統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 18: Nonparametric Methods

上課時間地點: 二 D56, 資訊 140306
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

Overview

1. The statistical methods for inference presented previously are _____.
2. The parametric methods begin with an _____ about the probability distribution of the _____ which is often that the population has a _____ distribution.
3. Based upon this assumption, statisticians are able to derive the _____ that can be used to make _____ about one or more parameters of the population, such as the population mean or the population standard deviation.
   (a) (Recall Chapter 9) An inference about a population mean that was based on an assumption that the population had a normal probability distribution with unknown parameters μ and σ.
   (b) Using the sample standard deviation s to estimate the population standard deviation σ.
   (c) The test statistic for making an inference about the population mean was shown to have a t distribution.

(d) The t distribution was used to compute confidence intervals and conduct hypothesis tests about the mean of a normally distributed population.

4. In this chapter we present _____ methods which can be used to make inferences about a population without requiring an assumption about the specific form of the population's probability distribution.

   (a) (First section) how the binomial distribution uses two categories of data to make an inference about a _____.

   (b) (Next three sections) how _____ data are used in nonparametric tests about two or more populations.

   (c) (Final section) use rank-ordered data to compute the _____ for two variables.

5. For this reason, these nonparametric methods are also called _____.

6. The computations used in the nonparametric methods are generally done with _____. Whenever the data are quantitative, we will transform the data into categorical data in order to conduct the nonparametric test.

18.1 Sign Test

Hypothesis Test About a Population Median

1. The _____ provides a nonparametric procedure for testing a hypothesis about the value of a _____.

2. If we consider a population where _____ is exactly equal to the median, the median is the measure of _____ that divides the population so that _____ of the values are greater than the median and _____ of the values are less than the median.

3. Whenever a population distribution is _____, the median is often preferred over the mean as the best measure of central location for the population.

4. Example The weekly sales of Cape May Potato Chips by the Lawler Grocery Store chain.

   (a) Lawler's management made the decision to carry the new potato chip product based on the manufacture's estimate that the _____ should be \$450 per week on a per store basis.

   (b) (Table 18.1) After carrying the product for three-months, Lawler's management requested the following hypothesis test about the population median weekly sales:

$$H_0 \ : \ \text{Median} \ = 450$$
$$H_a \ : \ \text{Median} \ \neq 450$$

TABLE 18.1 One-Week Sales of Cape May Potato Chips at 10 Lawler Grocery Stores

Store Number	Weekly Sales ($)	Store Number	Weekly Sales ($)
56	485	63	474
19	562	39	662
36	415	84	380
128	860	102	515
12	426	44	721

   (c) (Table 18.2) In conducting the sign test, we compare each sample observation to the _____ of the population median.

      i. If the observation is greater than the hypothesized value, we record a plus sign _____

      ii. If the observation is less than the hypothesized value, we record a minus sign _____

      iii. If an observation is exactly equal to the hypothesized value, the observation is _____ from the sample and the analysis proceeds with the smaller sample size, using only the observations where a plus sign or a minus sign has been recorded.

TABLE 18.2 Lawler Sample Data for the Sign Test About the Population Median Weekly Sales

Store Number	Weekly Sales ($)	Sign	Store Number	Weekly Sales ($)	Sign
56	485	+	63	474	+
19	562	+	39	662	+
36	415	−	84	380	−
128	860	+	102	515	+
12	426	−	44	721	+

(d) Note that there are 7 plus signs and 3 minus signs.

5. The assigning of the plus signs and minus signs has made the situation a _____ _____ application. The sample size _____ is the number of trials. There are two outcomes possible per trial, a _____ sign or a _____ sign, and the trials are independent. Let _____ denote the probability of a plus sign.

6. If the population median is 450, p would equal _____ as there should be 50% plus signs and 50% minus signs in the population. Thus, in terms of the binomial probability p, the sign test hypotheses about the population median are converted to the following hypotheses about the binomial probability p.

$$\begin{aligned} H_0 &: \text{Median } = 450 \\ H_a &: \text{Median } \neq 450 \end{aligned} \quad \Rightarrow \quad \underline{\hspace{4cm}}$$

(a) If H_0 cannot be rejected, we cannot conclude that p is different from 0.50 and thus we cannot conclude that the population median is different from 450.

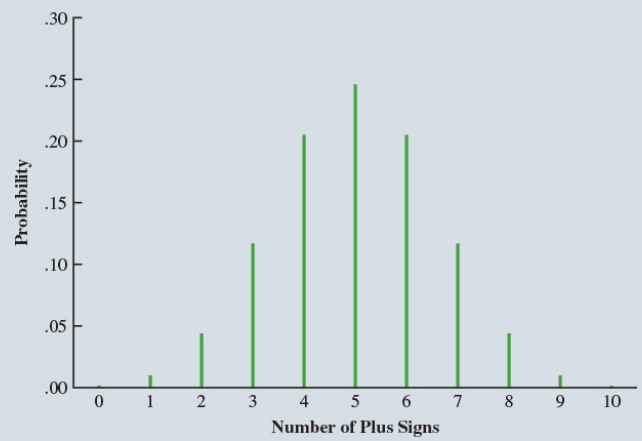
(b) If H_0 is rejected, we can conclude that p is not equal to 0.50 and thus the population median is not equal to 450.

7. (Table 5 in Appendix B)(Table 18.3)(Figure 18.1) With $n = 10$ stores or trials and $p = 0.50$, obtain the binomial probabilities for the number of plus signs under the assumption H_0 is true. (_____)

TABLE 18.3 Binomial Probabilities with $n = 10$ and $p = .50$

Number of Plus Signs	Probability
0	.0010
1	.0098
2	.0439
3	.1172
4	.2051
5	.2461
6	.2051
7	.1172
8	.0439
9	.0098
10	.0010

FIGURE 18.1 Binomial Sampling Distribution for the Number of Plus Signs When $n = 10$ and $p = .50$



(a) Use a 0.10 level of significance for the test.

(b) Since the observed number of plus signs for the sample data, 7, is in the upper tail of the binomial distribution, we compute the probability of obtaining 7 or more plus signs

_____.

(c) Since we are using a two-tailed hypothesis test, this upper tail probability is doubled to obtain the _____.

(d) With _____, we cannot reject H_0. In terms of the binomial probability p, we cannot reject $H_0 : p = 0.50$, and thus we cannot reject the hypothesis that the population median is \$450.

8. The one-tailed sign tests about a population median:

   (a) Formulated the hypotheses as an _____:

$$\begin{aligned} H_0 &: \text{Median} \leq 450 \\ H_a &: \text{Median} > 450 \end{aligned}$$

   (b) The corresponding p-value is equal to the binomial probability that the number of plus signs is _____ found in the sample.

   (c) This one-tailed p-value: _____.

   (d) If the example were converted to a lower tail test, the p-value would have been the probability of obtaining 7 or fewer plus signs.

   (e) The binomial probabilities provided in Table 5 of Appendix B can be used to compute the p-value when the sample size is _____.

   (f) With larger sample sizes, we rely on the _____ of the binomial distribution to compute the p-value; this makes the computations quicker and easier.

Use the Normal Distribution to Approximate the Binomial Probability

1. Example One year ago the median price of a new home was \$236,000. However, a current downturn in the economy has real estate firms using sample data on recent home sales to determine if the population median price of a new home is less today than it was a year ago.

   (a) The hypothesis test about the population median price of a new home is as follows:

$$\begin{aligned} H_0 &: \_\_\_\_\_ \\ H_a &: \_\_\_\_\_ \end{aligned}$$

   (b) We will use a 0.05 level of significance to conduct this test. A random sample of _____ recent new home sales found _____ homes sold for more than \$236,000, _____ homes sold for less than \$236,000, and _____ home sold for \$236,000.

(c) After deleting the home that sold for the hypothesized median price of \$236,000, the sign test continues with 22 plus signs, 38 minus signs, and a sample of _____.

(d) The null hypothesis that the population median is greater than or equal to \$236,000 is expressed by the binomial distribution hypothesis _____.

(e) If H_0 were true as an equality, we would expect _____ homes to have a plus sign.

(f) The sample result showing 22 plus signs is in the lower tail of the binomial distribution. Thus, the p-value is the probability of _____ when $p = 0.50$.

(g) While it is possible to compute the exact binomial probabilities for $0, 1, 2, \cdots$ to 22 and sum these probabilities, we will use the normal distribution approximation of the binomial distribution to make this computation easier.

2. **Normal approximation of the sampling distribution of the number of plus signs when $H_0 : p = 0.50$:** For this approximation (_____), the mean and standard deviation of the normal distribution are:

$$\text{Mean} : \mu = \underline{\qquad\qquad} \tag{18.1}$$

$$\text{Standard deviation} : \sigma = \underline{\qquad\qquad\qquad\qquad} \tag{18.2}$$

3. With $n = 60$ homes and $p = 0.50$, the sampling distribution of the number of plus signs can be approximated by a normal distribution with

$$\mu = 0.50n = \underline{\qquad\qquad} \qquad \sigma = \sqrt{0.25n} = \underline{\qquad\qquad\qquad}$$

4. The binomial probability distribution is discrete and the normal probability distribution is continuous. To account for this, the binomial probability of 22 is computed by the normal probability interval _____. The 0.5 added to and subtracted from 22 is called the _____ factor.

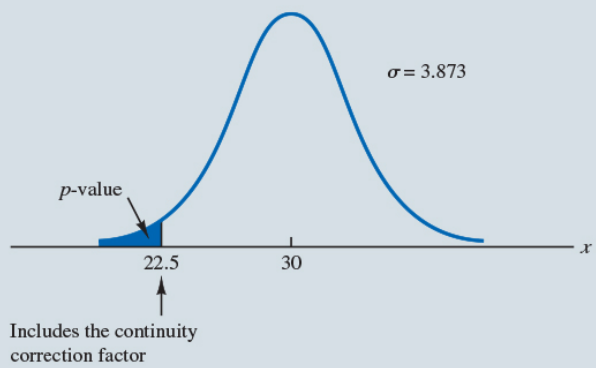
5. Thus, to compute the p-value for 22 or fewer plus signs we use the normal distribution with $\mu = 30$ and $\sigma = 3.873$ to compute the probability that the normal random variable, X, has a value less than or equal to 22.5.

6. (Figure 18.2) Using this normal distribution, we compute the p-value as follows:

   p-value = _____

7. With $0.0262 < 0.05$, we _____ the null hypothesis and conclude that the median price of a new home is _____ the \$236,000 median price a year ago.

FIGURE 18.2 Normal Distribution Approximation of the p-Value for the Sign Test About the Median Price of New Homes



Hypothesis Test with Matched Samples

1. (Recall Chapter 10) Using _____ and assuming that the differences between the pairs of matched observations were _____ distributed, the _____ distribution was used to make an inference about the difference between the means of the two populations.

2. Use the nonparametric sign test to analyze _____ data. the sign test enables us to analyze categorical as well as quantitative data and requires no assumption about the distribution of the differences.

3. This type of matched-sample design occurs in _____ when a sample of n potential customers is asked to compare two brands of a product such as coffee, soft drinks, or detergents. Without obtaining a quantitative measure of each individual's preference for the brands, each individual is asked to state a brand preference.

4. Example Sun Coast Farms produces an orange juice product called Citrus Valley. The primary competition for Citrus Valley comes from the producer of an orange juice known as Tropical Orange. In a consumer preference comparison of the two brands, 14 individuals were given unmarked samples of the two orange juice products. The brand each individual tasted first was selected randomly.

   (a) If the individual selected Citrus Valley as the more preferred, a _____ was recorded.

   (b) If the individual selected Tropical Orange as the more preferred, a _____ was recorded.

   (c) If the individual was unable to express a difference in preference for the two products, _____ was recorded.

5. (Table 18.4) Deleting the two individuals who could not express a preference for either brand, the data have been converted to a sign test with _____ signs and _____ signs for the _____ individuals who could express a preference for one of the two brands.

TABLE 18.4 Preference Data for the Sun Coast Farms Taste Test

Individual	Preference	Sign	Individual	Preference	Sign
1	Tropical Orange	−	8	Tropical Orange	−
2	Tropical Orange	−	9	Tropical Orange	−
3	Citrus Valley	+	10	No Preference	
4	Tropical Orange	−	11	Tropical Orange	−
5	Tropical Orange	−	12	Citrus Valley	+
6	No Preference		13	Tropical Orange	−
7	Tropical Orange	−	14	Tropical Orange	−

6. Letting _____ indicate the proportion of the population of customers who prefer Citrus Valley orange juice, we want to test the hypotheses that there is no difference between the preferences for the two brands as follows:

$$H_0 \ : \ \_\_\_\_\_$$
$$H_a \ : \ \_\_\_\_\_$$

7. If H_0 cannot be rejected, we cannot conclude that there is a difference in preference for the two brands. However, if H_0 can be rejected, we can conclude that the consumer preferences differ for the two brands.

8. (Table 18.5) We will conduct the sign test ($\alpha = 0.05$). The sampling distribution for the number of plus signs is a _____ distribution with $p = 0.50$ and $n = 12$. (_____)

TABLE 18.5 Binomial Probabilities with $n = 12$ and $p = .50$

Number of Plus Signs	Probability
0	.0002
1	.0029
2	.0161
3	.0537
4	.1208
5	.1934
6	.2256
7	.1934
8	.1208
9	.0537
10	.0161
11	.0029
12	.0002

9. Under the assumption H_0 is true, we would expect _____ plus signs. With only two plus signs in the sample, the results are in the _____ of the binomial distribution.

10. To compute the p-value for this two-tailed test, we first compute the probability of 2 or fewer plus signs and then _____ this value. Using the binomial probabilities of 0, 1, and 2 shown in Table 18.5, the p-value is

p-value = _____

11. We reject H_0. The taste test provides evidence that consumer preference _____ for the two brands of orange juice. We would advise Sun Coast Farms of this result and conclude that the competitor's Tropical Orange product is the more preferred. Sun Coast Farms can then pursue a strategy to address this issue.

12. Similar to other uses of the sign test, one-tailed tests may be used depending upon the application.

13. As the sample size becomes large, the _____ of the binomial distribution will ease the computation.

14. While the Sun Coast Farms sign test for matched samples used categorical preference data, the sign test for matched samples can be used with _____ data as well.

    (a) This would be particularly helpful if the _____ are _____ distributed and are _____.

    (b) In this case a positive difference is assigned a plus sign, a negative difference is assigned a negative sign, and a zero difference is removed from the sample.

    (c) The sign test computations proceed as before.

😍 **EXERCISES 18.1**: 1, 3, 6, 9

18.2 Wilcoxon Signed-Rank Test

1. (Recall Chapter 10) The parametric test for the _____ experiment requires quantitative data and the assumption that the _____ between the paired observations are normally distributed. The _____ can then be used to make an inference about the difference between the means of the two populations.

2. The _____ test is a nonparametric procedure for analyzing data from a _____. The test uses _____ but does not require the assumption that the differences between the paired observations are normally distributed.

3. It only requires the assumption that the _____ between the paired observations have a _____ distribution.

4. This occurs whenever the _____ of the two populations are the same and the focus is on determining if there is a difference between the _____ of the two populations.

5. Example Production Task Completion Times: Consider a manufacturing firm that is attempting to determine whether two production methods differ in terms of task completion time.

   (a) (Table 18.6) Using a matched-samples experimental design, 11 randomly selected workers completed the production task two times, once using method A and once using method B. The production method that the worker used first was randomly selected.

TABLE 18.6 Production Task Completion Times (Minutes)

	Method		
Worker	A	B	Difference
1	10.2	9.5	.7
2	9.6	9.8	−.2
3	9.2	8.8	.4
4	10.6	10.1	.5
5	9.9	10.3	−.4
6	10.2	9.3	.9
7	10.6	10.5	.1
8	10.0	10.0	.0
9	11.2	10.6	.6
10	10.7	10.2	.5
11	10.6	9.8	.8

   (b) A _____ difference indicates that method A required more time; a _____ difference indicates that method B required more time.

   (c) Do the data indicate that the two production methods differ significantly in terms of completion times? If we assume that the differences have a _____ distribution but not necessarily a normal distribution, the Wilcoxon signed-rank test applies.

   (d) In particular, we will use the Wilcoxon signed-rank test for the difference between the _____ completion times for the two production methods.

$$H_0 \ : \ \underline{\hspace{6cm}}$$

$$H_a \ : \ \text{Median for method A} - \text{Median for method B} \neq 0$$

   (e) If H_0 cannot be rejected, we will not be able to conclude that the median completion times are different. However, if H_0 is rejected, we will conclude that the median completion times are different.

6. The Wilcoxon signed-rank test steps ($\alpha = 0.05$):

   (a) Discard the difference of _____ for worker 8 and then compute the _____ _____ for the remaining 10 workers.

   (b) Rank these absolute differences from _____. The first (second) smallest absolute difference of 0.1 (0.2) for worker 7 (2) is assigned the rank of 1 (2). This ranking of absolute differences continues with the largest absolute difference of 0.9 for worker 6 being assigned the rank of 10. The _____ absolute differences of 0.4 (0.5) for workers 3 and 5 (4 and 10) are assigned the _____ of 3.5 (5.5).

   (c) (Table 18.7) Each rank is given the _____ of the original difference for the worker.

TABLE 18.7 Ranking the Absolute Differences and the Signed Ranks for the Production Task Completion Times

Worker	Difference	Absolute Difference	Rank	Signed Ranks Negative	Signed Ranks Positive
1	.7	.7	8		8
2	−.2	.2	2	−2	
3	.4	.4	3.5		3.5
4	.5	.5	5.5		5.5
5	−.4	.4	3.5	−3.5	
6	.9	.9	10		10
7	.1	.1	1		1
8	.0				
9	.6	.6	7		7
10	.5	.5	5.5		5.5
11	.8	.8	9		9
			Sum of Positive Signed Ranks		$T^+ = 49.5$

   (d) Let _____ denote the sum of the positive signed ranks ($T^+ = 49.5$). We will use T^+ as the Wilcoxon signed-rank test statistic.

   (e) **Sampling Distribution of T^+ for the Wilcoxon Signed-Rank Test**: If the medians of the two populations are equal and the number of matched pairs is 10 or more, the sampling distribution of T^+ can be approximated by a _____:

$$\text{Mean}: \mu_{T^+} = \underline{\qquad\qquad} \tag{18.3}$$

$$\text{Standard deviation}: \sigma_{T^+} = \underline{\qquad\qquad\qquad\qquad} \tag{18.4}$$

Distribution Form: Approximately normal for _____.

7. Example Production Task Completion Times:

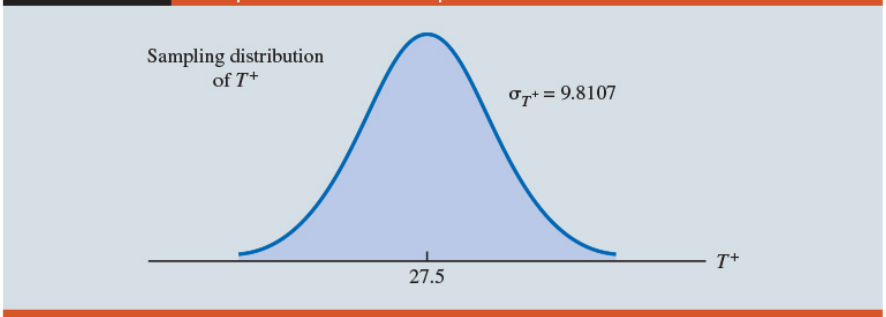
(a) After discarding the observation of a zero difference for worker 8, the analysis continues with the $n = 10$ matched pairs.

$$\mu_{T^+} = \frac{n(n+1)}{4} = \underline{\qquad\qquad} = 27.5$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad} = 9.8107$$

(b) (Figure 18.3) The sampling distribution of the T^+ test statistic.

FIGURE 18.3 Sampling Distribution of T^+ for the Production Task Completion Time Example



(c) Compute the two-tailed p-value for the hypothesis that the median completion times for the two production methods are equal. Since the test statistic $T^+ = 49.5$ is in the _____ of the sampling distribution, we begin by computing the upper tail probability _____.

(d) Since the sum of the positive ranks T^+ is discrete and the normal distribution is continuous, we will obtain the best approximation by including the _____ factor. Thus, the discrete probability of _____ is approximated by the normal probability interval, _____, and the probability that $T^+ \geq 49.5$ is approximated by:

$$P(T^+ \geq 49.5) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

(e) Using the standard normal distribution table and $z = 2.19$, we see that the two-tailed p-value = _____. With the p-value ≤ 0.05, we reject H_0 and conclude that the median completion times for the two production methods are not equal.

(f) With T^+ being in the upper tail of the sampling distribution, we see that method A led to the longer completion times. We would expect management to conclude that method B is the faster or better production method.

8. **One-tailed Wilcoxon signed-rank tests** are possible. For example, if initially we had been looking for statistical evidence to conclude method A had the larger median completion time than method B:

$$H_0 \ : \ \underline{\hspace{6cm}}$$
$$H_a \ : \ \text{Median for method A} - \text{Median for method B} > 0$$

9. (Recall Section 18.1) the sign test could be used for both a hypothesis test about a population median and a hypothesis test with matched samples.

10. The Wilcoxon signed-rank test can also be used for a nonparametric test about a _____. This test makes no assumption about the population distribution other than that it is _____.

11. If this symmetric assumption is appropriate, the Wilcoxon signed-rank test is the preferred nonparametric test for a population median. However, if the population is _____, the sign test is preferred.

12. With the Wilcoxon signed-rank test, the differences between the _____ and the _____ of the population median are used instead of the differences between the matched-pair observations.

13. NOTES+COMMENTS:

(a) The Wilcoxon signed-rank test for a population median is based on the assumption that the population is symmetric. With this assumption, the population _____ is equal to the population _____. Thus, the Wilcoxon signed-rank test can also be used as a test about the _____.

(b) There are several variations of the Wilcoxon signed-rank test that generally provide similar but not identical results. The test we use in section 18.2 is based on a _____ (which is much easier to calculate).

(c) JMP uses the exact Wilcoxon signed-rank test when $n \leq 20$ and a Student's t approximation when $n > 20$.

😍 **EXERCISES 18.2**: 12, 15, 17

18.3 Mann-Whitney-Wilcoxon (MWW) Test

1. (Recall Chapter 10) A hypothesis test (t-test) about the difference between the means of two populations using two independent samples:

   (a) This parametric test required _____ data and the assumption that both populations had a _____ distribution.

   (b) If population standard deviations σ_1 and σ_2 were unknown, the sample standard deviations s_1 and s_2 provided estimates of σ_1 and σ_2.

   (c) The t distribution was used to make an _____ about the difference between the means of the two populations.

2. We present a nonparametric test for the difference between two populations based on two independent samples. It can be used with either _____ data or _____ data and it does not require the assumption that the populations have a normal distribution.

3. Versions of the test were developed jointly by Mann and Whitney and also by Wilcoxon. As a result, the test has been referred to as the _____ and the _____. The tests are equivalent and both versions provide the same conclusion. We will refer to this nonparametric test as

the _____ test (e.g., a two-tailed test):

H_0 : The two populations are _____

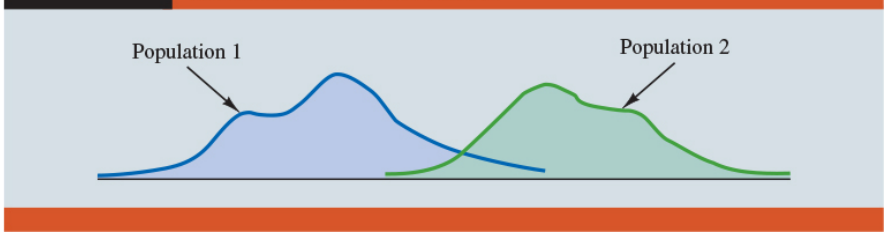
H_a : The two populations are not identical

(i.e., either population may provide the smaller or larger values.)

4. If H_0 is rejected, we are using the test to conclude that the populations are not identical and that population 1 tends to provide either _____ values than population 2.

5. (Figure 18.4) A situation where population 1 tends to provide smaller values than population 2. (Note that it is not necessary that all values from population 1 be less than all values from population 2.)

FIGURE 18.4 Two Populations Are Not Identical with Population 1 Tending to Provide the Smaller Values



6. First illustrate the MWW test using _____ with _____. Later, we will introduce a _____ approximation based on the _____ distribution that will simplify the calculations required by the MWW test.

7. Example Consider the on-the-job performance ratings for employees at a Showtime Cinemas 20-screen multiplex movie theater.

   (a) During an employee performance review, the theater manager rated all 35 employees from best (rating 1) to worst (rating 35) in the theater's annual report. Knowing that the part-time employees were primarily college and high school students, the district manager asked if there was evidence of a significant difference in performance for college students compared to high school students.

   (b) In terms of the population of college students and the population of high school students who could be considered for employment at the theater, the

hypotheses were:

H_0 : College and high school student populations are identical in terms of performance

H_a : College and high school student populations are not identical in terms of performance

(c) (Table 18.8) The theater manager's overall performance rating based on all 35 employees was recorded for each of these employees.

TABLE 18.8 Performance Ratings for a Sample of College Students and a Sample of High School Students Working at Showtime Cinemas

College Student	Manager's Performance Rating	High School Student	Manager's Performance Rating
1	15	1	18
2	3	2	20
3	23	3	32
4	8	4	9
		5	25

(d) (Table 18.9)(**The combined-sample ranks**) Use a 0.05 level of significance for this test and _____ the combined samples _____.

TABLE 18.9 Ranks for the Nine Students in the Showtime Cinemas Combined Samples

College Student	Manager's Performance Rating	Rank	High School Student	Manager's Performance Rating	Rank
1	15	4	1	18	5
2	3	1	2	20	6
3	23	7	3	32	9
4	8	2	4	9	3
	Sum of Ranks	14	5	25	8
				Sum of Ranks	31

(e) **Sum the ranks** for each sample as shown in Table 18.9. The sum of ranks for the first sample will be the test statistic W for the MWW test: $W = 4 + 1 + 7 + 2 = 14$.

(f) We will always follow the procedure of using the sum of the ranks for _____ as the _____.

8. Why the sum of the ranks will help us select between the two hypotheses: H_0: The two populations are identical and H_a: The two populations are not identical.

   (a) Letting C denote a college student and H denote a high school student, suppose the ranks of the nine students had the following order with the four college students having the four lowest ranks.

Rank	1	2	3	4	5	6	7	8	9
Student	C	C	C	C	H	H	H	H	H

   (b) Notice that this permutation or ordering separates the two samples, with the college students all having a _____ than the high school students.

   (c) This is a strong indication that the two populations are _____. The sum of ranks for the college students in this case is _____.

   (d) Now consider a ranking where the four college students have the four highest ranks.

Rank	1	2	3	4	5	6	7	8	9
Student	H	H	H	H	H	C	C	C	C

   This is another strong indication that the two populations are not identical. The sum of ranks for the college students in this case is _____.

   (e) Thus, we see that the _____ for the college students must be between 10 and 30. Values of _____ imply that college students have lower ranks than the high school students, whereas values of _____ imply that college students have higher ranks than the high school students.

   (f) Either of these extremes would signal the two populations are not identical. However, if the two populations are identical, we would expect a _____ in the ordering of the C's and H's so that the sum of ranks W is closer to the _____ of the two extremes, or nearer to _____.

9. (Figure 18.5)(Table 18.10) Making the assumption that the two populations are identical, we used a computer program to compute _____ for the nine students. For each ordering, we computed the _____ for the college students. This provided the probability distribution showing the exact sampling distribution of W.

FIGURE 18.5 Exact Sampling Distribution of the Sum of the Ranks for the Sample of College Students

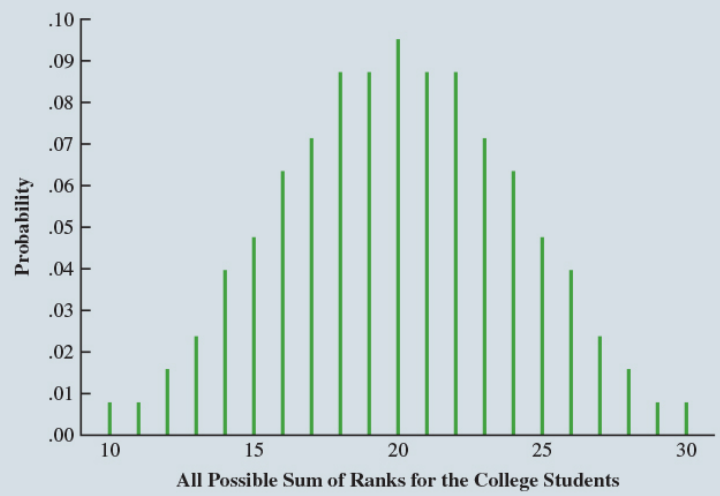


TABLE 18.10 Probabilities for the Exact Sampling Distribution of the Sum of the Ranks for the Sample of College Students

W	Probability	W	Probability
10	.0079	20	.0952
11	.0079	21	.0873
12	.0159	22	.0873
13	.0238	23	.0714
14	.0397	24	.0635
15	.0476	25	.0476
16	.0635	26	.0397
17	.0714	27	.0238
18	.0873	28	.0159
19	.0873	29	.0079
		30	.0079

10. Use the sampling distribution of W in Figure 18.5 to compute the p-value for the test. Table 18.9 shows that the sum of ranks for the four college student is _____. Because this value of W is in the _____ of the sampling distribution, we begin by computing the lower tail probability _____:

$$\begin{aligned} P(W \leq 14) &= P(10) + P(11) + P(12) + P(13) + P(14) \\ &= 0.0079 + 0.0079 + 0.0159 + 0.0238 + 0.0397 = 0.0952 \end{aligned}$$

11. The two-tailed p-value _____. With $\alpha = 0.05$ as the level of significance and p-value > 0.05, the MWW test conclusion is that we cannot reject

the null hypothesis that the populations of college and high school students are identical.

12. Use the same combined-sample ranking procedure and use the _____ distribution approximation of W to compute the p-value and draw the conclusion.

13. Example Third National Bank.

    (a) The bank manager is monitoring the balances maintained in checking accounts at two branch banks and is wondering if the populations of account balances at the two branch banks are identical.

    (b) (Table 18.11) Two independent samples of checking accounts are taken with sample sizes $n_1 = 12$ at branch 1 and $n_2 = 10$

TABLE 18.11 Account Balances for Two Branches of Third National Bank

Branch 1		Branch 2	
Account	Balance ($)	Account	Balance ($)
1	1095	1	885
2	955	2	850
3	1200	3	915
4	1195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1000
9	875	9	1050
10	1055	10	935
11	1025		
12	975		

    (c) (Table 18.12) The first step in the MWW test is to rank the combined data from the lowest to highest values. In that case of the two or more values are the same, the tied values are assigned the average rank of their positions in the combined data set.

TABLE 18.12 Assigned Ranks for the Combined Account Balance Samples

Branch	Account	Balance	Rank
2	6	750	1
2	5	800	2
1	7	805	3
2	2	850	4
2	7	865	5
1	9	875	6
2	1	885	7
2	3	915	8
1	5	925	9
2	10	935	10
1	8	945	11
1	6	950	12.5
2	4	950	12.5
1	2	955	14
1	12	975	15
2	8	1000	16
1	11	1025	17
2	9	1050	18
1	10	1055	19
1	1	1095	20
1	4	1195	21
1	3	1200	22

(d) (Table 18.13) The next step is to sum the ranks for each sample: 169.5 for sample 1 and 83.5 for sample 2 are shown. Thus, we have $W = 169.5$. When both samples sizes are _____, a normal approximation of the sampling distribution of W can be used.

TABLE 18.13 Combined Ranking of the Data in the Two Samples from Third National Bank

Branch 1			Branch 2		
Account	Balance ($)	Rank	Account	Balance ($)	Rank
1	1095	20	1	885	7
2	955	14	2	850	4
3	1200	22	3	915	8
4	1195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1000	16
9	875	6	9	1050	18
10	1055	19	10	935	10
11	1025	17		Sum of Ranks	83.5
12	975	15			
	Sum of Ranks	169.5			

14. Under the assumption that the null hypothesis is true and the populations are identical, the sampling distribution of the test statistic W is:

$$\text{Mean}: \underline{\qquad\qquad\qquad\qquad} \tag{18.5}$$

Standard deviation : _____ (18.6)

Distribution form: Approximately normal provided _____.

15. Since the test statistic W is discrete and the normal distribution is continuous, we will again use the _____ factor for the normal distribution approximation.

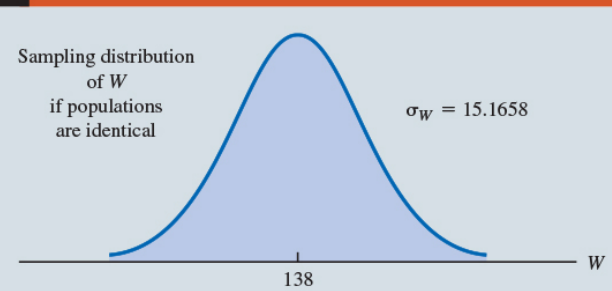
16. Example Third National Bank.

    (a) (Figure 18.6) Given the sample sizes _____, equations (18.5) and (18.6) provide the following mean and standard deviation for the sampling distribution:

$$\text{Mean} : \mu_W = (1/2)(12)(12 + 10 + 1) = 138$$

$$\text{Standard deviation} : \sigma_W = \sqrt{(1/12)(12)(10)(12 + 10 + 1)} = 15.1658$$

FIGURE 18.6 Sampling Distribution of W for the Third National Bank Example



    (b) With $W = 169.5$ in the _____ of the sampling distribution, we have the following p-value calculation:

$$P(W \geq 169.5) = \underline{\qquad\qquad\qquad\qquad}$$

    (c) Using the standard normal random variable and $z = 2.04$, the two-tailed p-value _____. With p-value ≤ 0.05, _____ and conclude that the two populations of account balances are not identical. The upper tail value for test statistic W indicates that the population of account balances at branch 1 tends to be _____.

17. Some applications of the MWW test make it appropriate to assume that the two populations have _____ and if the populations differ, it is only by a _____ in the location of the distributions.

18. If the two populations have the _____, the hypothesis test may be stated in terms of the difference between the two _____. Any difference between the medians can be interpreted as the shift in location of one population compared to the other. In this case, the three forms of the MWW test about the medians ($M_i, i = 1, 2$) of the two populations are as follows:

Two-Tailed Test	Lower Tail Test	Upper Tail Test
H_0: $M_1 - M_2 = 0$	H_0: $M_1 - M_2 \geq 0$	H_0: $M_1 - M_2 \leq 0$
H_a: $M_1 - M_2 \neq 0$	H_a: $M_1 - M_2 < 0$	H_a: $M_1 - M_2 > 0$

😍 **EXERCISES 18.3**: 18, 21

18.4 Kruskal-Wallis Test

1. (Recall Chapter 13, ANOVA) We considered a parametric test for three or more populations when we used _____ and assumed that the populations had normal distributions with the same standard deviations. Based on an independent random sample from each population, we used the _____ to test for differences among the _____.

2. The nonparametric _____ is based on the analysis of independent random samples from each of _____ populations. This procedure can be used with either _____ data or _____ data and does not require the assumption that the populations have normal distributions:

$$H_0 : \text{All populations are } \_\_\_\_\_$$
$$H_a : \text{Not all populations are identical}$$

3. If H_0 is rejected, we will conclude that there is a difference among the populations with one or more populations tending to provide _____ values compared to the other populations.

4. Example Performance Evaluation Ratings for 20 Williams Employees

   (a) (Table 18.14) Williams Manufacturing Company hires employees for its management staff from three different colleges. Recently, the company's personnel director began reviewing the annual performance reports for the management staff in an attempt to determine whether there are differences in the performance ratings among the managers who graduated from the three colleges. The performance rating shown for each manager is recorded on a scale from 0 to 100, with 100 being the highest possible rating.

TABLE 18.14 Performance Evaluation Ratings for 20 Williams Employees

College A	College B	College C
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

   (b) Suppose we want to test whether the three populations of managers are identical in terms of _____. We will use a 0.05 level of significance for the test.

   (c) (Table 18.15) The first step in the Kruskal-Wallis procedure is to _____ _____ from lowest to highest values. Note that we assigned the average ranks to tied performance ratings of 60, 70, 80, and 90.

TABLE 18.15 Combined Rankings for the Three Samples

College A	Rank	College B	Rank	College C	Rank
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5	Sum of Ranks	27	75	14
Sum of Ranks	95			Sum of Ranks	88

5. **The Kruskal-Wallis test statistic**:

$$\underline{\hspace{6cm}} \tag{18.7}$$

where

- k = the number of populations
- n_i =the number of observations in sample i
- $n_T = \sum_{i=1}^{k} n_i$ = the total number of observations in all samples
- R_i = the sum of the ranks for sample i

(a) Kruskal and Wallis were able to show that, under the null hypothesis assumption of identical populations, the sampling distribution of H can be approximated by a _____ distribution with _____ degrees of freedom.

(b) This approximation is generally acceptable if the _____ for each of the k populations are all _____.

(c) The null hypothesis of identical populations will be rejected if the test statistic H is large. As a result, the Kruskal-Wallis test is _____ expressed as an _____ test.

6. Example Performance Evaluation Ratings for 20 Williams Employees

(a) The value of the Kruskal-Wallis test statistic:

$$H = \frac{12}{20(21)}\left[\frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7}\right] - 3(20+1) = 8.92$$

(b) We find _____ has an area of 0.025 in the upper tail of the chi-square distribution and _____ has an area of 0.01 in the upper tail of the chi-square distribution.

(c) With $H = 8.92$ between 7.378 and 9.21, we can conclude that the p-value is between 0.025 and 0.01. Because p-value $\leq \alpha = 0.05$, we reject H_0 and conclude that the three populations are not all the same. The three populations of performance ratings are not identical and differ significantly depending upon the college.

(d) Because the sum of the ranks is relatively low for the sample of managers who graduated from _____, it would be reasonable for the company to either reduce its recruiting from college B, or at least evaluate the college B graduates more thoroughly before making a hiring decision.

7. In some applications of the Kruskal-Wallis test, it may be appropriate to make the assumption that the populations have _____ and if they differ, it is only by a _____ for one or more of the populations.

8. If the k populations are assumed to have the same shape, the hypothesis test can be stated in terms of the _____. In this case, the hypotheses for the Kruskal-Wallis test would be written as follows:

$$\begin{aligned} H_0 &: M_1 = M_2 = \cdots = M_k \\ H_a &: \text{Not all Medians are equal} \end{aligned}$$

9. NOTES+COMMENTS: The example in this section used quantitative data on employee performance ratings to conduct the Kruskal- Wallis test. This test could also have been used if the data were the _____ of the 20 employees in terms of performance. In this case, the test would use the ordinal data directly. The step of converting the quantitative data into rank-ordered data would not be necessary.

😍 **EXERCISES 18.4**: 26, 29

18.5 Rank Correlation

1. (Recall Chapter 3) The Pearson product moment correlation coefficient is a measure of the _____ between two variables using quantitative data.

2. The Spearman rank-correlation coefficient has been developed for a correlation measure of association between two variables when _____ are available:

$$\underline{\qquad\qquad\qquad\qquad} \tag{18.8}$$

where

n =the number of observations in the sample

x_i = the rank of observation i with respect to the first variable

y_i = the rank of observation i with respect to the second variable

_____

3. The Spearman rank-correlation coefficient ranges from _____ and its interpretation is similar to the Pearson product moment correlation coefficient for quantitative data.

4. A rank-correlation coefficient near _____ indicates a strong _____ _____ association between the ranks for the two variables.

5. Example Sales Potential and Actual Two-Year Sales Data

   (a) A company wants to determine whether individuals who had a greater potential at the time of employment turn out to have higher sales records. To investigate, the personnel director reviewed the original job interview reports, academic records, and letters of recommendation for 10 current members of the sales force.

   (b) After the review, the director ranked the 10 individuals in terms of their potential for success at the time of employment and assigned the individual who had the most potential the rank of 1.

   (c) (Table 18.16) Data were then collected on the actual sales for each individual during their first two years of employment. On the basis of the actual sales

records, a second ranking of the 10 individuals based on sales performance was obtained.

TABLE 18.16 Sales Potential and Actual Two-Year Sales Data

Salesperson	Ranking of Potential	Two-Year Sales (units)	Ranking According to Two-Year Sales
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

(d) (Table 18.17) Computation of the Spearman Rank-Correlation Coefficient for Sales Potential and Sales Performance

TABLE 18.17 Computation of the Spearman Rank-Correlation Coefficient for Sales Potential and Sales Performance

Salesperson	x_i = Ranking of Potential	y_i = Ranking of Sales Performance	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	−5	25
E	6	7	−1	1
F	3	4	−1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	−1	1
J	5	2	3	9
				$\Sigma d_i^2 = 44$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2+1)} = 1 - \frac{6(44)}{10(100-1)} = .733$$

(e) $r_s = 0.733$ indicates a _____ between the ranks based on potential and the ranks based on sales performance. Individuals who ranked higher in potential at the time of employment tended to rank higher in two-year sales performance.

6. Use the sample rank correlation r_s to make an inference about the population rank

correlation coefficient ρ_s:

$$H_0 : \underline{\qquad\qquad} \qquad H_a : \underline{\qquad\qquad}$$

7. (**S ampling distribution of** r_s) Under the assumption that the null hypothesis is true and the population rank-correlation coefficient is 0, the following sampling distribution of r_s can be used to conduct the test.

$$\text{Mean} : \underline{\qquad\qquad} \tag{18.9}$$

$$\text{Standard deviation} : \underline{\qquad\qquad\qquad\qquad} \tag{18.10}$$

Distribution form: Approximately normal provided _____

8. Example Sales Potential and Actual Two-Year Sales Data

   (a) The sample rank-correlation coefficient for sales potential and sales performance is _____. Using equation (18.9), we have _____, and using equation (18.10), we have _____.

   (b) With the sampling distribution of r_s approximated by a normal distribution, the standard normal random variable z becomes the test statistic with

   _____

   (c) Using the standard normal probability table and $z = 2.20$, we find the two-tailed p-value _____. With a 0.05 level of significance, p-value $\leq \alpha$. Thus, we _____ the null hypothesis that the population rank-correlation coefficient is zero.

   (d) The test result shows that there is a _____ rank correlation between potential at the time of employment and actual sales performance.

😍 **EXERCISES 18.5**: 32, 35

😍 **SUPPLEMENTARY EXERCISES**: 39, 41, 45