following parameters[1]:

$$\mu_W = \frac{n_1(n+1)}{2}$$

$$\sigma_W = \sqrt{\frac{n_1 n_2(n+1)}{12}} \qquad (12.2)$$

where $n = n_1 + n_2$. These formulas assume that no ties are present.

The following example approximates the Wilcoxon test through use of the normal distribution. Phyllis Jay Dolhinow made extensive observations on the dominance behavior of female langurs in Northern India. Dolhinow hypothesized that social position is largely a function of the individual female's status as a mother and also her phase in the reproductive cycle (Dolhinow 1972: 220). As field studies progressed, individual females became identifiable on sight, and Dolhinow was able to establish the female dominance hierarchy for the Kaukori langur troop (see Table 12.2). Assuming that all females suspected of pregnancy were in fact pregnant, do these data support the hypothesis that reproductive status is associated with position in the dominance hierarchy?

This comparison involves two groups, pregnant and not pregnant langurs, each of which has been ranked for dominance; hence, the Wilcoxon test is clearly in order. The rank ordering can be recast.

Original Data	Rank Ordering	Original Data	Rank Ordering
A	1	J	10
B	2	K	11
C	3	L	12
D	4	M	13
E	5	N	14
F	6	O	15
G	7	P	16
H	8	Q	17
I	9	S	18

The ranks of the pregnant females have been underlined. In this case, $n_1 = 7$ and $n_2 = 11$, values which are too large for Table A.6. The normal approximation to the Wilcoxon statistic will be used to derive a probability value for this event.

[1]For verification and discussion of these Wilcoxon parameters, see Wilcoxon (1947) and Adler and Roessler (1972: 179).

TABLE 12.2 Female dominance hierarchy of the Kaukori langur troop (modified slightly from Dolhinow 1972: tables 5–7). Individual "A" is most dominant, and "S" is most dominated.

Female	Reproductive Status	Female	Reproductive Status
A	Not pregnant	J	Pregnant
B	Pregnant	K	Not pregnant
C	Pregnant	L	Pregnant
D	Not pregnant	M	Pregnant
E	Pregnant	N	Not pregnant
F	Not pregnant	O	Not pregnant
G	Pregnant	P	Not pregnant
H	Not pregnant	Q	Not pregnant
I	Not pregnant	S	Not pregnant

The sum of the ranks in the first sample is

$$W_1 = 2 + 3 + 5 + 7 + 10 + 12 + 13 = 52$$

Because the sample size is relatively large, W_1 should be distributed roughly in normal fashion, with parameters given by Expression (12.2).

$$\mu_W = \frac{7(18+1)}{2} = 66.5$$

$$\sigma_W = \sqrt{\frac{7(11)(19)}{12}} = 11.04$$

The question now concerns the probability of obtaining results as deviant as $W_1 = 52$, where $\mu_w = 66.5$ and $\sigma_w = 11.04$. This is accomplished by using the normal approximation:

$$z = \frac{52 - 66.5}{11.04} = -1.31$$

which corresponds to an area of 0.0951. The two-tailed probability for the case of female langur domination is thus

$$p = 2(0.5000 - 0.4049) = 0.1902$$

H_0 must be retained for all $\alpha < 2(0.0951) = 0.1902$

The research conclusion is that this sample indicates no association between reproductive status and an individual's position in the female dominance hierarchy.

The normal approximation for the Wilcoxon test also holds when ties are present, but the following corrected formula must be used to compute the standard deviation:

$$\sigma_W = \sqrt{\frac{n_1 n_2 [n(n^2 - 1) - \Sigma T_i]}{12n(n - 1)}} \qquad (12.3)$$

where $n = n_1 + n_2$, and $T_i = (t_i - 1)t_i(t_i + 1)$ in which t_i = number of ties at rank i. This computation is illustrated in Example 12.4.

Example 12.3

In a classic study of the ecology of the American Southwest, Julian Steward (1937) demonstrated how ethnographic and archaeological data could jointly be focussed upon problems of general anthropological interest. Steward felt that the matrilineal clans of the Southwest could better be expressed as an adjustment to ecological pressures than through mere diffusion from neighboring areas. Steward argued that localized exogamous lineages had once been crowded during prehistoric times into large multilineage communities, probably due to increased population density. The unilateral groups devised ceremonies, totems, and other cultural devices which fostered group solidarity, thereby maintaining corporate identities. In time, the lineages thus evolved into clans. To support this thesis that small villages had once aggregated into large communities, Steward cited archaeological data showing that the number of habitation rooms increased through time in relation to ceremonial kivas. The data for the last two periods of Southwestern archaeology are presented below.

Do these archaeological data support Steward's hypothesis of a significant increase in the room:kiva ratio between Pueblo IV period (A.D. 1300–1700) and historic times?

Pueblo village growth (data from Steward 1955: 165–167).

Period	Site	House:Kiva Ratio
Pueblo IV	Tshirege, Rio Grande	60:1
PIV	Tsankawi, Rio Grande	30:1
PIV	Otowi, Rio Grande	90:1 (?)
PIV	Yapashi, Rio Grande	92:1 (?)
PIV	Kotyiti, Rio Grande	240:1 (?)
Pueblo V	Oraibi, Hopi	35.4:1
PV	Walpi, Hopi	34.2:1
PV	Sichumovi, Hopi	36:1
PV	Shipaulovi, Hopi	33.3:1
PV	Mishongnovi, Hopi	31.8:1
PV	Hano, Hopi	52.5:1
PV	Zuni	95.4:1
PV	Zuni	289.8:1

Although the data tabulated are expressed as ratios, the actual measurements are quite inconsistent: The Pueblo V figures are derived from

known family counts, while the Pueblo IV ratios consist of counts from archaeological excavations. To avoid misleading feelings of concreteness, the ratios have been reduced into rank orderings:

Original Data	Rank Ordering	Original Data	Rank Ordering
30	1	60	8
31.8	2	90	9
33.3	3	92	10
34.2	4	95.4	11
35.4	5	240	12
36.0	6	289.8	13
52.5	7		

The Pueblo IV data have been underlined, so $n_1 = 5$ and $n_2 = 8$. The sum of ranks is found to be

$$W_1 = 1 + 8 + 9 + 10 + 12 = 40$$
$$W_2 = 2 + 3 + 4 + 5 + 6 + 7 + 11 + 13 = 51$$

Because n_1 and n_2 are relatively small, Table A.6 could be consulted for the associated probability level. $C_{13,5}$ is found to be 1287, but there is no probability value listed for $U = 25$. This is because the probability is too large to bother listing. Nevertheless, let us find the exact probability by using the normal approximation to the Wilcoxon statistic:

$$\mu_W = \frac{5(14)}{2} = 35$$

$$\sigma_W = \sqrt{\frac{5(8)(14)}{12}} = 6.83$$

The standardized normal deviate is found to be

$$z = \frac{40 - 35}{6.83} = 0.73$$

The probability for a one-tailed alternative is

$$p = 0.5000 - 0.2673 = 0.2327$$

The result is not significant and H_0 is not rejected for $\alpha < 0.2327$. These data fail to lend support to the ecological hypothesis advocated by Steward. (In all fairness to Steward, however, he was interested primarily in a possible increase *before* Pueblo IV times, which turns out to be highly significant.)

Example 12.4

In their investigation of the relationship between child training practices and subsistence economy, Barry, Child, and Bacon (1959) hypothesized that child rearing among pastoral societies—in which food is extensively accumulated and stored—tends to value personality traits such as compliance and conservation. On the other extreme, societies with relatively little accumulation of food resources, particularly hunters and fishermen, were expected to reward individualism, assertiveness, and a venturesome attitude in youth. The authors selected a large cross-cultural sample to test their hypothesis, and the societies were scored on the relevant personality traits. Positive scores were awarded to groups with a relatively high degree of compliance, while assertion was rated negatively.

Do these cross-cultural findings support their hypothesis at the 0.01 level?

Because of the relatively large size of the samples, it becomes necessary to apply the normal approximation to the Wilcoxon two-sample test. The rank ordering appears as follows:

Animal Husbandry		Hunting, Fishing	
+13.5	Aymara	+ 4	Teton
+13.5	Tepoztlan	+ 1	Tahgan
+11.5	Lepcha	+ 0.5	Hupa
+ 8.5	Swazi	0	Chiricahua
+ 8.5	Tswana	0	Murngin
+ 8	Nyakyusa	0	Paiute
+ 8	Sotho	− 2	Arapaho
+ 7	Nuer	− 2	Kwakiutl
+ 7	Tallensi	− 2.5	Cheyenne
+ 6.5	Lovedu	− 2.5	Kaska
+ 6.5	Mbundu	− 2.5	Klamath
+ 6.5	Venda	− 2.5	Ojibwa
+ 6	Kikuyu	− 3	Ona
+ 6	Zulu	− 4	Aleut
+ 4.5	Pondo	− 6.5	Jicarilla
+ 4	Chagga	−10	Western Apache
+ 3	Ganda	−10.5	Siriono
+ 2.5	Chamorro	−11	West Greenland Eskimo
+ 2.5	Masai	−12	Aranda
+ 1	Chukchee	−12	Comanche
0	Tanala	−13.5	Crow
− 2.5	Thonga	−15	Manus
− 3	Araucanian		
− 3	Balinese		

Original data	Ranks	Original Data	Ranks
−15	1	1	25.5
−13.5	2	2.5	27.5
−12	3.5	2.5	27.5
−12	3.5	3	29
−11	5	4	30.5
−10.5	6	4	30.5
−10	7	4.5	32
− 6.5	8	6	34
− 4	9	6	34
− 3	11	6	34
− 3	11	6.5	36
− 3	11	6.5	36
− 2.5	15	6.5	36
− 2.5	15	7	38.5
− 2.5	15	7	38.5
− 2.5	15	8	40.5
− 2.5	15	8	40.5
− 2	18.5	8.5	42.5
− 2	18.5	8.5	42.5
0	21.5	11.5	44
0	21.5	13.5	45.5
0	21.5	13.5	45.5
0	21.5		
0.5	24		
1	25.5		

The sum of ranks for the hunting-fishing groups is

$$W_1 = 1 + 2 + 3.5 + 3.5 + 5 + 6 + 7 + 8 + 9 + 11 + 15 + 15 + 15 + 15 + 18.5 + 18.5 + 21.5 + 21.5 + 21.5 + 24 + 25.5 + 30.5$$
$$= 297.5$$

The parameters of the U distribution are

$$\mu_W = \frac{22(46+1)}{2} = 517$$

Because of the presence of ties, the standard deviation must be computed from Formula (12.3). The values of t_i are listed below.

Rank	No. of Ties	$(t-1)t(t+1) = T_i$
−12	2	1 · 2 · 3	6
− 3	3	2 · 3 · 4	24
− 2.5	5	4 · 5 · 6	120
− 2	2	1 · 2 · 3	6
0	4	3 · 4 · 5	60
1	2	1 · 2 · 3	6
2.5	2	1 · 2 · 3	6
4	2	1 · 2 · 3	6

Rank	No. of Ties	$(t_i - 1)t_i(t_i + 1) = T_i$
6	2	$1 \cdot 2 \cdot 3$ 6
6.5	3	$2 \cdot 3 \cdot 4$ 24
7	2	$1 \cdot 2 \cdot 3$ 6
8	2	$1 \cdot 2 \cdot 3$ 6
8.5	2	$1 \cdot 2 \cdot 3$ 6
13.5	2	$1 \cdot 2 \cdot 3$ 6
		$\Sigma T_i = 288$

The standard deviation is thus

$$\sigma_W = \sqrt{\frac{22 \cdot 24[46 \cdot (2116 - 1) - 288]}{12 \cdot 46 \cdot 45}}$$

$$= \sqrt{2058.9} = 45.38$$

The standardized normal deviate is thus

$$z = \frac{297.5 - 517}{45.38} = -4.84$$

The area associated with this extremely high value of z is too small to even appear in Table A.3, so the results are judged to be highly significant and H_o is rejected. This sample supports the hypothesis advanced by Barry, Child, and Bacon: that societies which rely upon stored food will tend to teach compliance while hunting–fishing groups tend to train their children toward more assertive behavior.

12.3 KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

Like the Wilcoxon procedure, the Kolmogorov-Smirnov test examines differences between two samples which have been measured into ordinal categories. Although the Wilcoxon test is still feasible with tied variates, the corrections for ties create considerable computational difficulties. The Kolmogorov–Smirnov test readily facilitates analysis of scales where many ties occur, yet avoids reducing the data to nominal relations, as does a conventional chi-square statistic.

The Kolmogorov–Smirnov test involves a rather simple underlying theory. Two ordinal level samples are involved, as in the Wilcoxon test. These two samples are arranged into a set of cumulative proportions; the procedure here is identical to that discussed in Section 3.3.4 when the cumulative curve (or *ogive*) was constructed. The null hypothesis of the Kolmogorov–Smirnov test asserts that the cumulative proportions of the first sample shall be essentially similar to those of the second sample. The larger the maximum absolute differences between the cumulative proportions, the less likely becomes H_o. The

distribution of the Kolmogorov–Smirnov statistic, D, is known and the critical values for the two-tailed Kolmogorov-Smirnov test are listed in Table A.8(b).

One of anthropology's thorniest problems is how to construct cross-cultural samples, an issue considered in detail in Chapter 15. The problem is whether the samples should be selected in a purely random fashion, or whether the universe should be "stratified" initially by continent (or culture area) and then sampled within the strata. Simple random sampling has certain statistical virtues, while careful stratification tends to eliminate the undesirable effects of cultural diffusion. Greenbaum (1970) has recently attempted to shed some light on this problem by statistically comparing samples generated through both methods. A simple random sample of 69 African societies was first selected from the total list of African cultures contained in the *Ethnographic Atlas* (Murdock 1967). Each society was rated on the variable *dependence upon agriculture*; these results appear in Table 12.3.

The 863 societies listed in the *Atlas* were then divided into 412 *cultural clusters*, each of which represents a grouping of highly similar societies which are known to have had extensive contact. So, each of the societies within a cluster are quite similar, and each cluster is dissimilar from its neighbors. A *stratified random sample* was then constructed of 69 African societies, with only one society permitted from each cluster (no clusters closer than 200 miles were permitted). This statistical method is designed to screen contamination due to proximity or diffusion, yet still produce a random sample. The stratified sample was also rated in Table 12.3 according to *dependence upon agriculture.*

Do these samples contain significant differences at the 0.01 level?

A nondirectional Kolmogorov–Smirnov two-sample test is relevant here, thus preserving the ordinal nature of the variable under study. The critical values of the two-tailed D statistic are given by Table A.8(b) to be

$$\begin{aligned} &\textit{0.05 level}: \quad 1.36\sqrt{\frac{n_1+n_2}{n_1 n_2}} \\ &\textit{0.01 level}: \quad 1.63\sqrt{\frac{n_1+n_2}{n_1 n_2}} \end{aligned} \tag{12.4}$$

TABLE 12.3 Comparison of simple random and stratified random samples for 69 African societies (data from Greenbaum 1970).

Dependence upon Agriculture, %	Random Sample		Stratified Sample		
	Raw	Cum., %	Raw	Cum., %	Difference
0–25	4	0.058	4	0.058	0.0
26–45	11	0.217	6	0.145	0.072
46–75	49	0.928	55	0.942	0.014
76–100	5	1.0	4		0.0
	69		69		

The *cumulative proportions* of each measurement class must first be computed. In the random sample given in Table 12.3, the proportion of societies depending less than 25 percent on agriculture is given by $4/69 = 0.058$. The proportion depending less than 45 percent on agriculture is $(4 + 11)/69 = 0.217$ and so forth. Similar computations are performed on the stratified sample, and the D statistic is simply the maximum deviation between the various pairwise comparisons. In this case, $D = 0.072$, the observed difference for the cumulative proportion for 0–45 percent dependence on agriculture.

The critical value of D at $\alpha = 0.01$ is computed from Expression (12.4) to be

$$D = 1.63 \sqrt{\frac{69 + 69}{69(69)}} = 0.277$$

The observed value of D falls short of the critical value, so the null hypothesis is not rejected. This experiment fails to show a significant difference between the two sampling schemes for the African continent.

To summarize the steps in using the Kolmogorov–Smirnov test:

Step I. *Statistical hypotheses:*

H_0: There is no difference in dependence on agriculture between random sampling and the stratified sampling.

H_1: There is a difference between the two sampling methods.

In a more rigorous sense, H_0 holds that D should be about equal to zero, while H_1 suggests that D will be significantly greater than zero.

Step II. *Statistical model:* The Kolmogorov–Smirnov model deals only with cumulative proportions. Under a true H_0, the respective unknown cumulative distributions should be identical, such that $D = 0$. The test assumes: (1) random sampling, (2) independent samples, (3) at least ordinal scale measurements, and (4) underlying continuous distribution of the variables.

Step III. *Significance level:* Let $\alpha = 0.01$ for a two-tailed test.

Step IV. *Region of rejection:* From Expression (12.4), any observed value of $D \geq 0.277$ falls into the critical area of the sampling distribution.

Step V. *Calculations and statistical decision:* The computed value of $D = 0.072$ does not exceed the critical value, so the samples appear to favor H_0 at $\alpha = 0.01$.

Step VI. *Nonstatistical decision:* The experiment demonstrated no significant difference between the sampling methods for African societies.

The Kolmogorov–Smirnov test can also be phrased in one-tailed fashion, and the significance of such tests is determined by converting the D statistic to chi-square, distributed with 2 degrees of freedom,

$$\chi^2 = 4D^2 \frac{n_1 n_2}{n_1 + n_2} \tag{12.5}$$

The standard chi-square tables can then be used to determine the critical value of χ^2 for the directional Kolmogorov–Smirnov test. Keep in mind, however, that although D has been converted to a chi-square distribution, the Kolmogorov–Smirnov procedure examined a quite different relationship among variables than did the chi-square test considered in Chapter 11.

An application of the directional version of the Kolmogorov–Smirnov test involves cross-cultural comparisons of *riddles*. Riddling behavior has a rather uneven distribution throughout the world, and a recent study by Roberts and Forman (1971) attempted to account for this unusual distribution. Among other hypotheses considered, Roberts and Forman examined the relationship between the presence of riddles and the level of political organization. A cross-cultural survey was conducted in which societies were rated on the presence/absence of riddling, and political integration was ranked along a seven-step ordinal scale ranging from "lack of political integration" to the "state" level (Table 12.4). Is riddling associated with a high level of political integration?

The Kolmogorov-Smirnov statistic preserves the ordinal ranking in the level of political integration, yet handles a case which contains too many ties for the Wilcoxon test. For a directional application at $\alpha = 0.01$ (df = 2), the critical value of the chi-square statistic is found to be $\chi^2 = 9.210$ (Table A.5).

The two samples are presented in Table 12.4 along with their cumulative proportions. The maximum deviation between ranks is found in the second category, in which $D = 0.307$. The chi-square statistic conversion for this value is given by Formula (12.5):

$$\chi^2 = 4(0.307)^2 \frac{45(101)}{45 + 101} = 11.730$$

This observed value of χ^2 exceeds the critical value of 9.210; hence, H_o is rejected. Based upon these data, Roberts and Forman concluded that riddles appear to be associated with a high level of political organization.

TABLE 12.4 Cross-cultural comparison of riddling and the level of political integration (data from Roberts and Forman 1971).

Level of Political Integration	Riddles		
	+	−	D
Absent	0 (0.0)	9 (0.089)	−0.089
Autonomous local	8 (0.178)	40 (0.485)	−0.307
Peace groups	2 (0.222)	2 (0.505)	−0.283
Dependent	2 (0.267)	3 (0.535)	−0.268
Minimal state	10 (0.489)	16 (0.693)	−0.204
Little state	7 (0.644)	6 (0.752)	−0.108
State	16 (1.0)	25 (1.0)	0.0
	45	101	

The rationale behind the Kolmogorov–Smirnov statistic depends upon the continuous distribution function and is beyond the present scope. The interested reader is referred to Conover (1971: chapter 6).

Example 12.5

When administering final examinations, I have often wondered whether the better students tend to work more quickly or more slowly than the poorer students. A good case can be made for either position. I often advise students to stick with their first hunch; "You either know it or you don't," I sagely counsel. But I have also seen unprepared students simply scan an examination, make some cursory guesses, and leave the classroom prematurely. I once kept track of the order in which students turned in their exams in an introductory anthropology course. Do the better students (those receiving A, B, or C grades) work at different rates than the poorer students?

	Good Students		Poor Students		
Time, minutes	*f*	cum.	*f*	cum.	*D*
0–40	5	0.179	3	0.143	0.036
41–45	3	0.286	0	0.143	0.143
46–50	9	0.607	4	0.333	0.274
51–55	6	0.821	8	0.714	0.107
56–60	5	1.0	6	1.0	0.0
	28		21		

The largest deviation between the cumulative proportions is found between 46 and 50 minutes, so $D = 0.274$.

This is a two-tailed test and the 0.01 critical value of D is given by Expression (12.4):

$$D = 1.63\sqrt{\frac{28+21}{28(21)}} = 0.470$$

The computed value of D falls short of this critical value, so H_o is not rejected. These data indicate that there is apparently no relationship between the time a student spends on an exam and the grade received.

12.4 TWO RELATED SAMPLES

Both the Wilcoxon and Kolmogorov-Smirnov two-sample tests tacitly assume that each sample is selected independently. That is, the selection for variate i in the first sample can in no way influence selection of variate i in the second

sample. But many situations occur in which two variates are *paired* to one another, as considered earlier in Section 10.8. As long as the paired p-variates are normally distributed, and both variables are measurable on a metric scale, then the t-test can be used to test the relationships. But when these conditions are not fulfilled, the following nonparametric alternatives to the paired t-test can be helpful.

12.4.1 The Sign Test

The *sign test* is probably the least complicated member of the nonparametric family of statistical tests. As the name implies, this test considers only the direction (that is, the *sign*) of differences and ignores the magnitude of these differences. The sign tests involves a "paired" research design in which the variates have not been selected independently; rather they are grouped a priori into pairs by criteria such as before–after, male–female, left–right, first born–second born, and so forth. Because only the direction of difference is considered by the sign test, variables can be measured only on an ordinal scale. The only assumption of the sign test, aside from random selection, is that, regardless of the level of measurement, the variable must have an underlying continuous distribution. Thus, we assume that should ties occur between scores, these ties will have resulted from errors of measurement rather than from any inherent equalities within the actual phenomena; this is a common assumption for rank-order statistics. An anthropological example will illustrate the computations involved in the sign test.

Archaeologists frequently employ some rather loose analogies to the ethnographic record. Some well-documented modern primitives are often considered to be "living fossils," functioning analogies which can be used to infer practices of prehistoric technology, division of labor, economics, and kinship structures. But these analogies can never be strictly assumed without first establishing some firm relationships within the ethnographic record itself. Archaeologists occasionally assume, for instance, that females are responsible for the pottery of the archaeological sites. This assumption has led to some rather sophisticated attempts to study ceramic design elements as clues to prehistoric patterns of postmarital residence, inheritance, and even corporate lineality.

The ubiquity of female potters in the ethnographic record serves as an illustration of how the sign test simplifies statistical analysis. A random sample of 22 societies was selected from the *Ethnographic Atlas*. The statistical population was operationally limited to pottery-making societies within aboriginal North America. The *Atlas* codes this variable into the following categories:

F: Females alone perform the activity, male participation being negligible.
G: Both sexes participate, but females do appreciably more than males.
E: Equal participation by both sexes without marked or reported differentiation in specific tasks.
N: Both sexes participate, but males do appreciably more than females.
M: Males alone perform the activity, female participation being negligible.

This scale of measurement is ordinal when applied to pottery making, but the actual degree of participation is clearly a continuous phenomenon which could

be measured in metric fashion if the data were of high enough quality. But because we are interested only in whether males or females make the pottery the data can be reduced into simply dichotomous categories:

\+ females participate more than males (categories $F + G$)
− males participate more than females (categories $N + M$)

There is no significance to the symbols + and −, for any other set of binary symbols would have served equally well. These data are listed in Table 12.5 (note that had the original five-step ranking been retained, one of the two sample tests would have been appropriate). When cases of equal dependence (*Atlas* coding E) occur in the sign test, these cases are dropped, so the original sample size has been reduced from 22 to only 20 societies.

The results of the survey indicate that 16 of the 20 societies have female potters. We wish now to determine the probability that such extreme results could be due to mere chance association. If these results were to prove statistically significant, then the ethnographic analogy might well hold for archaeological cases as well. Alternatively, one could argue that the four societies with male potters represents too large a deviation from expectation. This phrasing should strike a familiar bell, since this precise situation was discussed earlier in connection with the binomial theorem (Chapter 6). In fact, *the sign test is no more than a nonparametric application of the binomial theorem.* The sign test involves dichotomous relationships, so it is clear that the arithmetic mean is not a suitable measure of central tendency. The null hypothesis holds that if the two categories are independent, then roughly half of the signs should be minus and half should be plus. Female potters have been designated as plus, so a prevalence of female potters should result in a positive value of the *median.* This proposition could be tested using the normal approximation to the binomial distribution, although $n = 20$ makes the approximation somewhat questionable. Operating at a significance level of 0.01, the region of rejection in this sign test becomes that area under the normal curve which represents the extreme cases of the plus sign. Thus, the region of rejection is set at $p \leq 0.01$.

TABLE 12.5 Twenty randomly selected North American societies. A plus sign denotes female potters (data from *Ethnographic Atlas*, Murdock 1967).

Society	Potter	Society	Potter
Nunivak	+	Zapotec	−
Baffinland	−	Cochiti	+
Yokuts	+	Ponca	+
S. Ute	+	Klamath	−
Shivwits	+	Sanpoil	+
Kaibab	−	White Knife	+
Walapai	+	Chemehuevi	+
Oto	+	Arikara	+
Hano	+	Hidatsa	+
Tewa	+	Mixe	+

If H_o is true, then the mean and standard deviation under the binomial should be

$$\mu = np = \frac{1}{2}(20) = 10$$

$$\sigma = \sqrt{npq} = \sqrt{5} = 2.24$$

The sample of 20 cases produced 16 societies with positive signs, so it becomes necessary to find the area under the normal curve to the right of 15.5, as shown in Fig. 12.1.

The appropriate value of z is

$$z = \frac{15.5 - 10.0}{2.24} = 2.45$$

From Table A.3, the associated probability value is found to be

$$p = 0.50 - 0.4929 = 0.0071$$

This probability falls well within the region of rejection, so the null hypothesis is rejected. At the 0.01 level, these findings are consistent with the hypothesis that females tend to be potters in North America. But since 20 percent of the sample societies had male potters, one cannot unequivocably assume that all prehistoric potters were in fact female. The analogy remains probabilistic.

To summarize the sign test:

Step I. *Statistical hypotheses:*

$$H_o: \quad p \leq q \qquad H_1: \quad p > q$$

Step II. *Statistical model:* The normal approximation to the binomial distribution is applied under the following assumptions: (1) the sampling is random, (2) the Bernoulli random variables are independent, (3) the measurement scale is at least ordinal.

Step III. *Level of significance:* Let $\alpha = 0.01$ for a directional test.

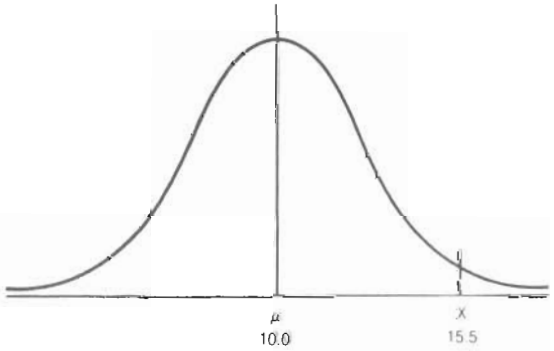

Fig. 12.1

Step IV. *Region of rejection:* The sign test is an exact test, so the critical probability value is defined directly by α. All probabilities ≤ 0.01 will reject H_o.

Step V. *Calculations and statistical decision:* The sample value was computed earlier to be $p = 0.0071$, so H_o is rejected.

Step VI. *Nonstatistical decision:* This sample is consistent with the hypothesis that prehistoric potters tended to be female.

Caution is in order when small samples are used in the sign test. When n is very small, the most extreme possible probability might still not fall within the region of rejection. Consider, for example, the case in which $n = 4$. The probability of obtaining all pluses is $(\frac{1}{2})^4 = 0.0625$. Of course this value will never exceed any conventional alpha level. The sample size must be at least larger than about 6, and larger values are desirable.

Because the sign test ignores the quantitative differences between variates, it is clear that the t-test utilizes more information and hence is more efficient. That is, the sign test will often fail to detect a difference which the t-test would have declared significant. Using the sign test thus increases the probability of a Type II error, and "power" is decreased. Thus, the parametric t-test is preferable *as long as the assumptions can be justified.* On the other hand, once results are judged significant by the sign test, these findings will generally be replicated by the more powerful tests.

The sign test is particularly well suited to cases involving *paired* variates. It is well known in social psychology, for example, that IQ scores can be strongly influenced by the environment during early childhood. Let us consider the hypothesis that urban-raised children will, on the average, fare better on IQ tests than children raised in a rural setting. To eliminate as many genetic (inherited) factors as possible, nine pairs of monozygotic twins were located; in each pair one twin was raised in the city, while the other was reared in rural conditions. The IQ scores are reproduced in Table 12.6.

A similar situation was investigated in Chapter 10 (Section 10.8) using the t-test for paired variates. But because of the vagaries of IQ testing, the investigator felt uneasy about the validity of using the raw IQ scores, so he elected to consider only the *absolute differences* between scores rather than consider the *magnitude*, which might be spurious. The data are now reduced to a form no longer applicable to the t-test.

The null hypothesis in this situation holds that the number of pluses and minuses should be roughly equal. This test is one-tailed because H_1 predicts that the city-reared children will receive superior IQ scores to the rural children. Alpha has been fixed at 0.05.

The data in Table 12.6 indicate that in six of the nine pairs, the urban-raised child received superior IQ scores. Are these findings significant enough to justify the research hypothesis?

If H_o is true, then the mean and standard deviations should be

$$\mu = np = \frac{1}{2}(9) = 4.5$$

$$\sigma = \sqrt{npq} = 1.5$$

TABLE 12.6 IQ scores from monozygotic twins raised in urban and rural environments.

Twin Code No.	IQ Scores		Differences	
	Rural	Urban	Raw	Sign
A	84	92	− 6	−
B	87	86	1	+
C	100	104	− 4	−
D	78	76	2	+
E	89	102	−13	−
F	96	92	4	+
G	115	123	− 8	−
H	108	112	− 4	−
I	72	76	− 4	−

Then z is found to be $(5.5-4.5)/1.5=0.67$ and Table A.3 indicates that $A=0.2486$. The resulting probability value thus is $p=0.2514$, a figure which is obviously not significant, and H_0 cannot be rejected. These results thus fail to support the research hypothesis that city and rural upbringing has a significantly positive effect upon IQ scores.

Example 12.6

In his monograph *Descendants of Immigrants* (1912), Franz Boas collected a wealth of comparative data, in addition to the stature information considered earlier. Boas was particularly interested in *hair color* because of its visibility and widespread significance as a racial indicator. Unfortunately, the modern techniques of physical anthropology for determining hair color (using standardized samples and reflectance spectrophotometry) were unavailable in 1908 when Boas was commissioned by the U.S. Immigration Commission to investigate the physical changes of immigrants. So Boas devised a method of measuring whereby the hair immediately over the forehead was ranked along an ordinal scale ranging from black to flaxen. Each color grade was assigned a number from 1 to 17. One set of Boas' data compared hair color of American-born and foreign-born Sicilian males (listed below) paired in age-graded classes.

Is there a significant difference in hair color?

This example raises some interesting problems of measurement. Because the hair-color categories consist of discrete ordered categories, the level of measurement is only ordinal, thereby negating use of the t-test for paired variates. The sign test allows comparison of two samples without assuming anything about the level of measurement, other than that Boas properly ranked hair color from dark to light.

Hair color of Sicilian males (data from Boas 1912: table IX).

Age Class, years	American-born	Foreign-born	Difference
5	10	9	+
6	10	12	−
7	10	10	0
8	11	9	+
9	8	8	0
10	8	10	−
11	8	9	−
12	8	8	0
13	7	8	−
14	8	8	0
15	7	7	0
16	5	8	−
17	6	6	0

The original set of 13 pairs is reduced to only $n = 7$, once the tied scores are removed. If H_o is true, the binomial parameters are

$$\mu = \frac{1}{2}(7) = 3.5$$

$$\sigma = \sqrt{7\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = 1.32$$

The standardized normal deviate is computed as usual.

$$z = \frac{2.5 - 3.5}{1.32} = -0.76$$

The corresponding area under the normal curve is $C = 0.2236$ and, because the test is two-tailed, $p = 2(C) = 0.4472$. Clearly, the null hypothesis remains inviolate. We conclude that Boas' data indicates no particular modification in hair color between American-born and foreign-born Sicilians.

12.4.2 The Wilcoxon Signed-Ranks Test

The sign test is useful when the assumptions of the paired *t*-test are untenable, but the sign test utilized only the directional relationships within a set of data, ignoring the magnitude of difference in every case. The *Wilcoxon Signed-Ranks Test* is a more powerful nonparametric tool which maintains the relative magnitude of difference between the ranked pair. The Wilcoxon method gives more weight to greater differences than to smaller ones, while the sign test records only which variate is larger, but not *how much* larger.

Let us examine the workings of the Wilcoxon test in another example

involving monozygotic twins. Some subtle physical differences sometimes are known to exist between first- and second-born twins; a first-born child, for instance, is often more dolichocephalic (roundheaded) than its twin, probably due to the fetal posture at birth. Some investigators likewise have noted a size difference between monozygotic twins. Since monozygotic twins are known to share an identical heritage and to be exactly of the same age, this difference in size can be ascribed only to some aspect of intrauterine conditions. An apparent factor seems to be the structure of the prenatal blood circulation because there is a "third circulation"—in addition to parental and fetal—by which blood actually passes from one twin to the other through the placenta. Other prenatal environmental factors could be due to variability in the uterine mucosa, and also variability in the size of placental vessels themselves. Any of these sources could result in one fetus receiving better nourishment than its twin.

With these factors in mind, it is possible to posit, following Gunnar Dahlberg, that the first-born twin tends to be larger than the second-born twin. Table 12.7 contains a sample of 16 sets of stature measurements collected by Dahlberg (1926). Do these data support the hypothesis that first-born twins tend to be larger?

The relations within these data could be tested by the *t*-test, as long as one is willing to assume an underlying normal distribution. But if we wish to handle these data in nonparametric fashion, then either the sign or Wilcoxon Signed-Ranks Test could be used. The Wilcoxon test is preferable here in order to preserve the magnitude of size differences between twins.

The one-tailed situation produces the following statistical hypotheses for the Wilcoxon Signed-Ranks test:

$$H_o: \text{ Median difference} \leq 0 \qquad H_1: \text{ Median difference} > 0$$

These hypotheses are comparable with those of the *t*- and sign tests, but each statistical test is based upon rather different assumptions and procedures.

The Wilcoxon test statistic is called T, defined as *the sum of differences of ranks with the least frequent sign*. Let X_i represent variates in the first sample and Y_i variates in the second sample. If fewer differences exist between $(X_i - Y_i)$, T is defined as the sum of these negative differences; otherwise the positive differences are summed to yield T.

The initial step in computing the Wilcoxon Signed-Ranks Test is to list the differences in stature within each pair, hence reducing the 16 pairs of variates in Table 12.7 to only 16 values of D. Whereas the sign test expressed these differences only in present/absent categories, the Wilcoxon paired test preserves the *ranks* of the differences (and, of course, the *t*-test preserves the actual quantitative magnitude). Two pairs of twins were exactly the same size, so these pairs are excluded from further consideration, and the sample size is reduced to $n = 14$. The absolute value of the remaining differences are then assigned rank orderings, with the smallest difference receiving the assigned rank of 1. Ties are handled as before by assigning the average of the tied ranks to each tied case.

Two different methods are available for determining the statistical significance of these pairwise ranks. As long as the number of cases are fewer than 50, Table A.7 can be used directly to define the critical region of rejection. In this

case, with $n = 14$, the critical level at the 0.05 level is found to be 25. This means that any observed sum of ranks less than or equal to 25 will be considered significant. Had this test been two-tailed, then the critical value of T would be found under $\alpha/2 = 0.025$; for $n = 14$, this critical value of $T = 21$.

The observed value of T is found in Table 12.7 to be 23, which is *less* than the critical value of $T = 25$; the critical number indicates "the maximum number of aberrant cases," so values less than or equal to the critical value are significant. H_o is rejected, and we conclude that Dahlberg's data on monozygotic twins are consistent with random fluctuations: that first-born twins do not appear to be significantly larger than the second-born.

Wilcoxon's T statistic is approximately distributed in normal fashion for samples of greater than about $n = 20$, so the results of the *Wilcoxon Signed Ranks Test* can be evaluated using a slightly modified version of the standardized normal deviate:

$$z = \frac{T - \mu}{\sigma} \tag{12.6}$$

where the parametric mean and standard deviation for $N = n$ cases are defined as

$$\mu = \frac{N(N+1)}{4}$$

$$\sigma = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

TABLE 12.7 Stature differences (in millimeters) between first- and second-born monozygotic twins (data from Dahlberg 1926: appendix I, table 1).

Stature, First Born X_i, mm	Stature, Second Born Y_i, mm	Difference $D = X_i - Y_i$	Rank $X_i > Y_i$	Rank $X_i < Y_i$
1014	1019	− 5		3.5
1186	1179	7	6	
1348	1334	14	10	
1357	1340	17	12	
1704	1709	− 5		3.5
1454	1434	20	13	
1592	1534	58	14	
1245	1261	−16		11
1380	1377	3	2	
1052	1058	− 6		5
1273	1262	11	9	
1426	1417	9	7.5	
1409	1400	9	7.5	
1253	1252	1	1	
1396	1396	0		
1219	1219	0		
			Σ of ranks = 82	$T = 23$

These parameters allow evaluation of the deviation of T relative to the familiar normal curve.

This large-sample method can be illustrated by returning to Boas' data on the physical changes of immigrants to the United States. These data were already analyzed by the paired version of the t-test (Example 10.8) in which a sample of American-born Bohemians were found to be significantly taller than those of foreign birth. The informants were paired in age grades for ages 4 through 20. These data are rank-ordered in Table 12.8, in which we find the value of the Wilcoxon statistic to be $T = 21$.

The parametric mean and standard deviation for a population of $N = n = 17$ pairs are

$$\mu = \frac{17(18)}{4} = 76.5$$

$$\sigma = \sqrt{\frac{17(18)(34+1)}{24}} = 21.12$$

The value of z can be computed from Expression (12.6) as

$$z = \frac{21 - 76.5}{21.12} = -2.63$$

The value of $z = -2.63$ corresponds to a probability figure of $p = 0.0043$ in Table A.3, a figure which is highly significant. H_0 is rejected and we conclude that American-born Bohemians seem to be notably taller than foreign-born Bohemians of the same racial stock. These findings agree with those of the t-test

TABLE 12.8 Comparison of stature between American- and foreign-born Bohemians (data from Boas 1912).

Age	American-born, cm	Foreign-born, cm	D	Ranks	
				$X_i > Y_i$	$X_i < Y_i$
4	99.4	98.0	+1.4	6	
5	105.7	101.0	+4.7	16	
6	110.7	110.6	+0.1	1	
7	116.0	111.7	+4.3	14.5	
8	122.5	118.2	+4.3	14.5	
9	128.5	128.1	+0.4	4	
10	132.7	135.1	−2.4		9.5
11	137.7	134.7	+3.0	12	
12	141.1	140.0	+1.1	5	
13	147.9	148.1	−0.2		2
14	152.3	150.4	+1.9	7	
15	155.5	155.2	+0.3	3	
16	162.7	160.7	+2.0	8	
17	167.6	165.0	+2.6	11	
18	175.0	167.7	+7.3	17	
19	171.2	167.0	+4.2	13	
20	168.6	171.0	−2.4		9.5
				Σ ranks = 132	$T = 21$

performed earlier, and the Wilcoxen Signed-Rank Test requires fewer procedural assumptions. Note further that since $n = 17$, Table A.7 could also have been used to compute an associated probability of $p < 0.005$.

12.5 THE KOLMOGOROV–SMIRNOV ONE-SAMPLE TEST

The Kolmogorov–Smirnov test compares observed and expected frequencies in a manner quite similar to the R × C chi-square test (Section 11.8). Both tests consider the "goodness of fit" between an expected distribution and the distribution of an actual random sample. The Kolmogorov–Smirnov test is preferable to χ^2 when the samples are small because the Kolmogorov–Smirnov method always provides an exact probability, regardless of n. Remember that chi-square assumes a sample size sufficient to satisfy the approximation to a continuous distribution of the χ^2 statistic.

This simple notion behind the Kolmogorov–Smirnov One-Sample Test can readily be illustrated using an archaeological example. Small quantities of Early Woodland (Black Sand phase) pottery sherds were found at the Macoupin site in the lower Illinois Valley (Rackerby 1973). Because the bulk of the cultural materials at Macoupin are Middle Woodland (Havana phase) in age, the excavators wanted to know whether these rare Early Woodland materials were associated with a particular stratigraphic level at the site or whether the aberrant sherds were simply strewn randomly throughout the site midden. The later Havana phase materials ran consistently from the surface to a depth of about 24 inches, while the Black Sand sherds seemed to concentrate in the upper levels: 0–6 inches, 11 sherds; 6–12 inches, 13 sherds; 12–18 inches, 11 sherds; 18–24 inches, 3 sherds. Can we justify the conclusion that the Black Sand sherds are uniformly distributed throughout the midden at the Macoupin site?

The Kolmogorov–Smirnov One-Sample Test can readily answer this question. The first step is to plot the observed sherd frequencies (labelled f in Table 12.9) by stratigraphic unit. Then the *cumulative* proportions of each stratigraphic unit are computed. The 0 to 6 inch level contained $11/38 \times 100 = 28.9$ percent of all the Black Sand sherds. The top *two* levels (0 to 6 and 6 to 12 inches) contained $(11 + 13)/38 \times 100 = 63.2$ percent of all these sherds, and so forth. Because we know that the Havana phase sherds were uniformly dispersed throughout the

TABLE 12.9 Stratigraphic placement of Black Sand pottery sherds at the Macoupin site (data from Rackerby 1973).

Stratigraphic Unit, inches	Frequency, f	Cumulative Frequency	Cumulative Proportion	Expected Proportion	Difference
0–6	11	11	11/38 = 0.289	0.250	0.039
6–12	13	24	24/38 = 0.632	0.500	0.132
12–18	11	35	35/38 = 0.921	0.750	0.171
18–24	3	38	38/38 = 1.0	1.0	0.0
	$n = 38$				

deposit, the null hypothesis states that the Black Sand should also distribute randomly throughout the levels of the site. So the expected cumulative proportion of the first stratum is 1/4 = 0.25, the expected cumulative proportion of the first two strata is 1/2 = 0.50, and so on. The main difference between this procedure and that of the familiar chi-square test is how the expected frequencies have been computed. The Kolomogorov–Smirnov method deals with expected cumulative proportions, while the χ^2 projects deal with expected absolute frequencies.

The final operation is to find the absolute differences between the observed and expected proportions (final column of Table 12.9). The Kolmogorov–Smirnov statistic, D, is merely the *maximum difference* between expected and observed proportions. In the example from the Macoupin site, D is found in the third row, the stratum consisting of 12 to 18 inches below the surface:

$$D = 0.171$$

The distribution of the Kolmogorov–Smirnov D statistic has been compiled in Table A.9. For the case of the 38 Black Sand potsherds, the critical value of D at the 0.05 level is

$$D = \frac{1.36}{\sqrt{38}} = 0.221$$

The observed D falls short of the critical level, so H_o is not rejected. Use of the Kolmogorov–Smirnov One-Sample Test allowed the excavator to conclude that "these data do not demonstrate stratigraphically that the Havana deposits are superimposed on the Black Sand deposit, but rather that there is considerable admixture of earlier material in later levels" (Rackerby 1973: 99).

Only the two-tailed version of the Kolmogorov–Smirnov One-Sample Test has been discussed here. The critical regions for the one-tailed option are poorly understood and have been omitted (see Siegel 1956:49) for appropriate references).

12.6 RUNS TEST

When a coin is tossed ten times, a "run" of ten heads is obviously a quite unlikely outcome.

HHHHHHHHHH

Whether this succession represents good or evil luck depends only on where one has placed his money, and Chapter 5 has considered methods to evaluate precisely the probabilities of such outcomes. But a second kind of departure from randomness has yet to be considered, a departure dealing only with *successions of events* rather than with their *relative frequency*. This section will consider a test to determine randomness in successive events.

Everyone has heard the riverboat gambler's expression "a run of luck." This sequence can consist of good luck—"Stick with me, baby, I can't lose!!"—or, more commonly, bad luck—"Somebody up there hates me." But in either case, a run of luck involves a sequence of events which deviates from expectation under randomness.

One possible outcome from tossing a coin ten times is

HHHHHTTTTT

In terms of strict frequency, the overall ratio 50:50 is the most likely outcome for a fair coin. But the occurrence of precisely five heads followed by exactly five tails is not a very likely event in terms of succession. In each of the ten independent trials, there are only two "runs": The sequence of five heads comprises the first run, followed by a second run of five tails. This is a very rare outcome. In fact, when we concentrate strictly upon sequence, there is only one more extreme outcome—a single run of all heads or all tails. These sequences have many fewer runs than are expected from chance phenomena.

At the other extreme, it is possible to have *too many* runs in a random sequence:

HTHTHTHTHT

The alternating head–tail sequence is a very unlikely event; in this case, a total of ten runs occurred. Clearly, the number of possible runs varies from one to *n* for any dichotomous variable. But the most likely number of runs is somewhere intermediate between the two extremes.

The *runs test* uses this simple concept of sequence to test for randomness. The more extreme (that is, the less frequent) the number of runs, the less likely it is that the sample is actually a random mix. The null hypothesis in this case is that the two dichotomous states are well mixed, that independent events should exhibit no tendency either to clump or to rigidly alternate. The computations of the runs tests can be illustrated by a simple example.

One particularly prolific family has spawned 12 children. While the frequencies are as expected, six boys and six girls, the order of birth seems rather odd, since the boys were born almost in sequence, followed by most of the female offspring.

MFMMMMMFFFFF

Does this sequence depart from randomness so far that we are entitled to question that the order of birth is randomly determined?

First it is necessary to find the total number of runs in the $(n_1 + n_2)$ births

M	F	MMMMM	FFFFF
1	2	3	4

There are only four runs in this sequence of 12 seemingly random events. The statistical dilemma is to decide whether four runs in 12 events is a rare event.

For small runs as this, the critical values have been compiled in Table A.16. Whenever an observed number of runs is less than or equal to the appropriate tabled value, then H_0 can be rejected at the 0.05 level. Strictly speaking, this is a test for *too few* runs, so the result is one-tailed. Tables for the alternative (*too many* runs) can be found in Siegel (1956).[2]

[2]As Blalock (1972: 252) has pointed out, this situation might cause some confusion unless care is taken with terminology. The runs test is one-tailed because we are considering only the possibility of *too few* runs. But, unlike most one-tailed tests, the direction has not been predicted, since either the variable *X* or *Y* could occur first. The sign test is a one-tailed test in which direction is not specified

Table A.10 indicates that for $n_1 = 6$ and $n_2 = 6$, the critical region for the 0.05 level is three or fewer runs. Because four runs were observed in the birth sequence above, H_o is not rejected at the 0.05 level. No doubt arises that the order of birth departs from a random sequence.

When either sample size exceeds 20, then Table A.10 is no longer applicable. But as the sample sizes increase, the distribution of r (the number of runs) approaches normality with mean and standard deviation as follows

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$
$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}} \tag{12.7}$$

where $n = n_1 + n_2$. This handy, yet slightly less cumbersome, computational procedure is illustrated in Example 12.7.

It is interesting to note that although the number of runs, r, approaches normality in the larger samples, the runs test remains nonparametric because the normal distribution still need not be assumed for the population of variates.

Example 12.7

The table below contains the annual rainfall tabulation for the period 1901–1950 for Sante Fe, New Mexico. If we operationally define "dry year" as one receiving 13 inches or less rainfall, can we say that wet and dry years appear to cluster (data from Schulman 1956: table 19A)?

Year	Rainfall, inches	Year	Rainfall, inches
1901	15.61	1920	18.56
1902	15.53	1921	14.37
1903	15.77	1922	13.67
1904	5.49	1923	10.75
1905	19.34	1924	13.63
1906	14.06	1925	8.14
1907	19.42	1926	15.83
1908	13.23	1927	13.20
1909	9.71	1928	14.70
1910	12.54	1929	13.60
1911	10.66	1930	17.14
1912	17.78	1931	15.47
1913	12.72	1932	16.90
1914	12.75	1933	14.23
1915	20.36	1934	12.88
1916	16.16	1935	13.71
1917	10.98	1936	12.32
1918	9.58	1937	19.48
1919	17.92	1938	11.49

Year	Rainfall, inches	Year	Rainfall, inches
1939	15.00	1946	11.30
1940	15.62	1947	14.17
1941	17.96	1948	16.06
1942	12.63	1949	15.41
1943	12.00	1950	12.31
1944	6.79	Mean	14.27
1945	13.03		

There are a total of 22 runs, with the number of dry years $n_1 = 17$ and the number of wet years $n_2 = 33$. The mean of the distribution of r is

$$\mu_r = \frac{2(17)(33)}{50} + 1 = 23.44$$

and the standard deviation is

$$\sigma_r = \sqrt{\frac{2(17)(33)(2\cdot 17\cdot 33 - 17 - 33)}{50^2(49)}}$$

$$= 3.12$$

The standardized deviate is thus

$$z = \frac{22 - 23.44}{3.12} = -0.46$$

This value is obviously not significant and we can conclude that these precipitation figures do not tend to cluster in wet and dry years.

12.7 SOME ASSUMPTIONS OF NONPARAMETRIC STATISTICS

The preceding nonparametric statistical techniques were considered in an almost negative fashion: The "Jones test for circular asymmetry" is useful because we don't have to assume X, Y, or even Z. But let us not be misled by the terms "nonparametric" or "distribution-free" to the erroneous conclusion that these techniques are somehow *nonassuming* or *assumption-free*. Nonparametric statistical tests make some very important assumptions which should not be ignored.

First of all, all nonparametric techniques of statistical inference assume that the sample was constructed through random sampling. Specifically, each element in the population must have had an *equal* and *independent* chance for selection.

In addition, many of the distribution-free tests involve comparing two samples, such as the chi-square test of a 2×2 table, Fisher's Exact Test, the Wilcoxon Signed-Ranks Test, and the Kolmogorov-Smirnov Two-Sample Test.

These two samples are assumed to be mutually independent in that the controls for selection of the first sample can in no way influence selection of the second sample. Violation of this independence can seriously change the computed levels of significance unless the test is specifically constructed to handle such dependence (as with the McNemar test).

Finally, the ordinal level tests assume that the underlying scales of measurement are actually continuous in nature. The units of observation are discrete because of the relative crude scales used for measurement: Either men or women make the pottery; either a group is heavily reliant upon fishing, or hardly dependent upon fishing, or they don't fish at all; either langur A dominates langur B or otherwise. Because of these crude categories of measurements, independent variates will occasionally be rated into the same category. The *Ethnographic Atlas*, for instance, rates both the Copper Eskimo and the Kaska as 36 to 45 percent dependent upon hunting. But even though these two societies are operationally considered to be "equal," we still must assume that there is really some slight difference which has simply gone undetected. This tie in ranking occurred because of our gross scale of measurement; given a suitably accurate measuring system, presumably we could detect a difference between the Copper Eskimo and the Kaska. Thus, in nonparametric testing, all ties are assumed to result from a gross system of measurement. So a moderate number of ties are permitted in the ordinal-level testing as long as ties are corrected by suitable formulas. Such is the conventional thinking about ties (for example, Siegel 1956).

But it must be mentioned that recent work on the problems of ties indicates that even when a high degree of agreement occurs in ordinal scales, there is almost no effect upon the computed level of significance (Conover 1971 and Noether 1972). Of course some nonparametric tests are more appropriate than others in the presence of ties. The Wilcoxon Two-Sample Test, for instance, becomes computationally undesirable as the ties increase, so one would do well to switch to an alternative rank-order test. The assumption of continuity is mentioned here only to warn prospective users that although one is commonly cautioned to assume an underlying continuity of measurement, such guidelines have little practical effect upon modern application of the rank-order statistics.

SUGGESTIONS FOR FURTHER READING

Blalock (1972: chapter 14)
Conover (1971: chapters 5, 6)
Siegel (1956)

EXERCISES

12.1 The male adults in two contiguous bands of hunter-gatherers were measured for stature (in centimeters):

Band A:	152	159	163	149	164
Band B:	156	167	169	155	172

Is there a significant difference in stature between these two bands (use the Wilcoxon Two-Sample Test)?

12.2 Five radiocarbon dates are available for each of two archaeological sites:

Site A:	A.D. 520	A.D. 490	A.D. 525	A.D. 540	A.D. 690
Site B:	A.D. 590	400 B.C.	A.D. 740	A.D. 730	A.D. 820

(a) Use the *t*-test to see whether site A is significantly older than site B.
(b) Use the Wilcoxon Two-Sample Test to test the same hypothesis.
(c) Which method is preferable? Why?

12.3 A census revealed the following mortality figures for two societies:

	Age at Death											
Society A:	62	54	78	56	45	58	64	63	63	34	53	45
Society B:	54	78	67	45	68	69	39	83	78	68	71	69

Does society B appear to be longer-lived than society A?

12.4 Two neighboring archaeological sites have been excavated and the projectile points from each analyzed.

Total Weight, grams	Site A	Site B
<1.0	2	0
1.0–1.9	15	6
2.0–2.9	10	13
3.0–3.9	3	6
4.0–4.9	5	5
5.0–5.9	2	6
≥6.0	0	3

In this area, points become lighter through time. Is site A later than site B? (Use the Kolmogorov-Smirnov Two-Sample Test.)

12.5 The following grades were assigned in a freshman anthropology course:

Grade	Total	Eventually Graduated	Did Not Graduate
A	35	28	7
B	130	82	48
C	90	62	28
D	52	38	14
F	18	3	15

Use the appropriate nonparametric method to determine whether those students with higher grades in this freshman course tended to graduate more frequently than those students receiving lower grades.

12.6 Fluted points are important time markers of the big-game hunting tradition in North America. McKenzie (1970) compiled the following data on the number of individual flutes per projectile points for two early types in Ohio.

Number of Flutes	Cumberland Points	Holcombe Points
0	11	10
1	22	8
2	4	0
3	1	4
4	0	0
>4	0	1

Is there a significant difference in the number of flutes between Cumberland and Holcombe points?

12.7 The following figures were obtained in a study of work habits among married couples.

	Hours Worked per Week	
Couple	Males	Females
A	39	41
B	51	42
C	23	35
D	45	39
E	67	54
F	39	43
G	42	46
H	51	51
I	32	36
J	40	42
K	56	53
L	41	41
M	43	45
N	37	39
O	40	41

Do these findings indicate that females tend to work more hours per week than males? (Use the sign test.)

12.8 Use the Wilcoxon Signed-Ranks Test to solve Exercise 12.7.