Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد

الدكتور فراس شاكر محمود

The Wilcoxon Tests

As mentioned earlier in the text, at times it is clear that the normality assumptions are not met and that other procedures, sometimes referred to as **nonparametric** or **distribution-free** methods, should be considered.

Example: Suppose some hypothesis, say, $H_0: m = m_0$, against $H_1: m \neq m_0$, is made about the unknown median, m, of a continuous-type distribution.

From the data, we could construct a $100(1-\alpha)\%$ confidence interval for m, and if m_0 is not in that interval, we would reject H_0 at the α significance level.

Now let X be a continuous-type random variable and let m denote the median of X. To test the hypothesis $H_0: m = m_0$ against an appropriate alternative hypothesis, we could also use a **sign test**. That is, if $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from this distribution, and if we let Y equal the number of negative differences among $X_1 - m_0, X_2 - m_0, \ldots, X_n - m_0$, then Y has the binomial distribution $b(n, 1/2)$ under H_0 and is the test statistic for the sign test. If Y is too large or too small, we reject $H_0: m = m_0$.

Example: Let X denote the length of time in seconds between two calls entering a call center.

Let m be the unique median of this continuous-type distribution. We test the null hypothesis $H_0: m = 6.2$ against the alternative hypothesis $H_1: m < 6.2$. If Y is the number of lengths of time between calls in a random sample of size 20 that are less than 6.2, then the critical region $C = \{y : y \geq 14\}$ has a significance level of $\alpha = 0.0577$.

A random sample of size 20 yielded the following data:

6.8	5.7	6.9	5.3	4.1	9.8	1.7	7.0
2.1	19.0	18.9	16.9	10.4	44.1	2.9	2.4
4.8	18.9	4.8	7.9				

Since $y = 9$, the null hypothesis is not rejected .

The sign test can also be used to test the hypothesis that two possibly dependent continuous-type random variables X and Y are such that $p = P(X > Y) = 1/2$.

To test the hypothesis $H_0 : p = 1/2$ against an appropriate alternative hypothesis, consider the independent pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

Let W denote the number of pairs for which $X_k - Y_k > 0$.When H_0 is true, W is $b(n, 1/2)$, and the test can be based upon the statistic W.

For example, say X is the length of the right foot of a person and Y the length of the corresponding left foot. Thus, there is a natural pairing, and here $H_0 : p = P(X > Y) = 1/2$ suggests that either foot of a particular individual is equally likely to be longer.

One major objection to the sign test is that it does not take into account the magnitude of the differences $X_1 - m_0, \ldots, X_n - m_0$.

We now discuss a **test of Wilcoxon** that does take into account the magnitude of the differences $|X_k - m_0|, k = 1, 2, \ldots, n$. However, in addition to assuming that the random variable X is of the continuous type, we must also assume that the pdf of X is symmetric about the median in order to find the distribution of this new statistic.

Because of the continuity assumption, we assume, in the discussion which follows, that no two observations are equal and that no observation is equal to the median.

We are interested in testing the hypothesis $H_0: m = m_0$, where $m0$ is some given constant. With our random sample $X_1, X_2, \ldots, X_n$, we rank the absolute values $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$ in ascending order according to magnitude.That is, for $k = 1, 2, \ldots, n$, we let R_k denote the rank of

$|X_k - m_0|$ among $|X_1 - m_0|, |X_2 - m_0|, \ldots, |X_n - m_0|$.

Note that $R_1, R_2, \ldots, R_n$ is a permutation of the first n positive integers, $1, 2, \ldots, n$. Now, with each R_k, we associate the sign of the difference $X_k - m_0$; that is, if $X_k - m_0 > 0$, we use R_k, but if $X_k - m_0 < 0$, we use $-R_k$. The Wilcoxon statistic W is the sum of these n signed ranks, and therefore is often called the **Wilcoxon signed rank statistic**.

<u>Example</u>: Suppose the lengths of n = 10 sunfish are

$$xi : \; 5.0 \;\; 3.9 \;\; 5.2 \;\; 5.5 \;\; 2.8 \;\; 6.1 \;\; 6.4 \;\; 2.6 \;\; 1.7 \;\; 4.3$$

We shall test $H_0: m = 3.7$ against the alternative hypothesis $H_1: m > 3.7$. Thus, we have

$x_k - m_0$:	1.3,	0.2,	1.5,	1.8,	−0.9,	2.4,	2.7,	−1.1,	−2.0,	0.6
$\lvert xk - m0 \rvert$:	1.3,	0.2,	1.5,	1.8,	0.9,	2.4,	2.7,	1.1,	2.0,	0.6
$Ranks$:	5,	1,	6,	7,	3,	9,	10,	4,	8,	2
$Signed\ Ranks$:	5,	1,	6,	7,	− 3,	9,	10,	− 4,	− 8,	2

Therefore, the Wilcoxon statistic is equal to

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25\,.$$

Incidentally, the positive answer seems reasonable because the number of the 10 lengths that are less than 3.7 is 3, which is the statistic used in the sign test .

If the hypothesis $H_0: m = m_0$ is true, about one half of the differences would be negative and thus about one half of the signs would be negative.

Hence, it seems that the hypothesis $H_0: m = m_0$ is supported if the observed value of W is close to zero. If the alternative hypothesis is $H_1: m > m_0$, we would reject H_0 if the observed $W = w$ is too large, since, in this case, the larger deviations $|X_k - m_0|$ would usually be associated with observations for which $x_k - m_0 > 0$.

That is, the critical region would be of the form $\{w: w \geq c_1\}$.

If the alternative hypothesis is $H_1: m < m_0$, the critical region would be of the form $\{w: w \leq c_2\}$. Of course, the critical region would be of the form $\{w: w \leq c_3 \ \ or\, w \geq c_4\}$ for a two-sided alternative hypothesis $H_1: m \neq m_0$.

In order to find the values of c_1, c_2, c_3, and c_4 that yield desired significance levels, it is necessary to determine the distribution of W under H_0.

Accordingly, we consider certain characteristics of this distribution.

When $H_0: m = m_0$ is true,

$$P(X_k < m_0) = P(X_k > m_0) = \frac{1}{2}\,, \qquad k = 1,2,\ldots,n.$$

Hence, the probability is 1/2 that a negative sign is associated with the rank R_k of $|X_k - m_0|$.

Moreover, the assignments of these n signs are independent because $X_1, X_2, \ldots, X_n$ are mutually independent. In addition, W is a sum that contains the integers $1, 2, \ldots, n$, each with a positive or negative sign. Since the underlying distribution is symmetric, it seems intuitively obvious that W has the same distribution as the random variable

$$V = \sum_{k=1}^{n} V_k,$$

where $V_1, V_2, \ldots, V_n$ are independent and

$$P(V_k = k) = P(V_k = -k) = \frac{1}{2}, \quad k = 1,2, \ldots, n\,.$$

That is, V is a sum that contains the integers 1, 2, . . . , n, and these integers receive their algebraic signs by independent assignments. Since W and V have the same distribution, their means and variances are equal, and we can easily find those of V.

Now, the mean of V_k is

$$E(V_k) = -k\left(\frac{1}{2}\right) + k\left(\frac{1}{2}\right) = 0\;;$$

Thus ,

$$E(W) = E(V) = \sum_{k=1}^{n} E(V_k) = 0.$$

The variance of V_k is

$$Var(V_k) = E(V_k^2) = (-k)^2\left(\frac{1}{2}\right) + (k)^2\left(\frac{1}{2}\right) = k^2\ .$$

Hence,

$$Var(W) = Var(V) = \sum_{k=1}^{n} Var(V_k) = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}\ .$$

We shall not try to find the distribution of W in general, since that pmf does not have a convenient expression.

However, we demonstrate how we could find the distribution of W (or V) with enough patience and computer support.

Recall that the moment-generating function of V_i is

$$M_k(t) = e^{t(-k)}\left(\frac{1}{2}\right) + e^{t(+k)}\left(\frac{1}{2}\right) = \frac{e^{-kt} + e^{kt}}{2},\ \ k = 1,2,\dots,n$$

Let $n\ =\ 2$; then the moment-generating function of $V_1\ +\ V_2$ is

$$M(t) = E\left[e^{t(V_1+V_2)}\right].$$

From the independence of V_1 and V_2, we obtain

$$M(t) = E(e^{tV_1})E(e^{tV_2})$$

$$= \left(\frac{e^{-t} + e^{t}}{2}\right)\left(\frac{e^{-2t} + e^{2t}}{2}\right)$$

$$= \frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4}\ .$$

This means that each of the points $-3, -1, 1, 3$ in the support of $V_1\ +\ V_2$ has probability 1/4.

Next let $n = 3$;

then the moment-generating function of $V_1 + V_2 + V_3$ is

$$M(t) = E\left[e^{t(V_1+V_2+V_3)}\right]$$

$$= E\left[e^{t(V_1+V_2)}\right]E\left(e^{tV_3}\right)$$

$$= \left(\frac{e^{-3t} + e^{-t} + e^{t} + e^{3t}}{4}\right)\left(\frac{e^{-3t} + e^{3t}}{2}\right)$$

$$= \frac{e^{-6t} + e^{-4t} + e^{-2t} + 2e^{0} + e^{2t} + e^{4t} + e^{6t}}{8}.$$

Thus, the points $-6, -4, -2, 0, 2, 4,$ and 6 in the support of $V_1 + V_2 + V_3$ have the respective probabilities $1/8, 1/8, 1/8, 2/8, 1/8, 1/8,$ and $1/8$.

Obviously, this procedure can be continued for $n = 4, 5, 6, \ldots$, but it is rather tedious.

Fortunately, however, even though $V_1, V_2, \ldots, V_n$ are not identically distributed random variables, the sum V of them still has an approximate normal distribution for large samples.

To obtain this normal approximation for V (or W), a more general form of the central limit theorem, due to Liapounov, can be used which allows us to say that the standardized random variable

$$z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}}$$

is approximately $N(0, 1)$ when H_0 is true.

We accept this theorem without proof, so that we can use this normal distribution to approximate probabilities such as

$P(W \geq c; H_0) \approx P(Z \geq z_\alpha; H_0)$ when the sample size n is sufficiently large.

<u>Example</u>: The moment-generating function of W or of V is given by

$$M(t) = \prod_{i=1}^{n} \frac{e^{-kt} + e^{kt}}{2} .$$

Using a computer algebra system such as *Maple*, we can expand $M(t)$ and find the coefficients of e_{kt}, which is equal to $P(W = k)$.

In Figure, we have drawn a probability histogram for the distribution of W along with the approximating $N[0, n(n + 1)(2n + 1)/6]\ pdf$ for $n = 4$ (a poor approximation) and for $n = 10$. It is important to note that the widths of the rectangles in the probability histogram are equal to 2, so the "half-unit correction for continuity" mentioned in Section 5.7 now is equal to 1.

<u>Example</u>: Let m be the median of a symmetric distribution of the continuous type.

To test the hypothesis $H_0: m = 160$ against the alternative hypothesis $H_1: m > 160$, we take a random sample of size $n = 16$. For an approximate significance level of $\alpha = 0.05, H_0$ is rejected if the computed $W = w$ is such that

$$z = \frac{w}{\sqrt{\frac{16(17)(33)}{6}}} \geq 1.645 ,$$

or

$$w \geq 1.645\sqrt{\frac{16(17)(33)}{6}} = 63.626\,.$$

Say the observed values of a random sample are 176.9, 158.3, 152.1, 158.8, 172.4, 169.8, 159.7, 162.7, 156.6, 174.5, 184.4, 165.2, 147.8, 177.8, 160.1, and 160.5.

In Table 1, the magnitudes of the differences $|x_k - 160|$ have been ordered and ranked.Those differences $x_k - 160$ which were negative have been underlined, and the ranks are under the ordered values.

For this set of data,

$$w = 1 - 2 + 3 - 4 - 5 + 6 + \cdots + 16 = 60.$$

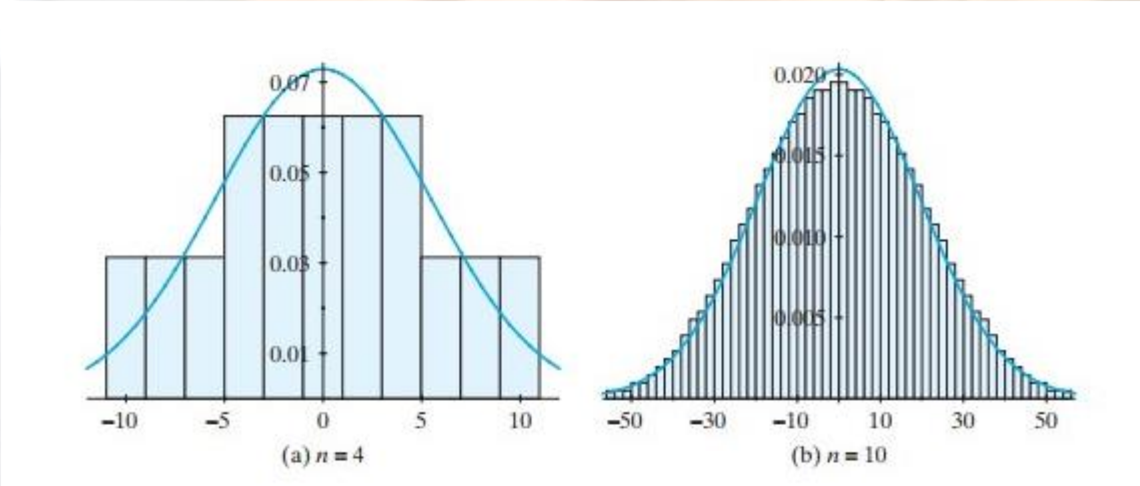

Figure 1

<u>Example</u>: The weights of the contents of $n_1 = 8$ and $n_2 = 8$ tins of cinnamon packaged by companies A and B, respectively, selected at random, yielded the following observations of *X* and *Y*:

x:	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
y:	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

The critical region for testing $H_0\colon m_X = m_Y$ against $H_1\colon m_X < m_Y$ is of the form $w \geq c$.

Since $n_1 = n_2 = 8$, at an approximate $\alpha = 0.05$ significance level H_0 is rejected if

$$z = \frac{w - 8(8+8+1)/2}{\sqrt{[(8)(8)(8+8+1)]/12}} \geq 1.645\,,$$

or

$$w \geq 1.645\sqrt{\frac{(8)(8)(17)}{12}} + 4(17) = 83.66\,.$$

To calculate the value of *W*, it is sometimes helpful to construct a **back-to-back stem-and-leaf display**. In such a display, the stems are put in the center and the leaves go to the left and the right.

(See Table 1.) Reading from this two-sided stem-and-leaf display, we show the combined sample in Table 2, with the Company B (*y*) weights underlined. The ranks are given beneath the values. From Table 2, the computed W is

$$w = 3 + 8 + 9 + 11 + 12 + 13 + 15 + 16 = 87 > 83.66.$$

Table 1 : Back-to-back stem-and-leaf diagram of weights of cinnamon

x Leaves		Stems	y Leaves	
	51	$11f$		
74	71	$11s$	72	
	95	$11\bullet$		
19	13	$12*$		
		$12t$	21	35
	45	$12f$	53	56
	78	$12s$	65	79
		$12\bullet$	98	

Table 2: Combined ordered samples

115.1	117.1	$\underline{117.2}$	117.4	119.5	121.3	121.9	$\underline{122.1}$
1	2	3	4	5	6	7	8
$\underline{123.5}$	124.5	$\underline{125.3}$	$\underline{125.6}$	$\underline{126.5}$	127.8	$\underline{127.9}$	$\underline{129.8}$
9	10	11	12	13	14	15	16

Thus, H_0 is rejected.

Finally, making a half-unit correction for continuity, we see that the p-value of this test is

$$p - value \ = \ P(W \ \geq \ 87)$$
$$= P\left(\frac{w-68}{\sqrt{90.0667}} \geq \frac{86.5-68}{\sqrt{90.667}}\right)$$
$$\approx \ P(Z \ \geq \ 1.943) \ = \ 0.0260.$$