1 Medical Statistics: Data analysis with SPSS

Omar Prieto
2019

1.1 Introduction

The dataset used to carry out this analysis was extracted from a study on 83 subjects diagnosed with malignant soft tissue sarcoma. The aim of the study was to look at the clinical utility of a bio-marker called NLR (neutrophil to lymphocyte ratio) in the predictive value for survival and recurrence.

In the following sections, I will show how to extract relevant information from the dataset.

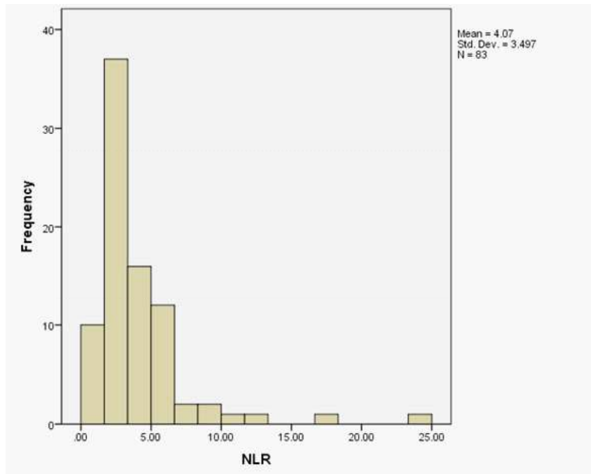
1.2 Data

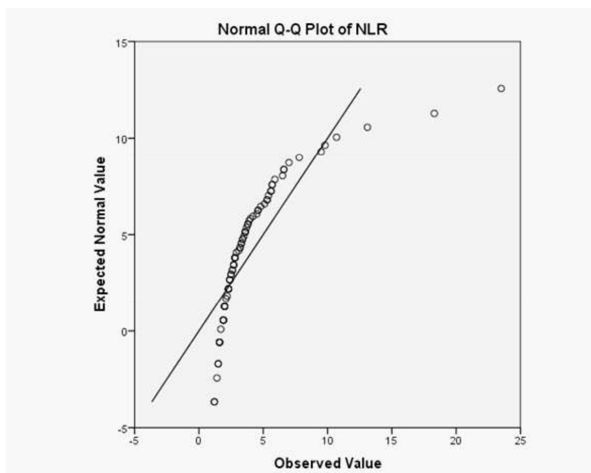
The data is coded as below:

- age - 0: $<$ 65 years old, 1: $>$ 65 years old
- sex - 0: male, 1: female
- anatomiclocation - anatomic location 0: lower, 1: upper
- grade - tumour histology grade 0: low, 1: high
- surgery - type of surgery 0: marginal, 1: wide
- size: size of tumour 0: $<$ 5 mm, 1: $>$ 5mm
- NLR: the NLR ratio in percent
- survival: outcome variable representing 0: survived, 1: deceased
- recurrence: outcome variable representing 0: no recurrence, 1: recurrence
- T_survival: Survival time in days
- T_recurrence: Recurrence time in days

1.3 Checking for normality

Having plotted a histogram and a Q-Q plot, a visual inspection of the NLR data was performed.





The histogram of the variable NLR shows that the data are positive skewed, which suggest that they could not be normally distributed. In addition, the Q-Q plot shows that the curve does not fit into a straight line; consequently there is a clear deviation from the normality, supporting the findings from the histogram. Lastly, a numerical test, in this case the Kolmogorov-Smirnov test, was carried out in order to evaluate the normality more formally.

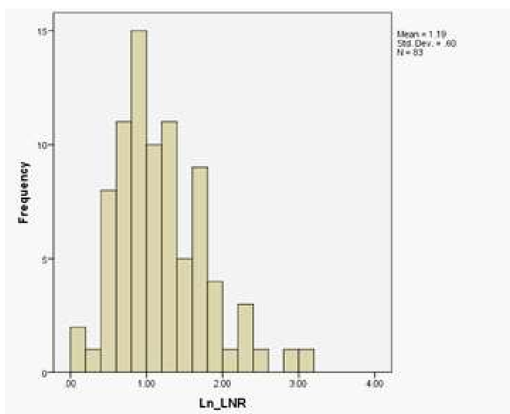
One-Sample Kolmogorov-Smirnov Test

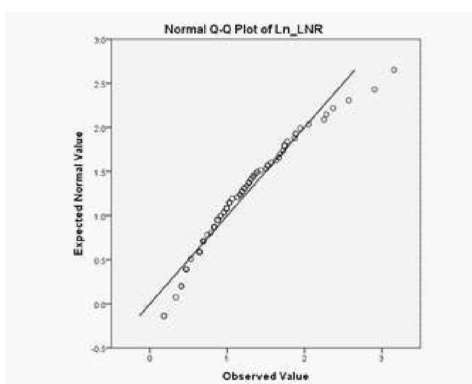
		NLR
N		83
Normal Parameters[a,b]	Mean	4.0675
	Std. Deviation	3.49683
Most Extreme Differences	Absolute	.206
	Positive	.206
	Negative	-.206
Kolmogorov-Smirnov Z		1.881
Asymp. Sig. (2-tailed)		.002

a. Test distribution is Normal.

b. Calculated from data.

The p-value estimated by SPSS is 0.002, so that there is strong evidence against the null hypothesis. Therefore the data significantly deviate from a normal distribution. Next, using a change of variable on NLP taking the natural logarithm, the same analysis was done over the new variable giving the following results:





One-Sample Kolmogorov-Smirnov Test

		Ln_LNR
N		83
Normal Parameters[a,b]	Mean	1.1912
	Std. Deviation	.60019
Most Extreme Differences	Absolute	.100
	Positive	.100
	Negative	-.059
Kolmogorov-Smirnov Z		.912
Asymp. Sig. (2-tailed)		.377

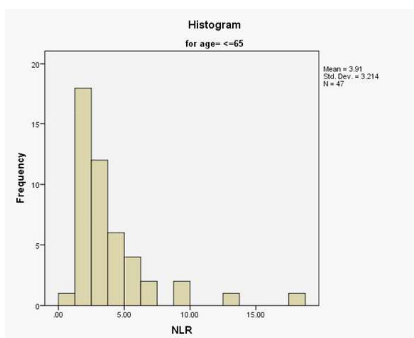
a. Test distribution is Normal.

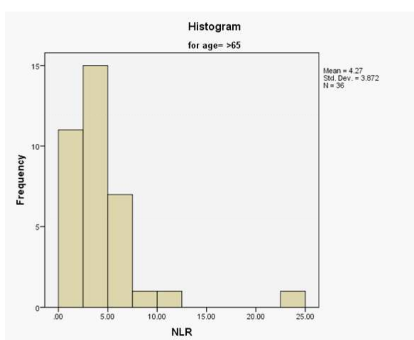
b. Calculated from data.

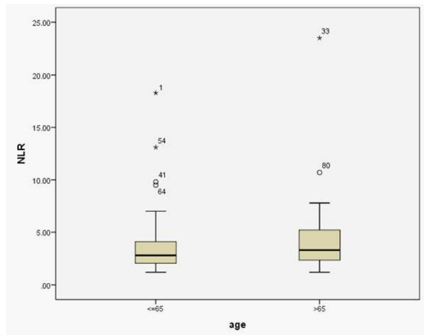
The histogram changed nevertheless still being positive skew. The QQ plot adjusts reasonably to a straight line, which suggest that these data (in this case, the ln of LNR) are normally distributed. The one sample Kolmogorov-Smirnov test for the new variable presents a P value of 0.377. Since the P value is greater than 0.1, there is little evidence against the null hypothesis, cosequently ln_LNR is normally distributed. Taking into account this result and the Q-Q plot, it possible to infer that the new variable (ln_LNR) is normally distributed.

1.4 Data Exploration

Age







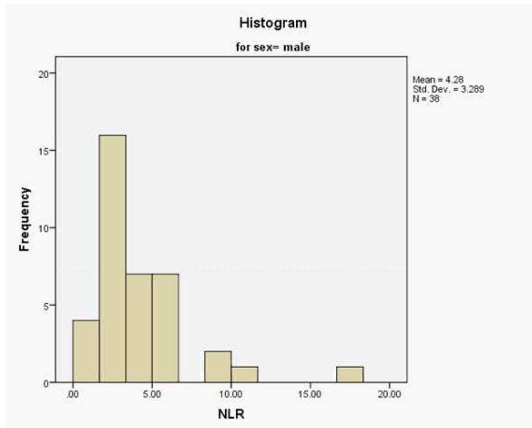
Descriptives

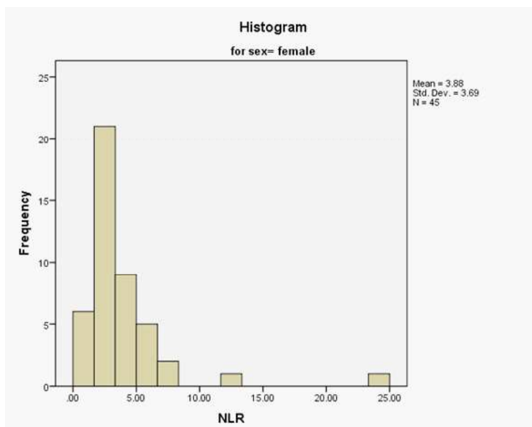
	age			Statistic	Std. Error
NLR	<=65	Mean		3.9128	.46887
		95% Confidence Interval for Mean	Lower Bound	2.9690	
			Upper Bound	4.8566	
		5% Trimmed Mean		3.4494	
		Median		2.8000	
		Variance		10.332	
		Std. Deviation		3.21441	
		Minimum		1.20	
		Maximum		18.30	
		Range		17.10	
		Interquartile Range		2.20	
		Skewness		2.752	.347
		Kurtosis		9.021	.681
	>65	Mean		4.2694	.64537
		95% Confidence Interval for Mean	Lower Bound	2.9593	
			Upper Bound	5.5796	
		5% Trimmed Mean		3.6802	
		Median		3.3000	
		Variance		14.994	
		Std. Deviation		3.87223	

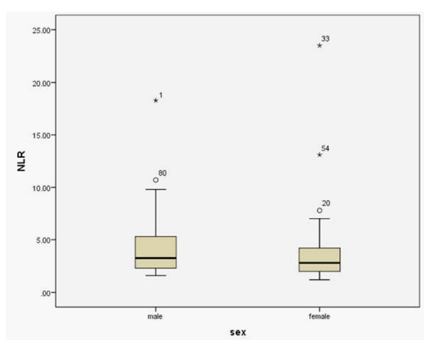
The shape of both histograms is positive skewed and they show that the mean of both groups differ slightly (3.91 and 4.27). On the other hand, the boxplot provides more information: since the boxplots shows significant overlap (including the median) no difference can be claimed between both groups.

Finally, the 95% CI for each mean value showed considerable overlap strengthen the idea that there is no clear difference between groups.

Sex



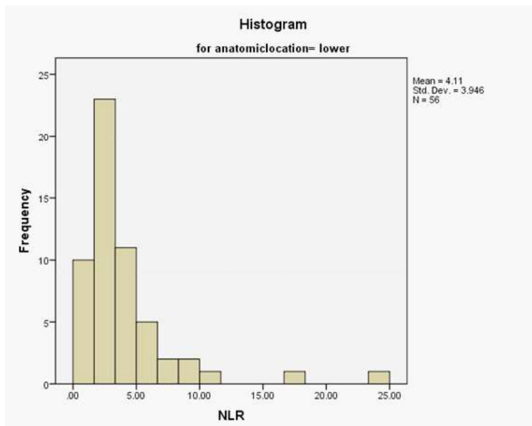


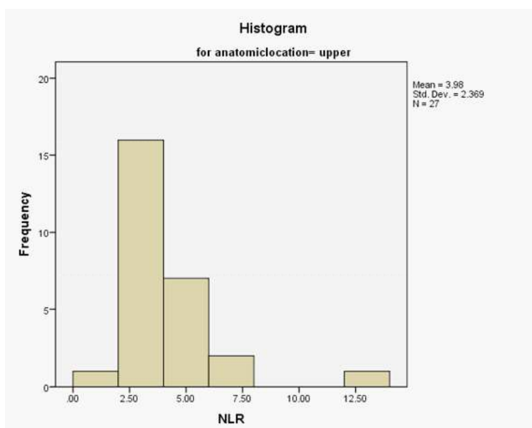


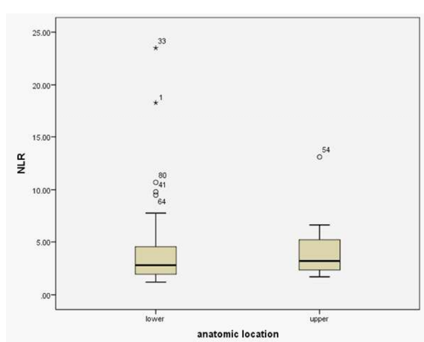
Male: Mean 4.2842, 95% CI (3.2032- 5.3652)
Female: Mean 3.8844, 95% CI (2.7758- 4.9931)

As occurred in the age classification, the histograms of NLR distribution by sex are positive skewed with similar mean as well. According to the boxplot it is possible to note that the median values in both, males and females distributions have almost the same value although there is a significant overlap in both distributions. The 95 % CI for the mean overlap as well so that there is no clear difference between groups.

Anatomic Location



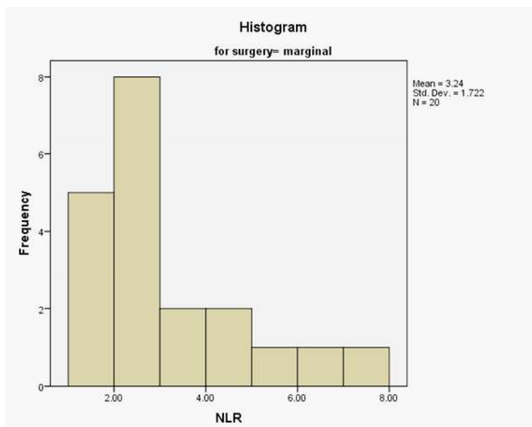


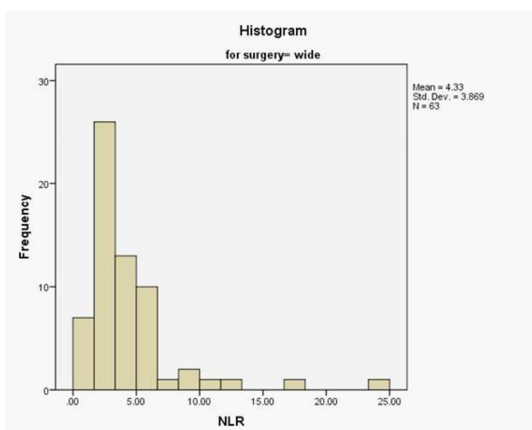


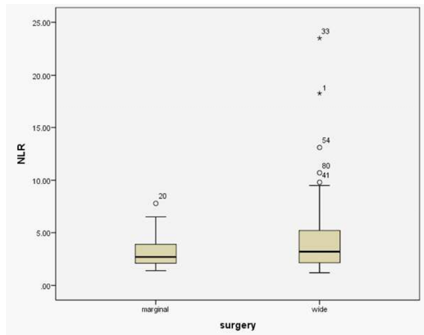
Upper: Mean 3.9778, 95% CI (3.0405- 4.9150)
Lower: Mean 4.1107, 95% CI (3.0540- 5.1675)

The histograms of NLR distribution by anatomic location (upper and lower) have similar shape and mean. As in the previous analysis, due to the overlap in boxplot and confidence intervals, it is possible to conclude that there is not substantial difference between these two groups.

Surgery



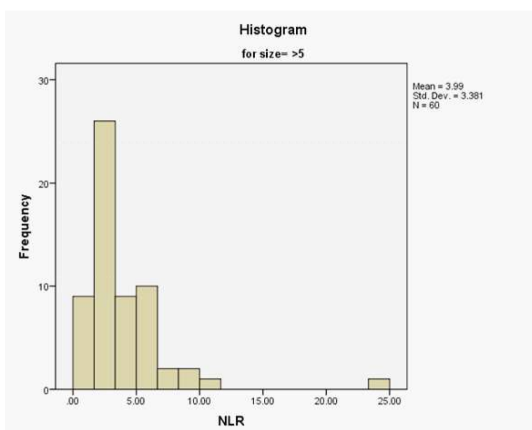


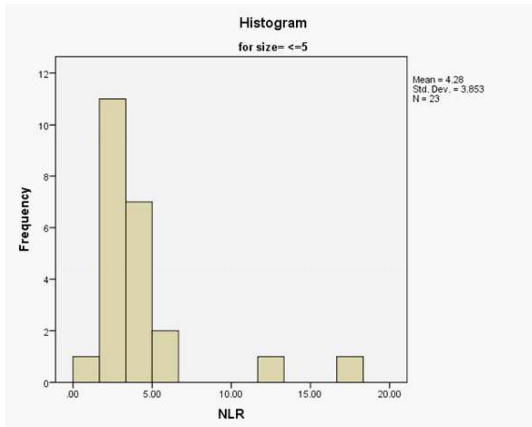


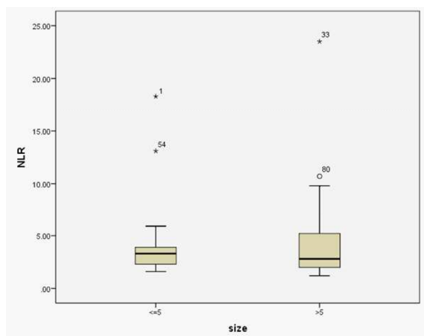
Marginal: Mean 3.2400, 95% CI (2.4342- 4.0458)
Wide: Mean 4.3302, 95% CI (3.3556- 5.3047)

The histograms show a slightly difference in mean, and they draw the same shape. There is a significant overlap in the 95% confidence intervals of the mean. Analysing the boxplots, the median is almost at the same level, which indicates that these values are similar (the classification of surgery wide has a slightly higher value than the classification marginal). Furthermore, the IQR and the distance between whiskers of the variable wide are larger than those of the marginal variable. This spread is predictable due to the high standard deviation compared with the SD of the other group.

Size



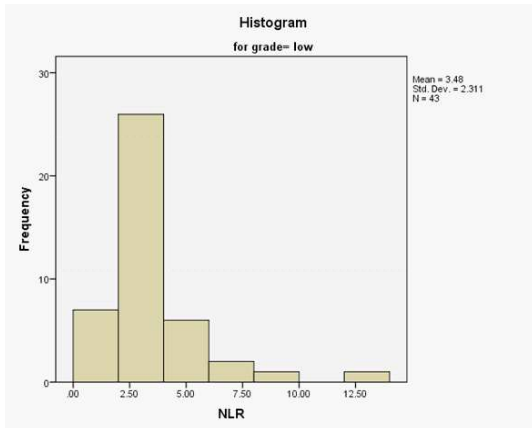


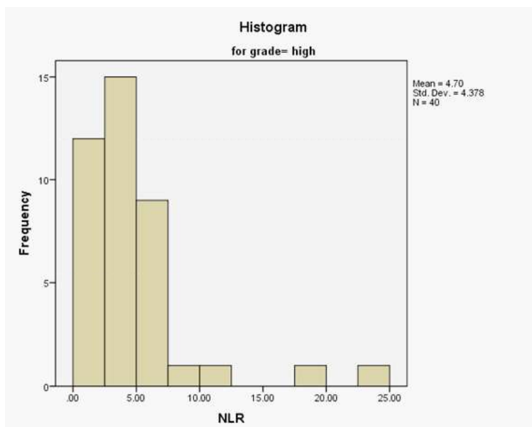


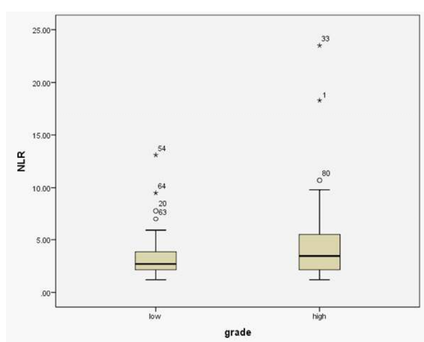
<=5 cm: Mean 4.2783, 95% CI (2.6120- 5.9446)
>5 cm: Mean 3.9867, 95% CI (3.1132- 4.8602)

As in the previous results, the boxplot shows considerable overlap between groups so that we cannot affirm if there is a clear difference in the NLR distribution between both groups.

Grade







Low: Mean 3.4791, 95% CI (2.7679- 4.1903)

High: Mean 4.7000, 95% CI (3.2999- 6.1001)

As occurred in the previous cases there is not substantial difference between groups of this variable.

1.5 Test of significance

A check for significance was performed. The independent-samples t-test can be used to determine whether the difference between two groups is statistically significant, however the assumption of normality is necessary. Due to the data is not normally distributed, a nonparametric alternative, in this case the Mann-Whitney U test, was carried out.

The test was used to test for significant differences in NLR with respect to the other variables (sex, age, etc....). The results are reported in the following tables.

Age

Ranks

	age	N	Mean Rank	Sum of Ranks
NLR	<=65	47	40.44	1900.50
	>65	36	44.04	1585.50
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	772.500
Wilcoxon W	1900.500
Z	-.676
Asymp. Sig. (2-tailed)	.499

a. Grouping Variable: age

Sex

Ranks

	sex	N	Mean Rank	Sum of Ranks
NLR	male	38	44.87	1705.00
	female	45	39.58	1781.00
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	746.000
Wilcoxon W	1781.000
Z	-.997
Asymp. Sig. (2-tailed)	.319

a. Grouping Variable: sex

Anatomic location

Ranks

	anatomic location	N	Mean Rank	Sum of Ranks
NLR	lower	56	39.96	2237.50
	upper	27	46.24	1248.50
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	641.500
Wilcoxon W	2237.500
Z	-1.114
Asymp. Sig. (2-tailed)	.265

a. Grouping Variable: anatomic location

Grade

Ranks

	grade	N	Mean Rank	Sum of Ranks
NLR	low	43	38.62	1660.50
	high	40	45.64	1825.50
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	714.500
Wilcoxon W	1660.500
Z	-1.327
Asymp. Sig. (2-tailed)	.185

a. Grouping Variable: grade

Surgery

Ranks

	surgery	N	Mean Rank	Sum of Ranks
NLR	marginal	20	37.88	757.50
	wide	63	43.31	2728.50
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	547.500
Wilcoxon W	757.500
Z	-.879
Asymp. Sig. (2-tailed)	.379

Size

Ranks

	size	N	Mean Rank	Sum of Ranks
NLR	<=5	23	44.15	1015.50
	>5	60	41.18	2470.50
	Total	83		

Test Statistics[a]

	NLR
Mann-Whitney U	640.500
Wilcoxon W	2470.500
Z	-.504
Asymp. Sig. (2-tailed)	.614

a. Grouping Variable: size

Null hypothesis: The distribution of NLR is the same across each category. In all cases, the P value is greater than 0.05, so that there is no difference in mean NLR score for the different groups classified by sex, age, grade, location size or surgery. This result strengthens the past results on the evaluation of histograms and boxplots.

It is important to note that the P value shown in the previous tables is the 'asymptotic', which means that the p-value approaches the real value as sample size increases. In the previous analysis the asymptotic p-value is a good enough approximation to the real pvalue (both groups in each variable are large enough). This was corroborated with the exact significance value displayed in the hypothesis summary test, which turns out to be the same that the asymptotic.

1.6 Survival Analysis

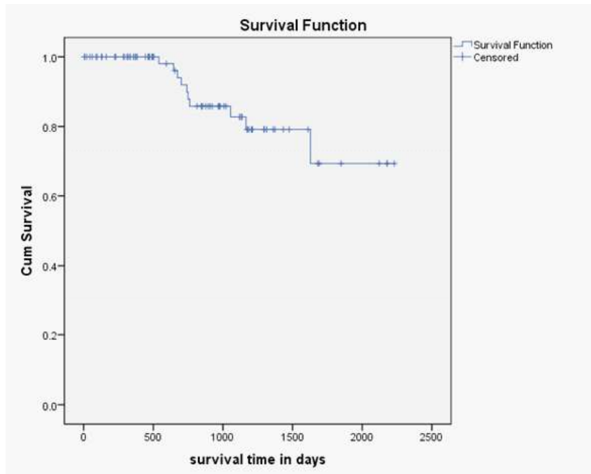
The Kaplan-Meyer is a non-parametric test used to estimate the probability of survival (i.e., not experiencing the event) past given time points, taking into consideration the presence of censored cases. The output obtained from the NLR dataset (whole cohort) is reported in the following tables and plot.

		Censored	
Total N	N of Events	N	Percent
83	10	73	88.0%

Means for Survival Time

Mean			
		95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound
1878.354	100.412	1681.546	2075.161

First table shows that the number of events (or deaths) is 10, from a population of 83 patients, being 73 participants who survived the duration of the study (88%). Second table states that the mean survival time is 1878 days with 95% CI (1682-2075 days).



The plot above is the survival function, which illustrates the probability of survival of the patients over a period of approximately 2500 days. The small vertical lines represent the censored cases, indicating the survival time of surviving patients before the end of the study.

At the beginning the probability is 1, which means that all the patients are alive. Next, approximately in the day 500 the first dead is reported, and it is followed by a decrease in the probability of survivor. The process repeats 9 times more until the end of the study and approximately 0.63 probability of survivor.

1.6.1 Difference between men and women

To investigate if there is any significant difference in survivor between men and women the Kaplan-Meyer test was used on these groups.

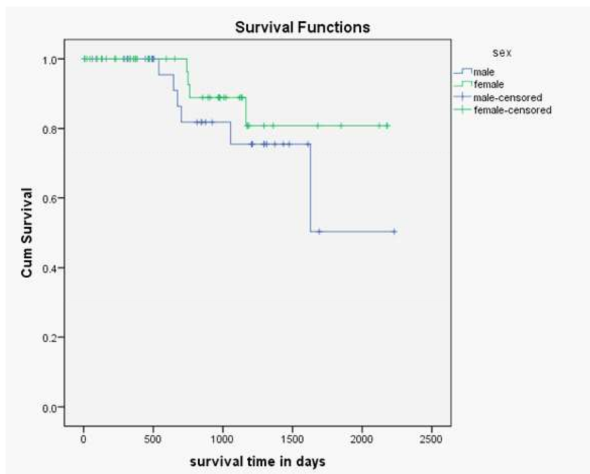
Case Processing Summary

			Censored	
sex	Total N	N of Events	N	Percent
male	38	6	32	84.2%
female	45	4	41	91.1%
Overall	83	10	73	88.0%

The percentage of censored cases present in males (84.2%) and females (91%) was reasonably similar.

Means for Survival Time

sex	Mean			
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
male	1716.161	174.698	1373.753	2058.568
female	1939.246	112.447	1718.850	2159.642
Overall	1878.354	100.412	1681.546	2075.161



The table 'Means for survival time' shows that the means between male and female patients are similar, although the female group has a slightly higher mean. Analysing visually the plot, it is possible to note that there is no a dramatic difference in cumulative survive between both groups, although females patients could have a slightly higher probability of survive (as shown after day 500). In order to assess the difference formally, a log-rank test was carried out.

Pairwise Comparisons

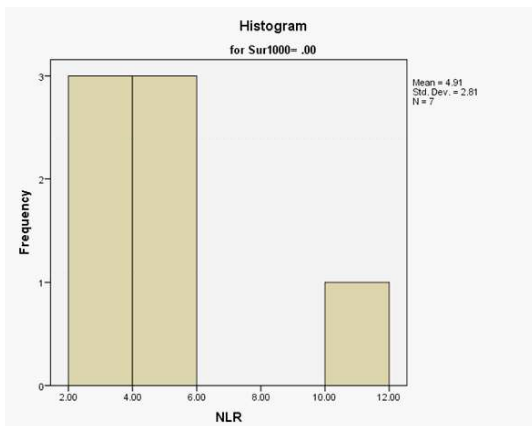
		male		female	
	sex	Chi-Square	Sig.	Chi-Square	Sig.
Log Rank (Mantel-Cox)	male			1.205	.272
	female	1.205	.272		

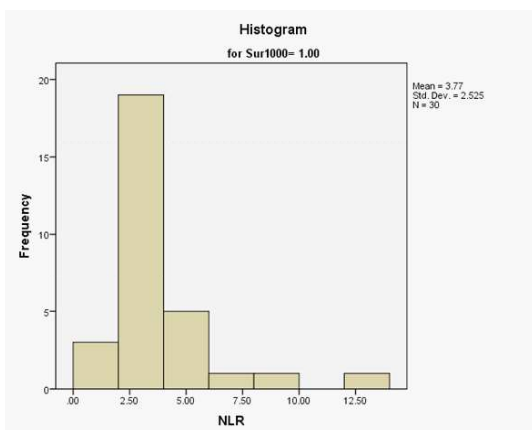
The log rank test was used to test the null hypothesis that there is no difference between the populations in the probability of a death at any time point. Due to the P value is 0.272, there is little evidence against the null hypothesis and therefore, the difference is not significant.

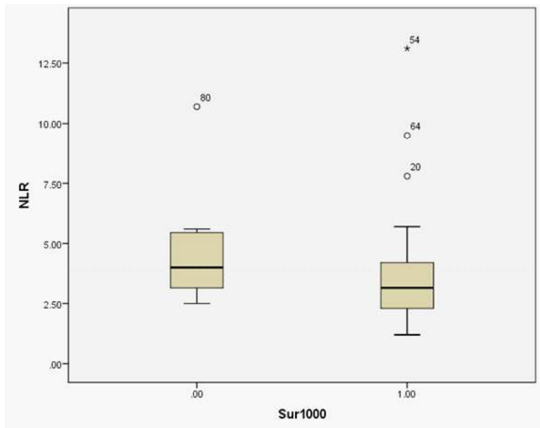
1.7 Univariate analysis

The univariate analysis was performed over a new variable (called Sur1000) defined as follows: 1 for the subjects alive or censored at a time ■ 1000 days, and 0 for those subjects who died at time < 1000 days.

For those subjects who is not possible to know if they experienced the event or not before or after 1000 days (if they were retired from the study for causes other than the cause of interest) they were classified as undefined.







The histograms do not present a clear distribution, particularly the histogram for sur1000=0, which was plotted using 7 element only. This reduction in the number of elements comes from the fact that undefined elements were removed from the set so that the sample size for analysis was reduced. The histogram of sur1000=1 kept more elements (n=30) but its mean presents a high standard deviation.

The boxplot shows that the median value of the sur1000=0 variable is slightly greater than the median value of the variable sur1000=1 and there is significant overlap including the medians, which could mean that there is no difference in both populations. This hypothesis can be tested properly using the Mann-Whitney test in order to obtain a P-value.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of NLR is the same across categories of Sur_1000.	Independent-Samples Mann-Whitney U Test	.138[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

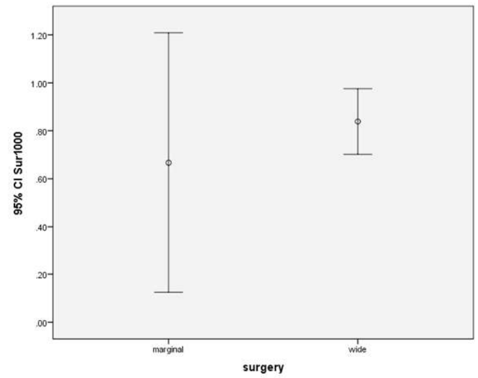
[1]Exact significance is displayed for this test.

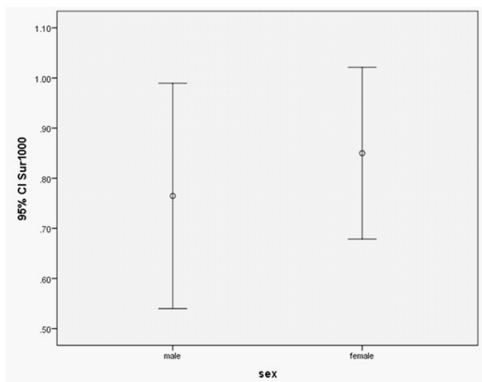
The hypothesis test summary shows the statistical significance for the Mann-Whitney test or P value, in this case 0.138, which means that there is little evidence against the null hypothesis and the distribution is the same.

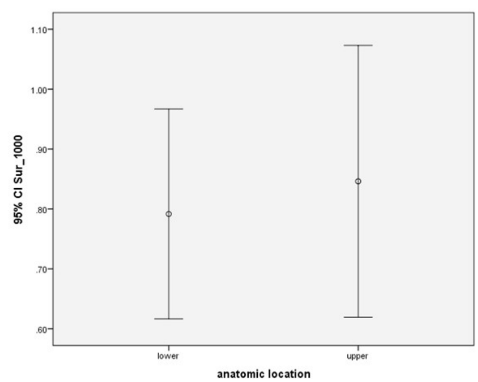
Therefore there is no difference in NLR distribution for subjects alive or censored at or past 1000 days (sur1000=1), and subjects who died before 1000 days (sur1000=0).

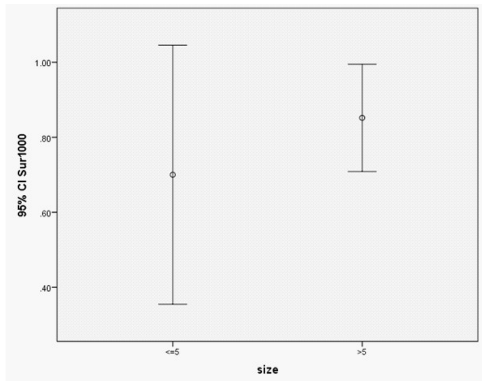
1.7.1 Association of survival at day 1000 with other variables

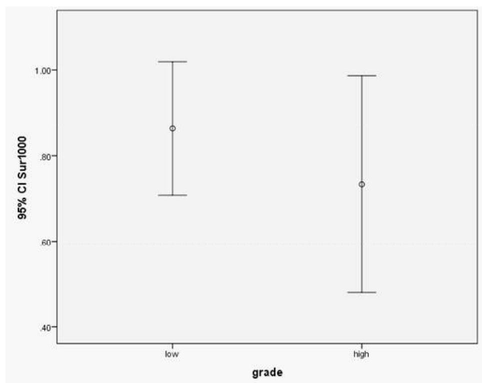
The aim is to analyse the association between survival at day 1000 and the variables: sex, grade, surgery, size, location. First, a visual inspection was made plotting error bars (95%CI) for each variable.

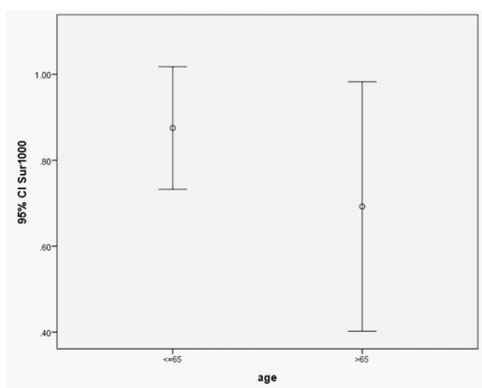

95% CI Sur1000
1.20
1.00
.80
.60
.40
.20
.00
marginal
wide
surgery


95% CI Sur1000
1.10
1.00
.90
.80
.70
.60
.50
male
female
sex


95% CI Sur_1000
1.10
1.00
.90
.80
.70
.60
lower
upper
anatomic location







All the graphs show large confidence intervals, so that there is a large error in the estimate. In addition there is substantial overlap in the error bars. This result suggests that there is not difference between the proportions of survivors at different factor levels. However, it is possible to use a statistical significance test to find more evidence. SPSS is able to test association between two variables (in this case, two variables with two groups in each one) using crosstabulation (a 2x2 contingency table).

The results of the test for the age variable is shown in the following tables as an example; the tables of the other variables are very similar.

Age

age * Sur1000 Crosstabulation

Count

		Sur1000		Total
		.00	1.00	
age	<=65	3	21	24
	>65	4	9	13
Total		7	30	37

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.835[a]	1	.176		
Continuity Correction[b]	.837	1	.360		
Likelihood Ratio	1.760	1	.185		
Fisher's Exact Test				.213	.179
Linear-by-Linear Association	1.785	1	.182		
N of Valid Cases	37				

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for age (<=65 / >65)	.321	.059	1.739
For cohort Sur1000 = .00	.406	.107	1.546
For cohort Sur1000 = 1.00	1.264	.853	1.872
N of Valid Cases	37		

Usually the chi-square with 1 degree of freedom could be used for the same propose, nonetheless in this practical, due to cell frequencies less than 5 are expected, the Fisher exact test is more accurate.

Therefore the results considered are those in the column "Fisher exact test statistic", under the null hypothesis that there is no difference in survival between both groups.

Next table summarises the results for all the variables.

Variable	Fisher's test statistics (2-sided)	Odds ratio (OR)	95% CI for OR
Age	0.213	0.321	(0.059-1.739)
Sex	0.680	1.744	(0.331-9.189)
Size	0.360	2.464	(0.441-13.755)
Grade	0.408	4.434	(0.082-2.309)
Surgery	0.315	2.600	(0.370-18.249)
Anatomic l.	1.00	1.447	(0.239-8.757)

The table includes the odds ratios and the 95% CI for OR. Due to the CI for OR includes 1 (in all cases) the finding is not statistically significant.

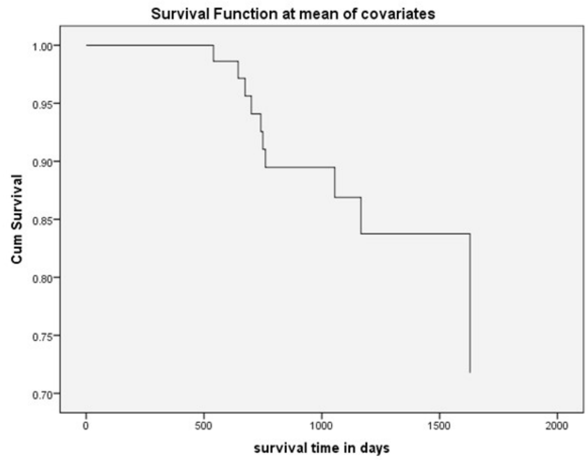
According to the previous table, there is no significant association between survival at 1000 days and the other variables.

1.8 Multivariate analysis

The Cox Regression procedure is useful for modeling the time to a specified event, based upon the values of given covariates. This model assesses relationship between survival time and covariates. In addition it assumes that the ratio of hazard in one group to another remains the same throughout the follow-up period; for example, that the hazard ratio for men versus women is constant over time. The Cox regression was carried out, using a process that involves steps to discriminate the significant predictors, i.e. using the Likehood Ratio test. SPSS outputs a table called variables in the equation showed below.

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	age	1.431	.792	3.261	1	.071	4.183
	sex	-.407	.799	.259	1	.611	.666
	anatomiclocation	-.171	.835	.042	1	.838	.843
	grade	1.277	.811	2.483	1	.115	3.587
	surgery	-.714	.860	.688	1	.407	.490
	size	-.522	.807	.419	1	.518	.593
	NLR	.080	.089	.809	1	.368	1.084
Step 2	age	1.406	.783	3.221	1	.073	4.079
	sex	-.466	.757	.378	1	.539	.628
	grade	1.214	.747	2.640	1	.104	3.365
	surgery	-.800	.753	1.128	1	.288	.449
	size	-.523	.808	.420	1	.517	.593
	NLR	.083	.089	.869	1	.351	1.086
Step 3	age	1.568	.769	4.157	1	.041	4.796
	grade	1.374	.704	3.811	1	.051	3.953
	surgery	-.680	.715	.905	1	.341	.507
	size	-.463	.799	.336	1	.562	.629
	NLR	.097	.086	1.277	1	.259	1.101
Step 4	age	1.438	.723	3.951	1	.047	4.213
	grade	1.277	.665	3.685	1	.055	3.587
	surgery	-.747	.702	1.132	1	.287	.474
	NLR	.111	.083	1.807	1	.179	1.118
Step 5	age	1.499	.718	4.357	1	.037	4.478
	grade	1.325	.662	4.015	1	.045	3.764
	NLR	.106	.081	1.697	1	.193	1.112
Step 6	age	1.312	.680	3.723	1	.054	3.712
	grade	1.282	.654	3.845	1	.050	3.605



According to the table: sex, size, anatomic location, surgery and NLR are not significant predictors (P>0.1), so they were removed from the variables in the equation during each step. Age and grade might have effect in survival since their beta values are positive, which is usually associated with increased hazard and decreased survival times. However, their P values at the end of the process are 0.054 and 0.050 (0.1 ■P<0.05) and its role affect survival is questionable.

To understand the effect of individual predictors on the hazard, it is necessary to take into account the hazard ratio Exp(B), which can be interpreted as the predicted change in the hazard for a unit increase in the predictor. In table (step 6) the variables age and grade have hazard ratio greater than the other variables, so that an increment by 1 unit increases the hazard 3.7 times more.

1.9 Bibliography

[1] Clinical implication of pre-treatment Nuetrophil lymphocyte ratio in soft tissue sarcoma,
O.K. Idowu, Q. Ding, A.F.G. Taktak, C.R. Chandrasekar, Q. Yin (2012)
Biomarkers, 17,6, 539-544