chapter 7 - non parametric

EXERCISES

6.1. Consider the data of Exercise 1.9.

a) Use the sign test to decide if the median amount of dust in the air is more than 1200 $\mu g/m^3$. Use $\alpha = 0.05$.

b) Assuming a symmetric population, use the signed rank test to decide if the mean amount of dust in the air is more than 1200. Use $\alpha = 0.05$.

c) Do the two tests give the same conclusion?

6.2. Consider the data of Exercise 1.11. Assuming cooling times are not symmetric, use the appropriate test to decide if the median cooling time is different from 50 minutes ($\alpha = 0.10$).

6.3. Consider the data of Exercise 1.13 and let $\alpha = 0.05$.

a) Use the sign test to decide if the median fluoride content of Saudi drinking water is less than 0.7 mg/L.

b) Assuming flouride contents are symmetric, use the signed rank test to decide if the mean fluoride content of such water is less than 0.7 mg/L.

6.4. Consider the data of Exercise 2.1 and let $\alpha = 0.01$.

a) Assuming that Na cation percents are not normally distributed but do have a symmetric distribution, decide if the mean Na cation percent of Qatif well water is more than 45 using both the sign and signed rank tests.

b) Compare the results in a) with the result obtained if normality is assumed.

6.5. Consider the data of Exercise 2.4 and let $\alpha = 0.05$.

a) Assuming sulfur contents are not normally distributed but do have a symmetric distribution, decide if the mean sulfur content of Middle Eastern oil is less than 2.6 using both the sign and signed rank tests.

b) Compare the results in a) with the result obtained if normality is assumed.

6.3. Two Sample Tests for Location

In Chapter 2 when we tested for two population means ass normality or having large samples, there were two broad cas one for independent samples and one for dependent sam Similarly even when we do not assume normality, these two must be handled differently. We consider independent sample Section 6.3.1 and dependent samples in Section 6.3.2 for vari for which the assumption of normality is not appropriate and large samples are not taken.

6.3.1. Independent Samples and the Rank Sum Test

We assume independent samples (sizes n_1 and n_2) from populations with continuous distributions for a variable mea on at least an ordinal scale and that these distributions d only in their location (if at all). Therefore, we assume variances. An appropriate test statistic is based on a qua known as the **rank sum**. To find the rank sum, *we combine the from both samples and rank them from 1, 2, ... , n_1 +* *increasing order* assigning the mean rank to any tied values. the rank sum is given by

W_1 = the sum of the ranks of the sample from population 1

Note that population 1 may be arbitrarily chosen to be population from which the smallest sample was taken.

To test $H_0: \eta_1 = \eta_2$ versus some alternative, an appro test statistic is the **rank sum statistic** (also called the Whitney **rank sum** statistic) which is defined as

$$W_s = W_1 - \frac{1}{2} n_1(n_1 + 1) .$$

Critical values for various cumulative probability valu (denoted as w.) for the distribution of the test stat

EXERCISES

1.9. Suppose we measure the amount of suspended dust in the air (in $\mu g/m^3$) in a sample of residential areas of Riyadh:

1100 1200 1300 1230 1130 1310 1250 1128 1190 1260

a) Using hand calculations, find the sample mean, variance and standard deviation of the amount of suspended dust in the air. Give the units for each measure.

b) Using MINITAB, find the mean and standard deviation for the data.

1.10. Suppose we measure the number of seeds per grape berry for a sample of a particular variety of grapes which have seeds:

3	2	2	3	3	2	4	2	2	3	2	4	4
3	2	2	3	4	2	2	3	3	4	2	3	4
3	4	2	4	3	2	3	4	3	3	4	2	3

a) Using hand calculations, find the sample mean, variance and standard deviation for the number of seeds. You may wish to make a frequency table first.

b) Using MINITAB, find the mean and standard deviation for the number of seeds.

1.11. Suppose we measure the time needed to cool victims of heat stroke during hajj to Makkah (in minutes) for a sample of heat stroke victims [Based on Al-Aska et al. (1987)]:

45 20 15 29 67 75 35 110 27 40 52 33 18 21

a) Using hand calculations, find the sample mean, variance and standard deviation for the time needed to cool such victims. Give the appropriate units for each measure.

b) Using MINITAB, find the mean and standard deviation for the data.

1.12. In a study on soils in Saudi Arabia [Al-Mustafa and Ayed (1989)], 22 soil samples from agricultural areas in the central region were taken. The percentage of clay in the soil was measured:

19.2	1.3	16.0	9.8	11.0	11.0	9.8	26.0
23.0	21.8	8.6	44.0	46.0	24.0	12.6	23.8
21.4	24.5	16.0	10.0	9.6	12.8		

a) Using hand calculations, find the sample mean, variance and standard deviation for the percentage of clay in the soil.

1.13. Suppose we measure the fluoride in drinking water (in mg/l) for a sample of 15 drinking water samples in Saudi Arabia

0.65	0.85	0.50	0.71	0.45
0.32	0.91	1.02	0.67	0.51
0.78	0.25	0.60	0.79	0.63

a) Using hand calculations, find the sample mean, variance, and standard deviation for the fluoride in drinking water in Saudi Arabia. Give the appropriate units for each measure.

b) Using MINITAB, find the mean and standard deviation for the data.

1.2.2 Measures for Qualitative Variables

When we have qualitative variables, the major summary statistic is called a **proportion**. A proportion is the fraction of a population or sample which have a certain characteristic. A percentage may be obtained by multiplying the proportion by 100. There are both population and sample measures:

$$\text{population proportion } \pi = \frac{\text{number in population with characteristic}}{N}$$

$$\text{sample proportion} \quad p = \frac{\text{number in sample with characteristic}}{n}$$

Note that proportions must be numbers between 0 and 1. In words, π is "the proportion in the population with the characteristic."

Example 1.9 Suppose we have a population of 20 students in a particular statistics course in a certain semester. At the end of the term, we record the final grade of each student

B+ B A B+ C D F A C D
C B+ A C C+ D+ D F F A

The proportion of students who received a grade of A is

$$\pi = \frac{4}{20} = 0.2 \quad (\text{or } 20\%)$$

and the proportion of students who failed (had a grade of F) is

6- Reject H_o if $R < R(\alpha/2,n) = R(0.025,6) \approx R(0.024,6) = 0$.

7- $R = 1$.

8- Fail to reject H_o at $\alpha = 0.05$.

9- We can not conclude that there is a difference in the mean number of adult pests trapped at the two times.

Note that for this example, both the sign and signed rank tests give the same conclusion although this need not be true in general.

chapter 7 - Non parametric tests
Two-samples

EXERCISES

6.6. Two varieties of tomato were grown under plastic house conditions. The fruit weight (in g) for independent samples of fruit of the two varieties gave [Based on means from Alsadon and Khalil (1993)]:

Variety 1:	125	143	150	156	135	132	145	147
Variety 2:	142	160	138	144	154	158	157	161

If we can not assume normality, test whether there is a difference in the median fruit weights of the varieties ($\alpha = 0.05$).

6.7. Consider the data of Exercise 2.R.4, and let $\alpha = 0.10$.

a) Without assuming a normal distribution, test whether the average moisture content before freezing is more than the average moisture content after freezing using both the sign and signed rank tests.

b) Compare the results obtained in a) with the result obtained if normality is assumed.

6.8. Consider the data of Exercise 2.28 and let $\alpha = 0.01$.

a) Without assuming normality, test whether the median phosphorus content of skim milk is less than the median phosphorus content of whole milk.

b) Compare the result obtained in a) with the result obtained if normal populations with equal variances are assumed.

6.9. Consider the data of Exercise 2.32 and let $\alpha = 0.10$.

a) Without assuming normality, test whether the median body wall thickness of the high energy level group is more than the median for the medium energy level group.

b) Compare the result obtained in a) with the result obtained if normal populations with equal variances are assumed.

6.10. Consider the data of Example 2.26 and let $\alpha = 0.05$.

a) Without assuming normality, test whether there is a difference in the median fat contents of soft and frozen ice cream.

b) Compare the result obtained in a) with the result presented in Example 2.26 which assumed normal distributions and equal variances.

6.11. Consider the data of Exercise 2.33 and let $\alpha = 0.10$.

a) Without assuming normality, test whether the median morning time spent in resting for male camels is less than the median afternoon time using both the sign and signed rank tests.

b) Compare the results in part a) with the result if normality is assumed.

n_1 — 7

n_2 — 6

Testing for Equal Variances to Pick the Case to Use For Two Means

When discussing two means (from normal populations with unknown variances), we had two different cases

1- variances unknown but equal

2- variances unknown and unequal.

These cases require different forms for test statistics and confidence intervals and have different distributions. If we are not told to assume equal or unequal variances, one procedure is to test for this first and based on the result of the variance test, choose the "correct" procedure for the means. That is, suppose we want to test

$$H_o: \mu_1 = \mu_2 \text{ versus some alternative}$$

but we do not know whether we have equal variances $(\sigma_1^2 = \sigma_2^2)$ or unequal variances $(\sigma_1^2 \neq \sigma_2^2)$. Then, we first test

$$H_o: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2 .$$

If we reject H_o, this means we conclude that the variances are unequal. Concerning testing or estimation for the difference in the two population means, we then choose the procedure which assumes unequal variances (and use t' values).

If we fail to reject H_o, this means that the variances were not found to be significantly different. Thus, in a test or confidence interval for means, we may assume that we have equal variances and use the procedure based on this assumption (using the pooled two-sample variance s_p^2).

*(6-10) Example 2.26 The fat content (as a %) of independent samples of soft and frozen ice cream was measured [El-Erian and Al-Shaikhli (1981)]:

Frozen:	8.9	11.5	12.4	11.5	12.2	7.2	8.6	8.2	8.3	7.6
	11.1	10.3	11.6	7.6	10.0	11.0	12.0	12.3	12.5	1.6
	12.6	14.8	7.0	5.9	6.7	12.0	7.4	9.0	9.0	
Soft:	1.8	2.1	0.0	1.0	0.0	9.2	2.0	5.9	2.1	2.0	9.8

Assuming approximate normal populations, test whether there is a difference in the average fat contents of soft and frozen ice cream. Use $\alpha = 0.05$.

Solution: Since we are not told about whether the variances are equal or unequal, we will first make a test for the equality of the variances.

1-*Data*: Variable-fat content

Populations- 1) all frozen ice cream

2) all soft ice cream

(in Riyadh in 1981)

$n_1 = 29$, $\bar{x}_1 = 9.6827586$, $s_1^2 = 7.5379064$ $\qquad \alpha = 0.05$

$n_2 = 11$, $\bar{x}_2 = 3.2636364$, $s_2^2 = 11.938545$

2-*Assumptions*: Assume normal populations.

3-*Hypotheses*: $H_o: \sigma_1^2 = \sigma_2^2$.

$H_a: \sigma_1^2 \neq \sigma_2^2$

4-*Test statistic*:

$$F = s_1^2 / s_2^2$$

5-*Distribution*:

F has a $F_{n_1-1, n_2-1} = F_{28,10}$ distribution if H_o is true.

6-*Decision rule*:

Reject H_o if $F < F_{\alpha/2,28,10} = F_{.025,28,10} = 1/F_{.975,10,28}$

$= 1/2.55 = 0.3922$

or if $F > F_{1-\alpha/2,28,10} = F_{.975,28,10} = 3.37$.

7-*Calculation*:

$$F = \frac{7.5379064}{11.938545} = 0.6314$$

8-*Decision*: Fail to reject H_o (at $\alpha = 0.05$).

9-*Conclusion*: We can not conclude that the variances of fat contents for the soft and frozen ice cream are different.

Since we did not find the variances to be different, we may treat them as equal in the desired test for the means. Thus, we now use

5-Conclusion: We can conclude that there was a difference in the average number of adult pests trapped in yellow sticky traps at the two different times.

For the confidence interval with $1-\alpha = 0.95$, the appropriate formula (based on the data and assumptions above) is

$$\bar{d} \pm t_{1-\alpha/2,n-1}\, s_d/\sqrt{n}$$

where $t_{1-\alpha/2,n-1} = t_{0.975,5} = 2.5706$. This gives

$$-31.83333 \pm 2.5706\ (29.491807/\sqrt{6})$$

$$-31.83333 \pm 30.949972$$

$$(-62.783305\ ,\ -0.8833615)$$

Interpretation: We are 95% confident that the average number caught at time 1 is from 0.9 to 62.8 less than the average number caught at time 2.

Note that we could have done the confidence interval first. Since the test has a different than alternative and the level α is the same, we could then make the test simply by checking whether 0 is in the confidence interval. Since it is not, we would reject the null hypothesis and conclude that μ_d is different from zero.

EXERCISES

2.28

2.28. The phosphorus content was measured for independent samples of skim and whole milk:

Whole:	94.95	95.15	94.85	94.55	94.55	93.40	95.05
	94.35	94.70	94.90				
Skim:	91.25	91.80	91.50	91.65	91.15	90.25	91.90
	91.25	91.65	91.00				

Assuming normal populations with equal variances,

a) Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk. Use $\alpha = 0.01$.

b) Find and interpret a 99% confidence interval for the difference in average phosphorus contents of whole and skim milk.

c) Could the confidence interval found in part b) be used to make the test in part a)? Why or why not?

2.29. In a study on chemical weed control for ... Arabia [Based on Tamim and Kadous (1984)], a ... (dimethyl tetrachloro terephthalate) was appl... rates - 6.70 and 8.97 kg/ha. The potato girth size was measured obtaining:

		Sample size	Mean	Standard devia...
Rate	6.70	10	37.82	4.89
	8.97	10	36.39	1.60

Assuming normal populations with unequal variances, ... whether there is a difference in the average potato size in the two rate groups. Use $\alpha = 0.10$.

2.30. Obesity, the condition of being very overweight, increa... person's risk for various health problems. One su... procedure used to deal with obesity is called bari... surgery and attempts to decrease the amount of food th... person can eat. In a study of obese Saudis operated o... bariatric surgery [Mofti and Al-Saleh (1992)], the we... of 31 obese Saudis were measured before and two years ... surgery:

Before	After	Before	After	Before	After
148	78	154	133	110	70
145	78	114	60	107	80
123	80	129	70	143	72
140	81	148	70	134	71
129	87	113	60	151	76
119	70	117	120	129	61
151	94	122	81	131	89
122	79	149	95	129	60
120	75	109	67	108	71
150	89	137	63		
102	70	154	83		

Find and interpret a 95% confidence interval for ... difference in the average weight of obese Saudis before ... two years after receiving bariatric surgery.

2.31. Bacteria can, under certain conditions, penetrate the sh... pores of eggs and may cause the egg not to hatch. In a st... on the bacterial contamination of hatching eggs [Based ... Barbour and Nabbut (1983)], the bacterial count was measu... for eggs from layer hens and for eggs from hens raised ... meat obtaining:

	Sample size	Mean
Layer	28	...

الفصل 7

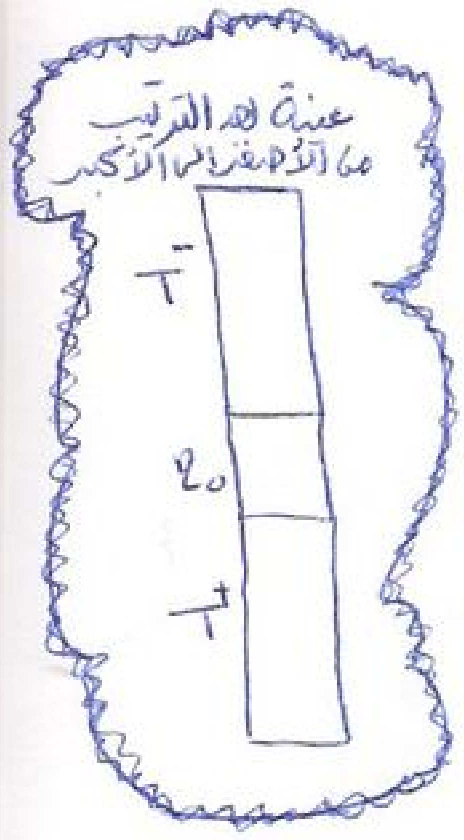
Sign test → 6.1(a)
→ 6.2, 6.3(a) (H.W)

Mann-whitney test → 6.6
→ 6.8(a) (H.W)

☺

① اختبار الإشارة sign test

sign test →
- مجتمع واحد بعينة حجمها n ،
- للبيانات الكمية والوصفية التي يمكن ترتيبها
- الاختبار حول الوسيط η



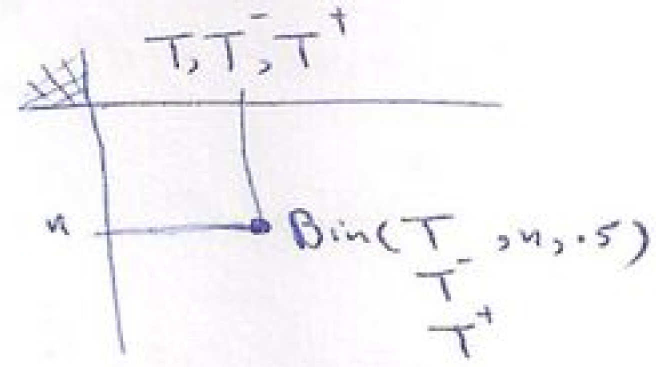
1. $H_0: \eta = \eta_0$ Vs $H_1: \eta \neq \eta_0$, $>$, $<$

2. الإحصاءة $=$ $\xrightarrow{\neq}$ $T = \min(T^-, T^+)$
   $\xrightarrow{>}$ $T^- =$ عدد القيم التي تقل عن η_0 في العينة
   $\xrightarrow{<}$ $T^+ =$ عدد القيم التي تزيد عن η_0 في العينة

3. نرفض H_0 عندما
   $\xrightarrow{\neq}$ $P\text{-value} = 2\,Bin(T, n, .5) < \alpha$
   $\xrightarrow{>}$ $P\text{-value} = Bin(T^-, n, .5) < \alpha$
   $\xrightarrow{<}$ $P\text{-value} = Bin(T^+, n, .5) < \alpha$

حيث n هي حجم العينة بعد استبعاد القيم التي تساوي η_0 و لإيجاد القيم الجدولية تكون كما يلي :

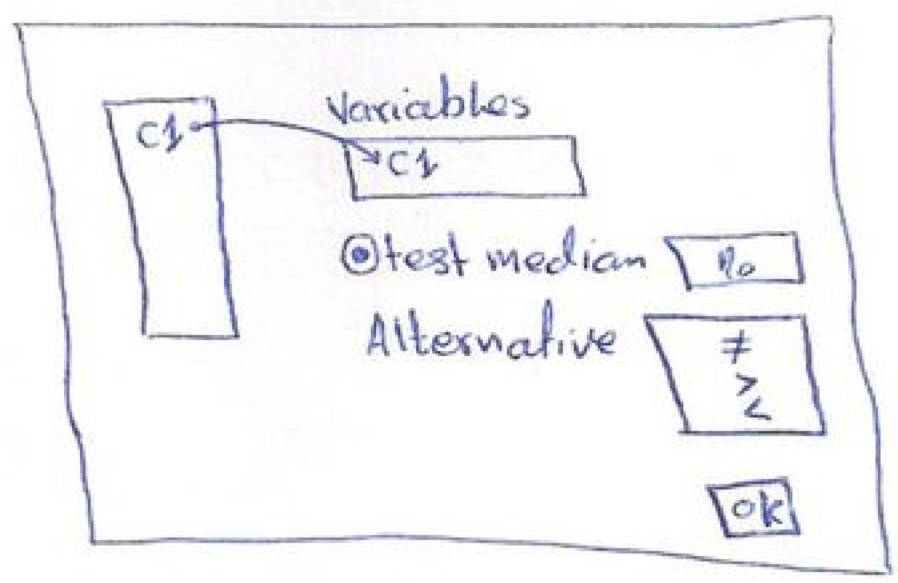
at .5



باستخدام minitab :

اضع البيانات في عمود واحد وليكن اسمه C1 ثم

stat → nonparametrics → 1-sample sign



سوف يظهر لنا التالي :

sign test For Median: C1

sign test on median $= \eta_0$ VS $\neq \eta_0$, $>$, $<$

n	T^-	$n(\eta_0)$	T^+	P-value

ملاحظة : تطبق هذه الطريقة على البيانات الكمية أما البيانات الوصفية القابلة للترتيب فلها طريقة أخرى ...

② اختبار مجموع الرتب Mann-whitney test

- مجتمعين مستقلين بعينتين حجمهما n_1 و n_2 وكلاهما أقل من 30 وليس لهما توزيع طبيعي
- للبيانات الكمية والوصفية التي يمكن ترتيبها
- اختبار حول الوسيطين η_1 و η_2

$\boxed{1}$ $H_0: \eta_1 = \eta_2$ vs $H_1: \eta_1 \neq \eta_2$ ($>$, $<$)

$\boxed{2}$ $W_S = W_1 - \frac{n_1(n_1+1)}{2}$,

W_1 : مجموع الرتب للعينة الأولى

n_1 : حجم العينة الأولى

		rank
1		
2		
⋮		
n_1+n_2		

أخذ العينتين وجعلتها كعينة واحدة و ترتيبها من الأصغر إلى الأكبر هذا إذا كانت كمية أما الوصفية فترتيبها فعلى المتعارف عليه

تمييز العناصر التي من العينة الأولى بعلامة ولتكن ✓

ملاحظة: الرقم المكرر الذي له أكثر من رتبة فإنه بغض النظر من أي عينة تكون رتبتها الموحدة = $\frac{\text{مجموع الرتب}}{\text{عددها}}$

$\boxed{3}$ نرفض H_0 عندما

$\neq$: $W_S < W_{\frac{\alpha}{2}, n_1, n_2}$ or $W_S > W_{1-\frac{\alpha}{2}, n_1, n_2} = n_1 n_2 - W_{\frac{\alpha}{2}, n_1, n_2}$

reject H_0 ⋮ accept H_0 ⋮ reject H_0

$W_{\frac{\alpha}{2}, n_1, n_2}$ $\qquad$ $W_{1-\frac{\alpha}{2}, n_1, n_2} = n_1 n_2 - W_{\frac{\alpha}{2}, n_1, n_2}$

$>$: $W_S > W_{1-\alpha, n_1, n_2} = n_1 n_2 - W_{\alpha, n_1, n_2}$

accept H_0 ⋮ reject H_0

$W_{1-\alpha, n_1, n_2} = n_1 n_2 - W_{\alpha, n_1, n_2}$

$<$: $W_S < W_{\alpha, n_1, n_2}$

reject H_0 ⋮ accept H_0

W_{α, n_1, n_2}

p-value $< \alpha$

لإيجاد القيمة الجدولية تكون كما يلي:

$W_{\alpha, n, m}$ ↕ $\frac{\alpha}{2}$

n or m	p	m or n

باستخدام minitab : نضع العينة الأولى في عمود C1 والعينة الثانية في عمود C2 ثم stat → nonparametrics → Mann-whitney

C1, C2	first sample [C1]
	second sample [C2]
	confidence level [$1-\alpha$ or $(1-\alpha)\%$]
	alternative [$\neq$ / $>$ / $<$]
	[ok]

سوف يظهر لنا التالي:

	N	median
C1	n_1	
C2	n_2	

$W_1 =$

test $\eta_1 = \eta_2$ vs $\eta_1 \neq \eta_2$ ($\gtrless$) is significant at p-value

ETA1 ETA2

ملاحظة: تطبق هذه الطريقة على البيانات الكمية أما البيانات الوصفية القابلة للترتيب فلها طريقة أخرى..

✓

(6.1)

ch.7

حل التمارين

1. $H_0: \eta = 1200$ Vs $H_1: \eta > 1200$, $\alpha = .05$

2.

1100, 1128, 1130, 1190	$T^- = 4$
1200	η_0
1230, 1250, 1260, 1300, 1310	$T^+ = 5$

$n_1 = 10$

$n = 9$

$T^- = 4 =$ إحصاءة الإختبار ∴

3. $Bin(T^-, n, .5) = Bin(4, 9, .5) = .5 = $ P-value

∴ as $Bin(4, 9, .5) = .5 > \alpha = .05$

so we accept H_0

باستخدام Minitab

sign test of median = 1200 Vs > 1200

	N	T^-= below	$n(\eta_0)$= equal	T^+= above	P	Median
C1	$n_1 = 10$	4	1	5	.5	1215

6.6

1. $H_0: \eta_1 = \eta_2$ vs $H_1: \eta_1 \neq \eta_2$, $\alpha = .05$

2. $W_S = W_1 - \frac{n_1(n_1+1)}{2}$

$= 51 - 36 = 15$

			rank
1	125	✓	1
2	132	✓	2
3	135	✓	3
4	138		4
5	142		5
6	143	✓	6
7	144		7
8	145	✓	8
9	147	✓	9
10	150	✓	10
11	154		11
12	156	✓	12
13	157		13
14	158		14
15	160		15
16	161		16

$n_1 = 8$

$W_1 = 1+2+3+6+8+9+10+12 = 51$

3. $W_{\frac{\alpha}{2}, n_1, n_2} = W_{.025, 8, 8} = 14$

$W_{1-\frac{\alpha}{2}, n_1, n_2} = n_1 n_2 - W_{\frac{\alpha}{2}, n_1, n_2} = 8(8) - 14 = 64 - 14 = 50$

$\therefore$ as $14 < W_S < 50$

$\therefore$ accept H_0

باستخدام Minitab

	N	median
C1	$n_1 = 8$	144
C2	$n_2 = 8$	155.5

$W = 51$

Test of ETA1=ETA2 vs ETA1 not= ETA2 is significant at .0831

$\therefore$ as we see that P-value $= .0831 > \alpha = .05$

so, we accept H_0