Assignment 3

Due Date: **23 January 2017**
Weighting: **25%**

- Answering the questions in this assignment should not be your first attempt at these types of questions. It is essential that you work through practice exercises from the tutorial sheets and/or text book first.
- This assignment is important in providing feedback and helping to establish competency in essential skills.
- Answer all the questions. The questions are not of equal weight; some questions are worth much more than others.
- The questions relate to material up to and including Module 10.
- Before starting this assignment read *Notes Concerning Assignments* under the *Introductory Material* link on the StudyDesk.
- When you are asked to comment on a finding, usually a short paragraph is required.
- In many cases the SPSS output contains much more information than is required for a correct and complete answer. In those cases just reproducing the output may not attract any marks. **Make sure you report only the information from the SPSS output relevant to your answer.**
- Unless instructed otherwise, show all working and formulae used in calculating confidence intervals and performing hypothesis tests. (Answers may of course be checked where possible using SPSS).
- In order to obtain full marks for any question you must **show all working**.
- Submission is via the link on the StudyDesk.
- This assessment item consists of 4 questions.

Question 1 (22 marks)

A low level of HDL (high density lipoprotein) is associated with an increased risk of heart disease. It is known that 14% of women in the general population have a low HDL level. There is concern that the proportion of African American *women* (living in Virginia) with a low HDL level is higher than this. Use the information in the dataset *diabetes16.sav* to answer the following questions:

(a) (2 mark) **Using SPSS**, calculate the proportion of African American *women* in the study who have a low HDL level (i.e. a high risk of heart disease).

(b) (20 marks) **Without using SPSS**, determine whether there is evidence to support the theory that the proportion of African American *women* (living in Virginia) with a low HDL level is higher than 14%.
Perform a hypothesis test to statistically justify your answer by completing the following:

i. State the appropriate hypotheses (define any symbols used).
ii. Check the conditions and assumptions for this test.
iii. Calculate the test statistic for this test.
iv. Calculate the *P*-value for this test.
v. Interpret the *P*-value and write a meaningful conclusion in the context of this situation.

Question 2 (22 marks)

Use the information in the dataset *diabetes16.sav* to answer the following questions. You should use SPSS to calculate the sample statistics you will need to do this question, but for parts (b) and (c) you are required to do all other calculations by hand, using a calculator.

(a) (7 marks) Check the appropriate conditions and assumptions needed to calculate either a confidence interval or hypothesis test in relation to the population mean weight of African American *males* in Virginia. **Include an appropriate graph to support your answer**.

(b) (6 marks) Estimate the population mean weight of African American *males* in Virginia, using a 95% confidence interval (**show all working**).

(c) (9 marks) From historical data, it is known that the mean weight of African American *males* is 90 kg. Perform a hypothesis test to see if there is evidence to support a suspicion that the mean weight of African American *males* in Virginia is different to this.

In performing this test:

i. State appropriate hypotheses (define any symbols used).
ii. Calculate the value of a suitable test statistic for this test.
iii. Calculate the *P*-value of this test.
iv. Write a meaningful conclusion at the 5% level of significance.

Question 3 (24 marks)

Use the information in the dataset *diabetes16.*sav to answer the following questions. You should use SPSS to calculate any sample statistics you will need to do this question, but for parts b(ii) and (c) you are required to do all other calculations by hand, using a calculator.

Data was collected at two different locations in Virginia – Birmingham and Louisa. As a researcher you are interested to see whether there is a difference in mean glycosylated haemoglobin percentage between residents at the two locations.

(a) (4 marks) Check the appropriate conditions and assumptions needed to perform a hypothesis test comparing the population mean glycosylated haemoglobin percentage for residents of the two locations. **Include an appropriate graph to support your answer**.

(b) (14 marks) Using an appropriate statistical test, determine if, on average, there is a difference in glycosylated haemoglobin percentage between people living in Birmingham and those living in Louisa. In performing the test, include:

i. State appropriate hypotheses, clearly defining all symbols.
ii. Calculate a suitable test statistic (you can use the results from part (a) in this calculation.
iii. Find the *P*-value of the test (and include the degrees of freedom).
iv. Interpret the *P*-value and write a meaningful conclusion in the context of the question.
v. **Now use SPSS to check** your results for this hypothesis test. Attach or copy and paste the relevant output from SPSS for this test to your assignment.
vi. Briefly comment on how the test statistic and *P*-value from SPSS output are similar to or differ from your hand calculations.

(c) (6 marks) Estimate, with 90% confidence, the population mean difference in glycosylated haemoglobin percentage between people living in Birmingham and those living in Louisa. Ensure you explain the confidence interval in the context of the question.

Question 4 (32 marks)

Use the information in the dataset *diabetes16.*sav to answer the following questions.

Diastolic blood pressure was measured at the beginning of the study (dia_bp_1) and then again four weeks later (dia_bp_2). As researcher you want to know, if, on average, the diastolic blood pressure after four weeks is less than the diastolic blood pressure at the time of entry into the study.

(a) (16 marks) Use a **parametric** test to answer this question by completing the following (parts i. to v. are to be completed **without the aid of SPSS**, although summary statistics, i.e. mean and standard deviation, required for the test can be found using SPSS):

i. State appropriate hypotheses (define any symbols used).
ii. State (but do not check) the assumptions for carrying out this test. Describe the assumptions in the context of this question.
iii. Calculate the value of a suitable test statistic for this test.
iv. Calculate the *P*-value of this test.
v. Interpret the *P*-value and describe the outcome of the test in the context of this question.
vi. **Now use SPSS** to carry out the analysis. Copy and paste the relevant SPSS output to your assignment. Do these results agree with those found in part iv? (Hint: comment on the p-value).

(b) (16 marks) Describe an alternative statistical test that could be used to answer this question.

Include in your answer:

i. the name of the test,
ii. the conditions/assumptions required for this test (in the context of the question),
iii. a definition of the test statistic that would need to be calculated to perform this test,
iv. the relative advantages and/or disadvantages of this test compared with the test you conducted to answer part (a),
v. the circumstances under which you would use this test in preference to the one used in part (a).
vi. **Now use SPSS** to carry out the analysis. Copy and paste the relevant SPSS output into your assignment.
vii. State and interpret the *P*-value **from the SPSS output** and describe the outcome of the test in the context of this question.

Question 1 (22 marks)

(a)

The cross tabulation table showing the association between African American women and HDL rating is given below

			HDL Rating		Total
			Low HDL - High risk of heart disease	High or Average HDL - Low risk of heart disease	
Gender	Male	Count	27	34	61
		% within Gender	44.3%	55.7%	100.0%
		% within HDL Rating	64.3%	35.8%	44.5%
	Female	Count	15	61	76
		% within Gender	19.7%	80.3%	100.0%
		% within HDL Rating	35.7%	64.2%	55.5%
Total		Count	42	95	137
		% within Gender	30.7%	69.3%	100.0%
		% within HDL Rating	100.0%	100.0%	100.0%

From the above table, we see that the proportion of African American women in the study who have a low HDL level is 0.1974 or 19.74%

(b).

In order to determine whether the proportion of African American women having low LDL level is greater than 14%, we perform One proportion Z test

(i)

Null Hypothesis: H_0: $P \leq 0.14$

That is, the proportion of African American women having low LDL is not greater than 14%

Alternate Hypothesis: H_a: $P > 0.14$

That is, the proportion of African American women having low LDL is greater than 14%

Level of Significance: Let the level of significance be $\alpha = 0.05$

(ii)

The assumptions that needs to be satisfied to perform one proportion z test is given below

- The samples taken for the study should be selected at random from the population of interest
- Samples selected each other should be independent of each other
- The success samples selected should neither be too small nor be large enough
- $np \geq 10$
- $n(1 - p) \geq 10$

Here, we see that the sample represents the subset of data collected from a health study on the prevalence of obesity, diabetes, and other cardiovascular risk factors in African Americans living in Virginia and the sample size is 137. Therefore, the samples are selected at random and are independent of each other. 76 out of 137 samples represents African American Women

Here, $np = 0.197 * 76 = 15$ and $n * (1 - p) = 0.803 * 76 = 61$

(iii)

The z test statistic is calculated by using the formula given below

$$Z = \frac{p - P}{\sqrt{\frac{P * (1 - P)}{n}}} = \frac{0.1974 - 0.14}{\sqrt{\frac{0.14 * (1 - 0.14)}{76}}} = 1.4413$$

(iv)

The p – value of z test statistic is

$P(Z_{0.05} > 1.4413) = 0.0747$ (By referring normal distribution table)

(v)

Since the p – value of z test statistic is greater than 0.05, we fail to reject the null hypothesis at 5% level of significance. Therefore, is no statistical evidence to conclude that the proportion of African American women having low LDL is greater than 14%

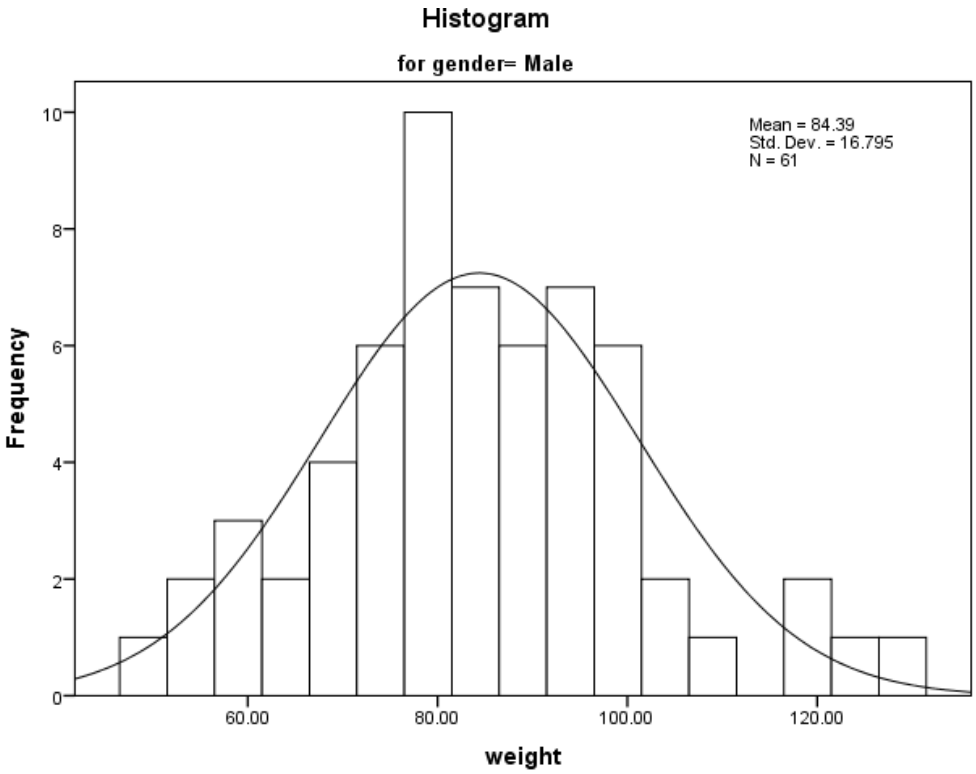
Question 2 (22 marks)

(a)

Assumptions:

1) The population variance for the sample taken into consideration should be known, or
2) The sample size should be at least 30

Here, we see that the sample of African American men is 61 and therefore, we can use z test to compute the confidence interval for the mean weight of African American Men in Virginia



Histogram for mean weight of African American male respondents was constructed to validate the assumption of normality. Going through the histogram, it is observed that the distribution of weight of African American male respondents has equal tail widths on both sides of the normal curve indicating that the distribution of weight of African American male respondents follows normal distribution

(b)

The 95 % confidence interval for the population mean is given by:

$$\left(\bar{x} - Z_{\alpha/2} * \frac{s}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} * \frac{s}{\sqrt{n}}\right)$$

The table value of z is taken from normal distribution table

$\mathbf{Z_{\alpha/2} = Z_{0.025} = 1.96}$ (by referring normal table)

$$\left(84.39 - 1.96 * \frac{16.795}{\sqrt{61}}, 84.39 + 1.96 * \frac{16.795}{\sqrt{61}}\right) = (80.092, 88.695)$$

That is, the 95 % confidence interval for mean weight of African American men is **(80.092, 88.695)**

(c)

In order to determine whether the mean weight of African American men differ significantly from 90 kg, we perform one mean z test

(i)

Null Hypothesis: H_0: $\mu = 90$

That is, the mean weight African American men do not differ significantly from 90 kg

Alternate Hypothesis: H_a: $\mu \neq 90$

That is, the mean weight African American men differ significantly from 90 kg

Level of Significance: Let the level of significance be $\alpha = 0.05$

(ii)

The assumptions that needs to be satisfied to perform one mean z test is given below

- The samples taken for the study should be selected at random from the population of interest
- Samples selected each other should be independent of each other
- The sample size taken should be sufficiently large (n >= 30)

Here, we see that the sample represents the subset of data collected from a health study on the prevalence of obesity, diabetes, and other cardiovascular risk factors in African Americans living in Virginia and the sample size is 137. Therefore, the samples are selected at random and are independent of each other. 61 out of 137 samples represents African American men

The z test statistic is calculated by using the formula given below

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{84.3934 - 90}{\frac{16.795}{\sqrt{61}}} = -2.6073$$

(iii)

The p – value of z test statistic is

P ($Z_{0.05}$ > |-2.6073|) = 0.0091 (By referring normal distribution table)

(iv)

Since the p – value of z test statistic is less than 0.05, we reject the null hypothesis at 5% level of significance. Therefore, is a statistical evidence to conclude that the mean weight of African American men differ significantly from 90 kg

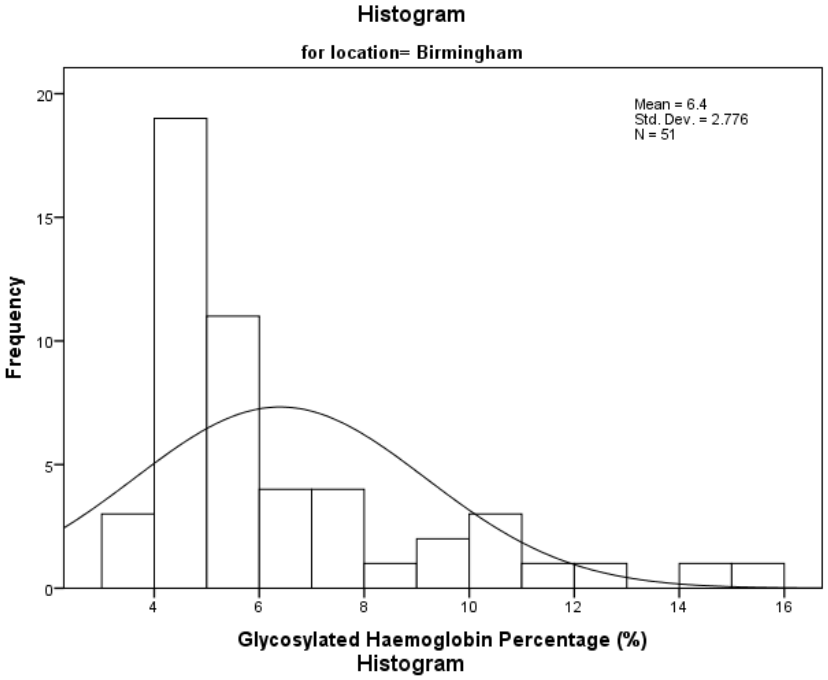
Question 3 (24 marks)

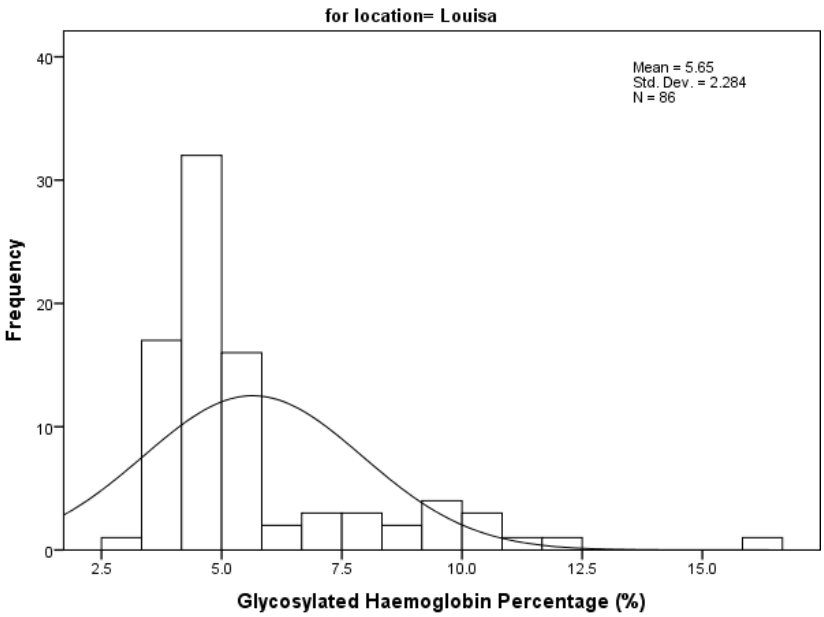
(a) The assumptions to perform independent sample t test is given below

1. The samples taken into consideration should be independent of each other

2. The distribution of two samples taken into consideration should follow normal distribution

3. Homogeneity of variances





Going through the histogram, we see that the distribution of glycosylated hemoglobin percentage for both locations seems to be extended longer towards the right side of normal curve, indicating that the distribution of glycosylated hemoglobin percentage for both locations is skewed right and hence the assumption of normality is violated. Hence, performing independent sample t test is not appropriate

(b) In order to determine whether there is a significant difference in mean glycosylated hemoglobin percentage between two locations, we perform independent sample t test

(i)

Null Hypothesis: H_0: $\mu_{Birmingham} = \mu_{Louisa}$

That is, there is no significant difference in the mean glycosylated hemoglobin percentage between two locations

Alternate Hypothesis: H_a: $\mu_{Birmingham} \neq \mu_{Louisa}$

That is, there is a significant difference in mean glycosylated hemoglobin percentage between two locations

Level of Significance: Let the level of significance be $\alpha = 0.05$

(ii)

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{6.4 - 5.65}{2.48\sqrt{\frac{1}{51} + \frac{1}{86}}} = 1.712$$

(iii)

The p – value of z test statistic is

P ($t_{0.05,135} > |-1.712|$) = 0.089 (By referring t distribution table)

(iv)

Since the p – value of z test statistic is greater than 0.05, we fail to reject the null hypothesis at 5% level of significance. Therefore, is no statistical evidence to conclude that there is a significant difference in mean glycosylated hemoglobin percentage between Birmingham and Louisa locations

(v)

The SPSS output is given below

Group Statistics

	Location	N	Mean	Std. Deviation	Std. Error Mean
Glycosylated Haemoglobin Percentage (%)	Birmingham	51	6.40	2.776	.389
	Louisa	86	5.65	2.284	.246

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	90% Confidence Interval of the Difference	
									Lower	Upper
Glycosylated Haemoglobin Percentage (%)	Equal variances assumed	2.658	.105	1.712	135	.089	.749	.438	.024	1.475
	Equal variances not assumed			1.629	89.706	.107	.749	.460	-.015	1.514

(vi)

On comparing the SPSS output with the independent sample t test statistic calculated manually, we observe that the t test value and the p – value remains the same.

(c)

The 90 % confidence interval for the population mean is given by:

$$\left(\bar{x} - t_{\alpha/2,(n1+n2-1)} * \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,(n1+n2-1)} * \frac{s}{\sqrt{n}} \right)$$

The table value of t is taken from normal distribution table

$\mathbf{t_{\alpha/2,135} = t_{0.05,135} = 1.656}$ (by referring t distribution table)

$$\left((6.4-5.65)-1.656*2.478*\sqrt{\frac{1}{51}+\frac{1}{86}},(6.4-5.65)+1.656*2.478*\sqrt{\frac{1}{51}+\frac{1}{86}}\right)=(-0.015,1.514)$$

The 90% confidence interval for the mean difference in glycosylated hemoglobin percentage between Birmingham and Louisa locations is (- 0.015, 1.514). Since the value 0 falls in the 90% confidence interval, there is no statistical evidence to conclude that there is a significant difference in mean glycosylated hemoglobin percentage between Birmingham and Louisa locations

Question 4 (32 marks)

(a)

(i)

Null Hypothesis: H_0: $\mu_d = 0$

That is, the mean diastolic blood pressure after four week is not less when compared to the mean diastolic blood pressure recorded at the time of the study

Alternate Hypothesis: H_a: $\mu_d > 0$

That is, the mean diastolic blood pressure after four week is less when compared to the mean diastolic blood pressure recorded at the time of the study

(b)

Assumptions

The dependent variable should be continuous

The distribution of the dependent variables taken into consideration should follow normal distribution table

There should not be any outliers in the dataset

(iii) The table given below shows the workings of paired t test statistic

t-Test: Paired Two Sample for Means

	dia_bp_1	*dia_bp_2*
Mean	94.14599	92.54745
Variance	129.1991	130.279
Observations	137	137
Pearson Correlation	0.786033	
Hypothesized Mean Difference	0	
df	136	
t Stat	2.511038	
P(T<=t) one-tail	0.006604	
t Critical one-tail	1.656135	
P(T<=t) two-tail	0.013208	
t Critical two-tail	1.977561	

(iv)

The value of t test statistic is 2.511 and its corresponding p – value is 0.0064

(v)

Since the p – value of t test statistic is less than 0.05, we reject the null hypothesis at 5% level of significance. Therefore, we conclude that the mean diastolic blood pressure after four week is less when compared to the mean diastolic blood pressure recorded at the time of the study

(vi)

The SPSS output is given below

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	First Diastolic Blood Pressure Reading (mmHg)	94.1460	137	11.36658	.97111
	Second Diastolic Blood Pressure Reading (mmHg)	92.5474	137	11.41398	.97516

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	First Diastolic Blood Pressure Reading (mmHg) - Second Diastolic Blood Pressure Reading (mmHg)	1.59854	7.45127	.63661	.33961	2.85747	2.511	136	.013

On comparing the SPSS output with the independent sample t test statistic calculated manually, we observe that the t test value and the p – value remains the same.

(b)

(i) An alternate test that will test whether the mean diastolic blood pressure after four week is less when compared to the mean diastolic blood pressure recorded at the time of the study is Wilcoxon signed rank test

(ii) Since this is a non-parametric test, no assumptions are required to perform this test

(iii)

The following steps need to be performed for performing Wilcoxon signed rank test

- The values with no differences should be removed
- The remaining differences should be arranged in ascending order of magnitude ignoring the signs
- Ranks should be assigned
- Average should be taken for equal ranks
- Now, calculate the number of positives (T+) and number of negatives (T-)
- If no differences found between two groups, then then T+ & T- should be same
- Take the smaller sum of T+ and T- and represent it as T
- Compare this T with the critical value obtained using Wilcoxon signed rank test table
- If the value of t test statistic is greater than the critical value of t, we reject null hypothesis, else, we fail to reject null hypothesis

(iv)

The main advantage is when the parametric assumptions are violated, then paired t test is not appropriate and in this situation, Wilcoxon signed rank test is more appropriate to compare two matched groups

The main disadvantage is that, this test is not effective for larger sample sizes

(v) When the parametric assumptions are violated, then paired t test is not appropriate and in this situation, Wilcoxon signed rank test is more appropriate to compare two matched groups

(vi)

The SPSS output is given below

Ranks

		N	Mean Rank	Sum of Ranks
Second Diastolic Blood Pressure Reading (mmHg) - First Diastolic Blood Pressure Reading (mmHg)	Negative Ranks	66[a]	58.51	3861.50
	Positive Ranks	45[b]	52.32	2354.50
	Ties	26[c]		
	Total	137		

a. Second Diastolic Blood Pressure Reading (mmHg) < First Diastolic Blood Pressure Reading (mmHg)

b. Second Diastolic Blood Pressure Reading (mmHg) > First Diastolic Blood Pressure Reading (mmHg)

c. Second Diastolic Blood Pressure Reading (mmHg) = First Diastolic Blood Pressure Reading (mmHg)

Test Statistics[a]

	Second Diastolic Blood Pressure Reading (mmHg) - First Diastolic Blood Pressure Reading (mmHg)
Z	-2.226[b]
Asymp. Sig. (2-tailed)	.026

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Since the p – value of t test statistic is less than 0.05, we reject the null hypothesis at 5% level of significance. Therefore, we conclude that the median diastolic blood pressure after four week is less when compared to the mean diastolic blood pressure recorded at the time of the study