SOLUTIONS TO ASSIGNMENT 2

Question 1

(a) The purpose of the study is to compare the survival times of patients, who have advanced cancers of the breast, bronchus, colon, ovary or stomach, and then examine whether patient survival differs with respect to the organ affected by the cancer.

The study is an example of an observational study. Therefore no causal inferences about the effect of a particular cancer type on survival time can be made from the statistics alone. Any observed differences in the survival times may be due also to some confounding variables like stage of their disease, age, gender, lifestyle, functional ability, psychological profile, etc. Given the study design (no random assignment of subjects to the five groups) it is not possible to separate the effects of the cancer type and the other variables on the survival time.

The patients have been selected randomly from the list of cancer patients in one hospital. In a strict sense, the statistical results can be generalized to the populations of the five cancer patients in this particular hospital. Any extrapolation to the whole populations of the five cancer patients in all hospitals must be based on the assumption that these patients are as representative of the whole populations. This is not a necessarily a bad assumption. However, as the cancer treatments may vary from hospital to hospital, formally it is not possible to extend the inferences to the populations of cancer patients in other hospitals.

(b) There are many other factors that might have affected the survival times of the cancer patients, such as age, lifestyle (smoking habit, eating habit, etc), province of residence, socio-economic status and so on.

In order to make a meaningful comparison, it is important to assure that patients in the five groups have similar age, gender, and socio-economic status distributions. Moreover, the patients in the five cancer groups should be matched according to similar stage of the disease determined by some very rigid criteria.

(c) Developing consistent criteria of untreatability across the five cancer types may be the most challenging part of the study. Indeed, some cancer types (i.e. prostate cancer) exhibit very unique characteristics that are not easily comparable to other cancer types. However, if criteria for some cancers are more rigid than for the other cancer types, the outcome of the study may be easily affected. For example, colon cancer with its high mortality rate may be easily categorized as untreatable though breast cancer with high survival rate may be subjected to less rigid criteria.

Question 2

(a) The descriptive statistics for the five types of cancers are displayed below:

	ORGAN			Statistic	Std.Error
SURVIVAL	**Breast**	Mean		1250.4545	272.50966
		95% Confidence Interval for Mean	Lower Bound	643.2652	
			Upper Bound	1857.6439	
		5% Trimmed Mean		1204.2828	
		Median		1035.0000	
		Variance		816876.673	
		Std. Deviation		903.81230	
		Minimum		224.00	
		Maximum		3108.00	
		Range		2884.00	
		Interquartile Range		1085.00	
		Skewness		.968	.661
		Kurtosis		.366	1.279
	Bronchus	Mean		305.7059	62.37280
		95% Confidence Interval for Mean	Lower Bound	173.4815	
			Upper Bound	437.9303	
		5% Trimmed Mean		280.9510	
		Median		223.0000	
		Variance		66136.221	
		Std. Deviation		257.16963	
		Minimum		37.00	
		Maximum		1020.00	
		Range		983.00	
		Interquartile Range		315.50	
		Skewness		1.528	.550
		Kurtosis		2.521	1.063
	Colon	Mean		457.4118	67.71393
		95% Confidence Interval for Mean	Lower Bound	313.8646	
			Upper Bound	600.9589	
		5% Trimmed Mean		443.6242	
		Median		380.0000	
		Variance		77948.007	
		Std. Deviation		279.19170	
		Minimum		120.00	
		Maximum		1043.00	
		Range		923.00	
		Interquartile Range		448.00	
		Skewness		.811	.550
		Kurtosis		-.278	1.063

Ovary	Mean		851.0000	245.18320
	95% Confidence Interval for Mean	Lower Bound	220.7365	
		Upper Bound	1481.2635	
	5% Trimmed Mean		824.9444	
	Median		622.5000	
	Variance		360688.800	
	Std. Deviation		600.57373	
	Minimum		301.00	
	Maximum		1870.00	
	Range		1569.00	
	Interquartile Range		975.75	
	Skewness		1.142	.845
	Kurtosis		.422	1.741
Stomach	Mean		86.0000	96.04900
	95% Confidence Interval for Mean	Lower Bound	76.7272	
		Upper Bound	495.2728	
	5% Trimmed Mean		254.6111	
	Median		124.0000	
	Variance		119930.333	
	Std. Deviation		346.30959	
	Minimum		25.00	
	Maximum		1112.00	
	Range		1087.00	
	Interquartile Range		358.50	
	Skewness		1.627	.616
	Kurtosis		1.893	1.191

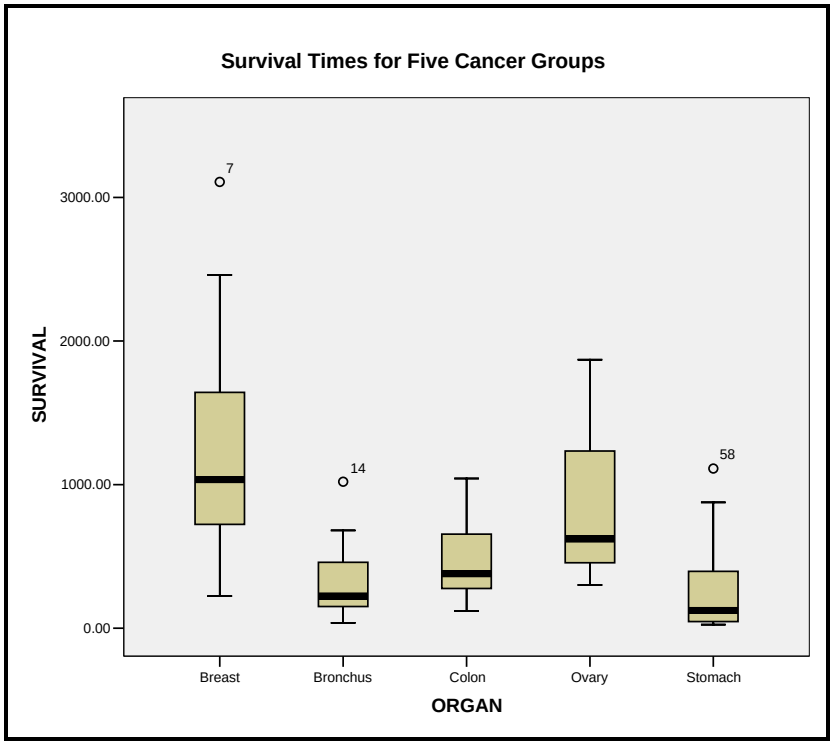
In order to compare the means and standard deviations of the five cancer groups, we extracted the statistics from the descriptive statistics output as following.

Organ	Breast	Bronchus	Colon	Ovary	Stomach
Sample Size	11	17	17	6	13
Mean	1250.4545	305.7059	457.4118	851	286
St. Deviation	903.8123	257.16963	279.1917	600.57373	346.30959

According to the table, we can see that sample sizes are relatively small for all cancer groups. The sample sizes vary from the sample size of 6 for ovary cancer group to 17 for each of the bronchus and colon cancer groups.

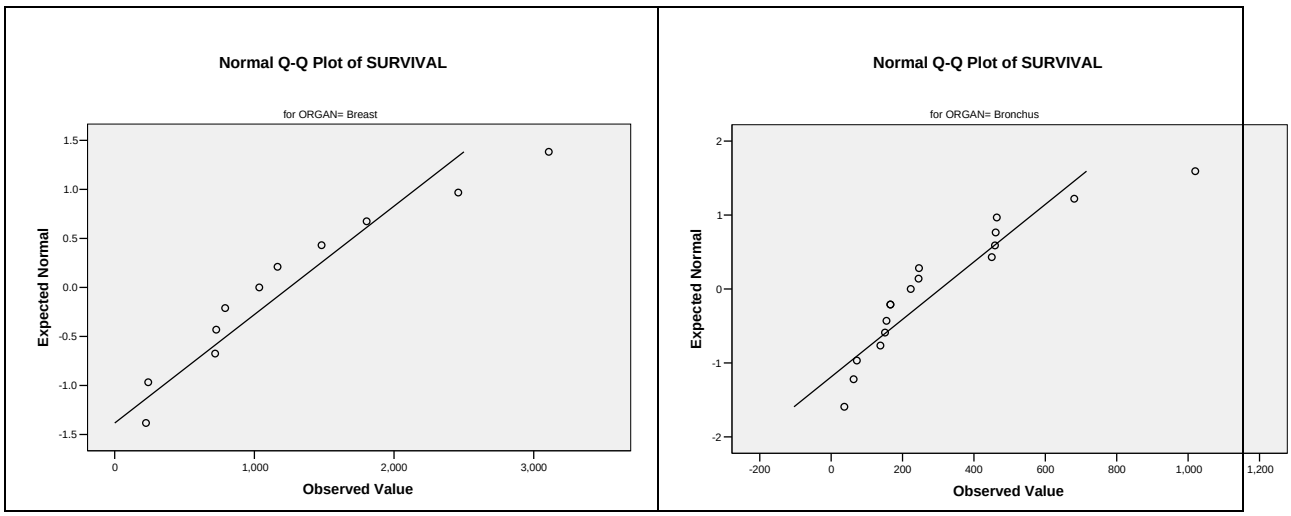
The average survival time for the five cancer groups varies from the lowest of 286 days to the highest of 1250.4545 days. The breast cancer has the largest average survival time among five types of cancers and the colon cancer has the lowest. The standard deviations vary from the lowest of 257.16963 to the highest of 903.8123. Therefore, there are very large differences in the magnitude of standard deviations for the five cancer groups.

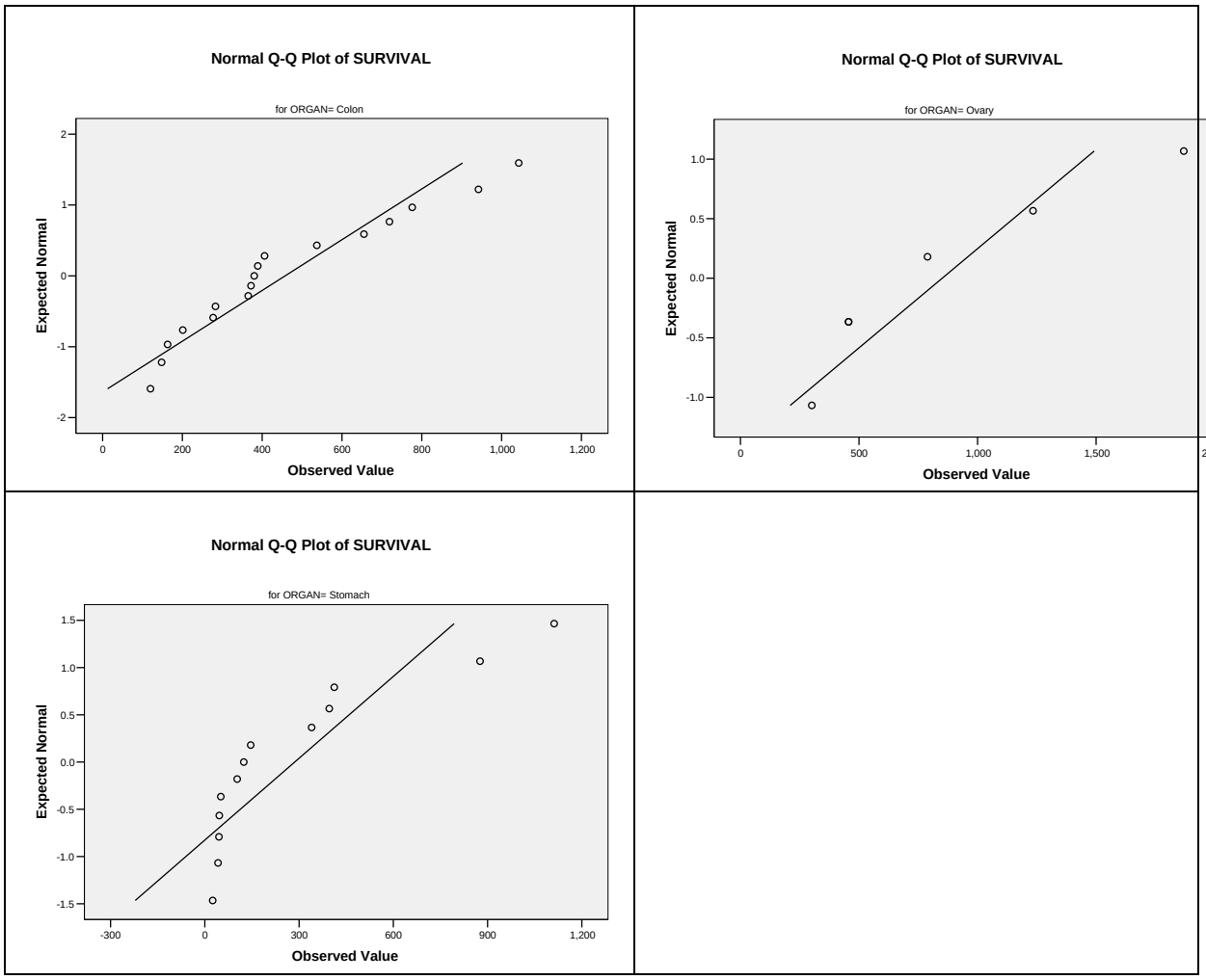
(b) The side-by-side boxplots are displayed below:



The above side-by-side boxplots show that the median survival times of bronchus, colon, ovary and stomach cancers are considerably lower than the median survival time of the breast cancer. The spreads of the distributions are different among five cancer groups. Breast and ovary cancer groups have relatively larger spreads than other cancer groups. All five distributions are seriously skewed to the right. And there are outliers in the breast, bronchus and stomach cancer groups.

(c) The normality plots of the five distributions displayed below indicate some serious departures from a straight-line pattern. There plots show that all the distributions are skewed to the right. Notice that given the relatively small sample size, it is difficult to verify the assumption for the data.





Question 3

(a) The descriptive statistics for the log-transformed data are displayed below:

ORGAN				Statistic	Std.Error
LNSURV	**Breast**	Mean		6.8531	.25461
		95% Confidence Interval for Mean	Lower Bound	6.2858	
			Upper Bound	7.4204	
		5% Trimmed Mean		6.8671	
		Median		6.9422	
		Variance		.713	
		Std. Deviation		.84444	
		Minimum		5.41	
		Maximum		8.04	
		Range		2.63	
		Interquartile Range		.92	
		Skewness		-.550	.661
		Kurtosis		-.230	1.279

Bronchus	Mean		5.3894	.21386
	95% Confidence Interval for Mean	Lower Bound	4.9360	
		Upper Bound	5.8427	
	5% Trimmed Mean		5.4027	
	Median		5.4072	
	Variance		.778	
	Std. Deviation		.88178	
	Minimum		3.61	
	Maximum		6.93	
	Range		3.32	
	Interquartile Range		1.16	
	Skewness		-.254	.550
	Kurtosis		-.270	1.063
Colon	Mean		5.9400	.15706
	95% Confidence Interval for Mean	Lower Bound	5.6071	
		Upper Bound	6.2730	
	5% Trimmed Mean		5.9480	
	Median		5.9402	
	Variance		.419	
	Std. Deviation		.64758	
	Minimum		4.79	
	Maximum		6.95	
	Range		2.16	
	Interquartile Range		1.07	
	Skewness		-.179	.550
	Kurtosis		-.811	1.063
Ovary	Mean		6.5458	.28171
	95% Confidence Interval for Mean	Lower Bound	5.8216	
		Upper Bound	7.2699	
	5% Trimmed Mean		6.5375	
	Median		6.3966	
	Variance		.476	
	Std. Deviation		.69004	
	Minimum		5.71	
	Maximum		7.53	
	Range		1.83	
	Interquartile Range		1.20	
	Skewness		.376	.845
	Kurtosis		-1.251	1.741

Stomach	Mean		4.9679	.34675
	95% Confidence Interval for Mean	Lower Bound	4.2124	
		Upper Bound	5.7234	
	5% Trimmed Mean		4.9514	
	Median		4.8203	
	Variance		1.563	
	Std. Deviation		1.25021	
	Minimum		3.22	
	Maximum		7.01	
	Range		3.80	
	Interquartile Range		2.18	
	Skewness		.298	.616
	Kurtosis		-1.255	1.191

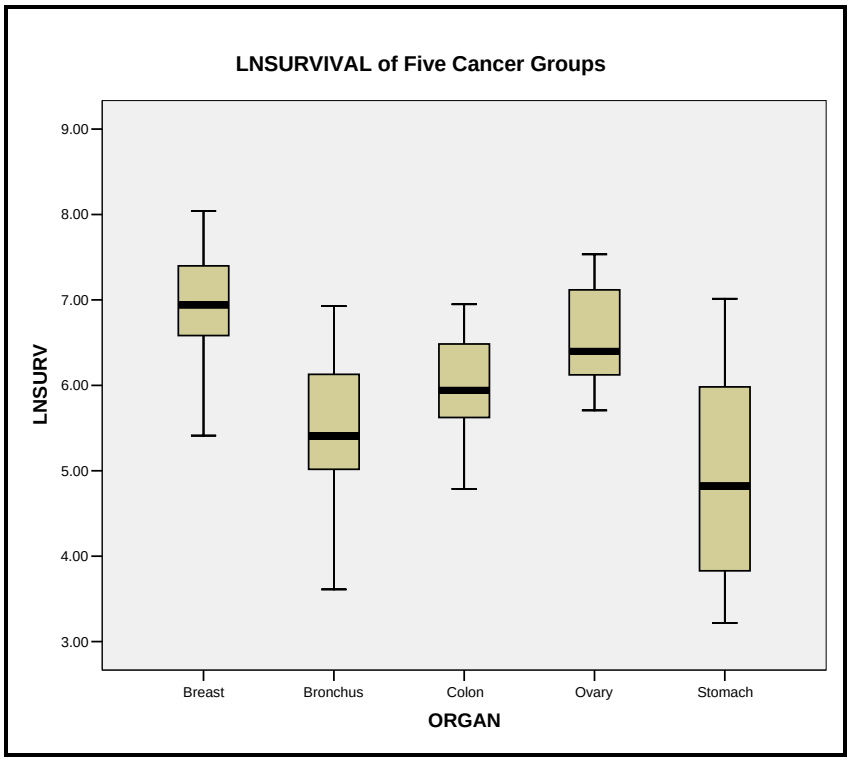
Similarly to the part (a) of Question 2, in order to compare the means and standard deviations of the five cancer groups, we extracted the statistics from the descriptive statistics output as following.

Organ	Breast	Bronchus	Colon	Ovary	Stomach
Sample Size	11	17	17	6	13
Mean	6.8531	5.3894	5.9400	6.5458	4.9679
St. Deviation	0.84444	0.88178	0.64758	0.69004	1.25021

According to the table, we can see that the average survival time on the log scale varies from the lowest of 4.9679 days to the highest of 6.8531 days. The breast cancer has the largest average log survival time among five types of cancers. As the natural logarithm is a monotone function, the order of survival times for the five cancers is the same as for the data on the original scale of measurement.

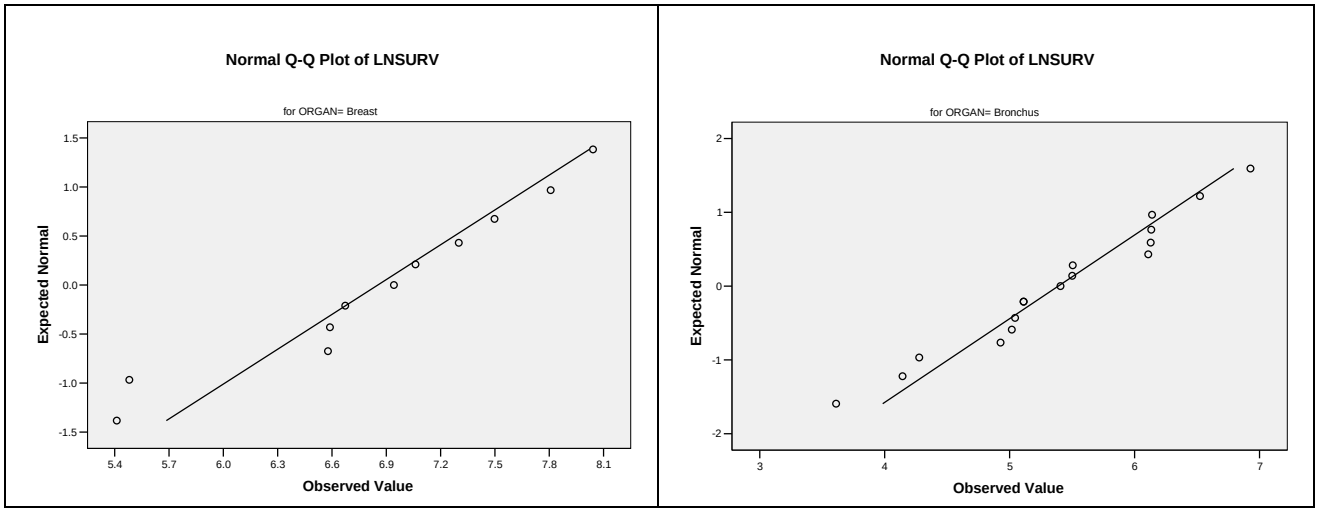
The standard deviations vary from the lowest of 0.64758 to the highest of 1.25021. Given the relatively small sample sizes, there are relatively small differences in the magnitude of standard deviations.

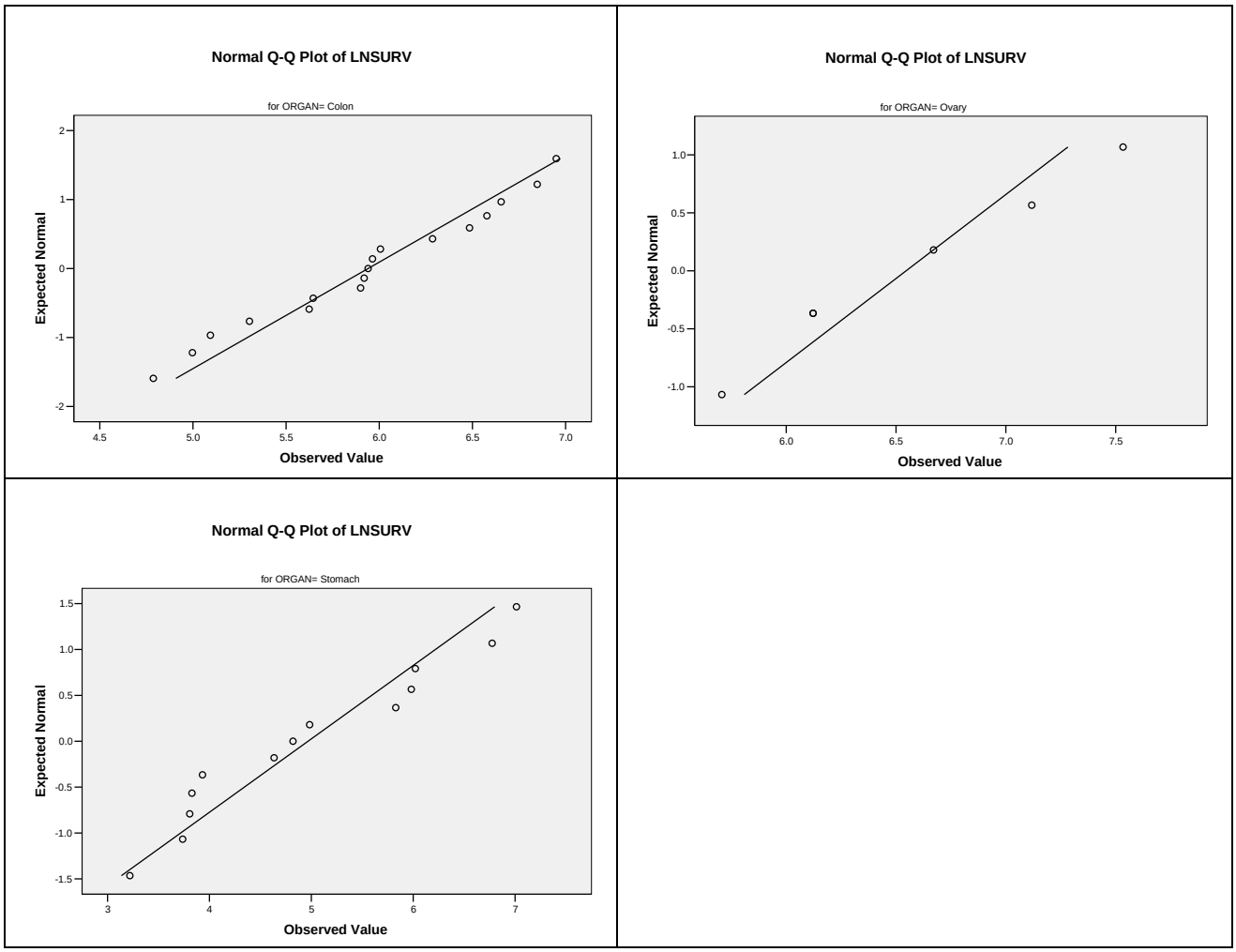
(b) The side-by-side boxplots for the log-transformed data are displayed below:



The median for the stomach cancer group is considerably smaller than the medians for other cancer groups. The spreads of the distributions are similar after the log-transformation among breast, bronchus, colon and ovary cancer groups. Stomach cancer group has a relatively larger spread than other cancer groups. All five distributions are approximately symmetric. Moreover, there are no outliers for all cancer groups.

(c) The Q-Q plots for the log-transformed data are shown below.





As you can see the log transformation successfully removed the skewness present in the data on the original scale of measurement. There are no systematic deviations from a straight-line pattern in all plots, except for the ovary cancer group. However, remember that we just have 6 observations in the ovary group. Because of the very small sample size, the skewness is difficult to evaluate. Therefore, there is no reason to suspect that the normality assumption is seriously violated. It is reasonable to conclude that the assumption of normality is feasible for the log-transformed data in all groups.

Question 4

(a) In general, the assumption of equal standard deviations in the populations is crucial. Nevertheless, ANOVA is robust to minor violations of the assumption. However, serious problems may occur if one of the populations has a very different standard deviation.

According to the side-by-side boxplots, there are no considerable differences in the spread among the five distributions on the log-scale. Based on the above summaries, the ratio of the largest to the smallest standard deviation does not exceed 2. Notice that only stomach cancer group has the standard deviation that is substantially larger than the standard deviation of any of the remaining four groups, though it is based on a sample of only 13 observations.

The Levene's test with the p-value of 0.057 is also consistent with the assumption of equal variances. The output for the Levene's test is provided below:

Test of Homogeneity of Variances

LNSURV

Levene Statistic	df1	df2	Sig.
2.438	4	59	.057

(b) In general, normality assumption is not critical. The normality plots of the other four distributions obtained in Question 3 do not indicate any systematic deviations from a straight-line pattern in any of the plots. Considering the very small sample size in the ovary cancer group, it is very difficult to make any claims about normality for the group.

Question 5

In order to see whether there is any evidence that patients affected by certain cancers survived longer than patients affected by other cancers, we apply *One-Way ANOVA* feature in SPSS. Let μ_i be the mean survival time (on the log scale) of the i-th cancer group, where i=1,2,...,5.

(a) Define the null and alternative hypotheses as follows:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (there is no difference in the mean survival time (on the log scale) among the five cancer groups),

H_a: there are differences in the mean survival time (on the log scale) among the five cancer groups (at least two log-survival means differ).

(b) The ANOVA output is displayed below:

ANOVA

LNSURV

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.730	4	6.932	8.626	.000
Within Groups	47.418	59	.804		
Total	75.148	63			

According to the above output, the sum of squared residuals from fitting the reduced (equal means) model is 75.148. The sum of squared residuals from fitting the full (separate means) model is 47.418. The pooled estimate of the variance is 0.804. The value of the F-statistic is 8.626 and the F statistic follows an F distribution with 4 and 59 degrees of freedom. The p-value of the test is reported as zero.

There is convincing evidence that there are significant differences among the group means. In other word, on the average, patients affected by certain cancers survived longer than patients affected by other cancers.

(c) Given the sums of squared residuals from fitting the full (separate means) and reduced models, we calculate the extra sum of squares as follows:

Extra sum of squares = Residual sum of squares (reduced) - Residual sum of squares (full) or

Extra sum of squares = 75.148-47.418 = 27.73

Pooled estimate of the variance = Residual sum of squares (full) / degree of freedom (full) or

Pooled estimate of the variance = 47.418/59 = 0.8037

The formula for the value of the test statistic F is

$$F = \frac{(Extra\_sum\_of\_squares)/(Extra\_degrees\_of\_freedom)}{pooled\_estimate\_of\_variance}$$

Thus the value of the test statistic F is therefore

$$F = \frac{27.73/(63-59)}{0.8037} = 8.626$$

This value of F is consistent with the value obtained by SPSS.

Let x_{ij} denote the jth observation in the ith group, $\bar{x}_{i.}$ be the sample mean for the ith group, and $\bar{x}_{..}$ be the sample mean for all data.

Then, given x_{ij} , the residuals for the full model (five-mean model) can be calculated by x_{ij} - $\bar{x}_{i.}$, and the residuals for the reduced model (one-mean model) can be calculated by x_{ij} - $\bar{x}_{..}$.

For example, let us calculate the residuals for the first observation in the dataset, which belongs to the breast cancer group with LNSURV=6.9422,

We know that the sample mean of LNSURV is 5.81 ($\bar{x}_{..}$ = 5.81), and the sample mean of LNSURV for the breast cancer group is 6.8531 ($\bar{x}_{1.}$ =6.8531).

Then, the corresponding residual for the full model is 6.9422-6.8531=0.0891, and the corresponding residual for the reduced model is 6.9422-5.81=1.1322.

Similarly, for case #12, it belongs to the bronchus group with LNSURV=6.5236. Given the sample mean of log-survival time for the bronchus group is 5.3894, the corresponding residual for the full model is 6.5236-5.3894=1.1342, and the corresponding residual for the reduced model is 6.5236-5.81=0.7136.

Question 6

After the ANOVA F-test in Question 5 establish that patients with certain cancers have a greater mean survival time than patients with other cancers, you will examine which cancer patients differ in their mean survival times from others by using the Scheffe's multiple-comparison procedure in SPSS.

The output of the Scheffe's Procedure is displayed below:

Multiple Comparisons

Dependent Variable: LNSURV

Scheffe

(I) CODE	(J) CODE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	1.46371*	.34690	.003	.3606	2.5668
	3.00	.91303	.34690	.155	-.1901	2.0161
	4.00	.30731	.45499	.977	-1.1395	1.7541
	5.00	1.88515*	.36727	.000	.7173	3.0530
2.00	1.00	-1.46371*	.34690	.003	-2.5668	-.3606
	3.00	-.55068	.30749	.529	-1.5285	.4271
	4.00	-1.15640	.42571	.132	-2.5101	.1973
	5.00	.42144	.33030	.803	-.6289	1.4718
3.00	1.00	-.91303	.34690	.155	-2.0161	.1901
	2.00	.55068	.30749	.529	-.4271	1.5285
	4.00	-.60573	.42571	.731	-1.9594	.7480
	5.00	.97212	.33030	.084	-.0782	2.0224
4.00	1.00	-.30731	.45499	.977	-1.7541	1.1395
	2.00	1.15640	.42571	.132	-.1973	2.5101
	3.00	.60573	.42571	.731	-.7480	1.9594
	5.00	1.57784*	.44246	.020	.1709	2.9848
5.00	1.00	-1.88515*	.36727	.000	-3.0530	-.7173
	2.00	-.42144	.33030	.803	-1.4718	.6289
	3.00	-.97212	.33030	.084	-2.0224	.0782
	4.00	-1.57784*	.44246	.020	-2.9848	-.1709

*. The mean difference is significant at the .05 level.

The above table provides both the pair-wise confidence intervals for $\mu_i - \mu_j$ and the observed levels of significance of the corresponding tests. Altogether there are 10 confidence intervals. It is well known that a 5% level two-sided significance test rejects a null hypothesis defined as H_0: $\mu_i - \mu_j = 0$ exactly when a 95% confidence interval for $\mu_i - \mu_j$ does not contain zero. This is the reason why the non-rejection of the null hypothesis is associated with a 95% confidence interval containing zero.

Out of 10 confidence intervals, 7 contain zero and 3 do not contain zero. Equivalently, in 7 cases the test failed to establish a difference between the means. More precisely, there are significant differences between the cancer group 1 vs. 2 (breast vs. bronchus), 1 vs. 5 (breast vs. stomach), and 4 vs. 5 (ovary vs. stomach).

LNSURV

Scheffe[a,b]

CODE	N	Subset for alpha = .05		
		1	2	3
5.00	13	4.9679		
2.00	17	5.3894	5.3894	
3.00	17	5.9400	5.9400	5.9400
4.00	6		6.5458	6.5458
1.00	11			6.8531
Sig.		.180	.069	.234

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 11.058.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Based on the above table, we conclude that, on average, patients with breast cancers survived significantly longer than patients with bronchus or stomach cancers. And patients with ovary cancers survived significantly longer than those who had stomach cancers.

In addition, patients with bronchus, colon or stomach cancers are indistinguishable in their mean lnsurvival times from one another. Similarly, patients with bronchus, colon or ovary cancers are indistinguishable in their mean lnsurvival times from one another, and patients with breast, colon or ovary cancers are indistinguishable in their mean lnsurvival times from one another.

Question 7

In order to determine whether patients with advanced cancer of breast, colon or ovary tend to live longer than patients with bronchus or stomach cancers, the null and alternative hypotheses in terms of the population parameters of interest are defined as follows.

$$H_0: (\mu_1 + \mu_3 + \mu_4)/3 - (\mu_2 + \mu_5)/2 = 0$$
$$H_a: (\mu_1 + \mu_3 + \mu_4)/3 - (\mu_2 + \mu_5)/2 > 0$$

The equation in the null hypothesis is equivalent with

$$\frac{1}{3}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{3}\mu_3 + \frac{1}{3}\mu_4 - \frac{1}{2}\mu_5 = 0$$

In order to avoid converting the fraction 1/3 to approximate decimal form 0.333, it is very recommended to express the above contrast in the equivalent form:

$$2\mu_1 - 3\mu_2 + 2\mu_3 + 2\mu_4 - 3\mu_5 = 0$$

The operation also multiplies the value of the contrast and its standard deviation by 6 but does not affect the value of the test statistic and the p-value of the test.

The values 2, -3, 2, 2, and -3 are the coefficients for $\mu_1, \mu_2, \mu_3, \mu_4$ and μ_5 respectively.

In order to carry out a contrast in SPSS, it is necessary to click *Contrasts* in the *One-Way ANOVA* dialog box and input the corresponding coefficient for each cancer group. The outputs are shown below.

Contrast Coefficients

	CODE				
Contrast	1.00	2.00	3.00	4.00	5.00
1	2	-3	2	2	-3

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
LNSURV	Assume equal variances	1	7.6059	1.41388	5.379	59	.000
	Does not assume equal	1	7.6059	1.47280	5.164	34.685	.000

According to the outputs given above, the one-sided p-value of the contrast is reported as zero (0/2=0) no matter equal variances assumed or not. Thus there is convincing evidence to claim that, on average, patients with advanced cancer of breast, colon or ovary tend to survive significantly longer than patients with bronchus or stomach cancers

Question 8

Define the null and alternative hypotheses as follows:

H_0 : all cancer populations possess the same survival time distribution
H_a : the survival time distribution differs with respect to the organ affected by cancer

The p-value of the Kruskal-Wallis test is displayed below is reported as zero.

We conclude that there is convincing evidence that the survival time distribution differs with respect to the organ affected by cancer. The conclusion is consistent with the outcome of the ANOVA F-test in part (b) of Question 5.

LAB ASSIGNMENT 2 MARKING SCHEMA

Proper Header: 10 points

Question 1

(a) Purpose of the study: 2 points
Causal inferences: 2 points
Inferences to the populations: 2 points
(b) Confounding variables: 2 points
How to control the confounding variables: 2 points
(c) Developing consistent criteria for untreatability: 2 point

Question 2

(a) Descriptive Statistics for each group: 5 points
Comparison of means, and standard deviations: 2 points
(b) Side-by-side boxplots: 4 points
Center, spread, and shape: 3 points
Outliers: 1 point
(c) Normality plots: 5 points
Comments about normality: 2 points

Question 3

(a) Descriptive Statistics for each group: 5 points
Comparison of means, and standard deviations: 2 points
(b) Side-by-side boxplots: 4 points
Center, spread, and shape: 3 points
Outliers: 1 point
(c) Normality plots: 5 points
Comments about normality: 2 points

Question 4

(a) Comment on side-by-side boxplots: 2 point
Levene't test: 2 point
(b) Normality assumption: 2 points

Question 5

(a) Null and alternative hypotheses: 3 points
(b) ANOVA output: 3 points
Sums of squares of residuals from fitting the full and reduced model: 2 points
Pooled estimate of the variance: 2 points
Value of the F-statistic and p-value: 2 points
Null distribution: 2 points
Conclusion in plain language: 2 points
(c) Calculating the value of F by hand: 4 points
Calculating the residuals for full and reduced model: 4 points

Question 6

Scheffe's procedure output: 3 points
Comments: 2 points

Question 7

Defining appropriate contrast: 3 points
Contrast output: 2 points
Conclusion: 2 points

Question 8

P-value: 1 point
Conclusions: 2 points
Consistent: 1 point

TOTAL= 112

Lab developed by Grace Liang
Grant: G227150109 – Acc1Time Science
Revised by Henryk Kolacz