<SIDS31081 Statistic Analysis Report>

The statistic analysis of survival ratio on TITANIC

17, December 2009

Yoonsun OH

Adrian VLAD

Contents

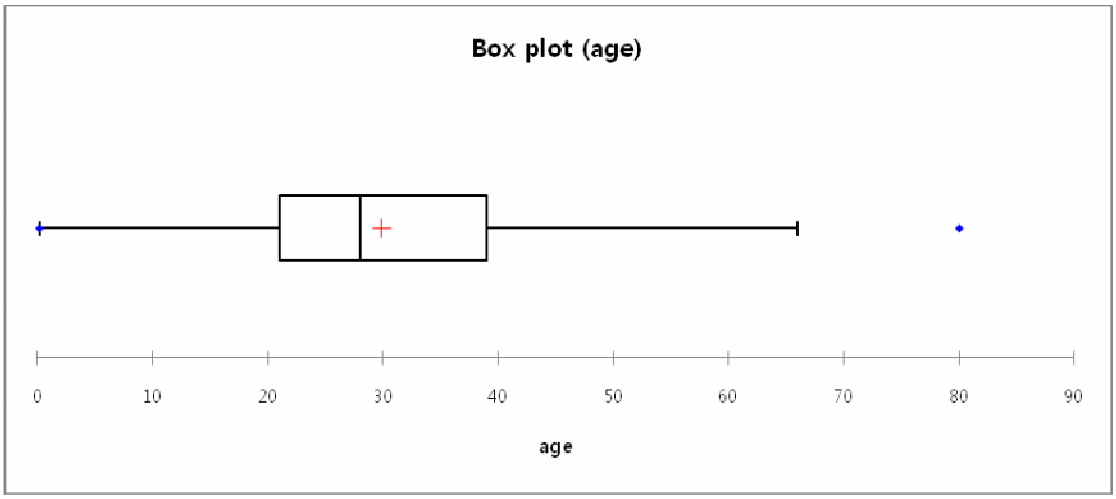
1. General information

There were various kinds of passengers on TITANIC in terms of the factors, which are age, gender and ticket fare; moreover, survival ratio is one of the key quantitative elements to analyze the information of TITANIC data.

1.1. Age

Statistic	age
1st Quartile	21.000
Median	28.000
3rd Quartile	39.000
Mean	29.881
Kurtosis (Pearson)	0.141
Kurtosis (Fisher)	0.147

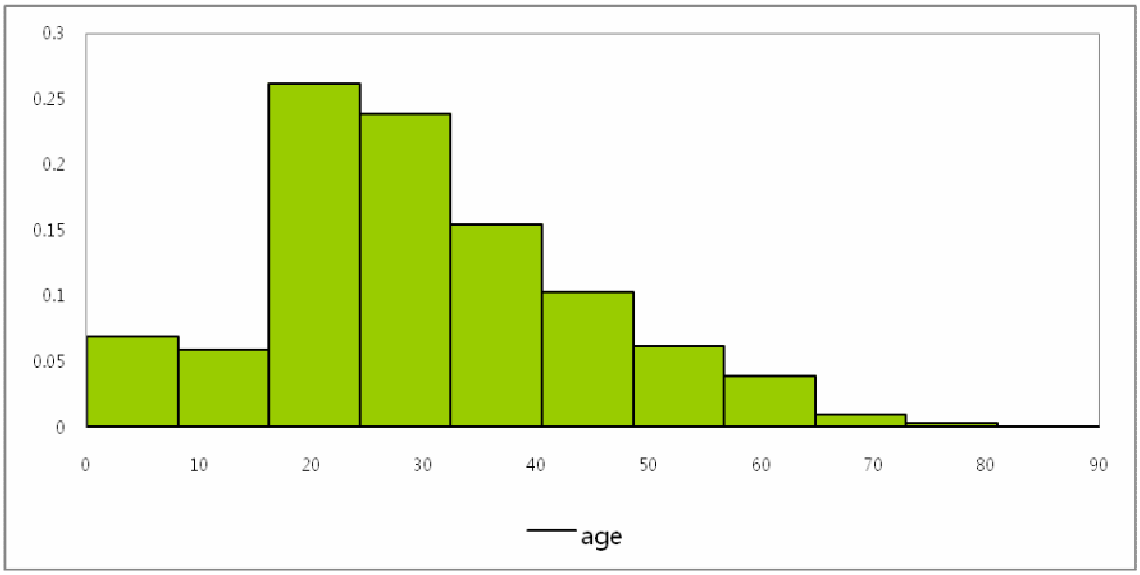
Average age of the passengers of TITANIC was **29.88** years old. Considering the fact that 25% of the passengers were younger than 21 (1st quartile) and only 25% of passengers were older than 39 (100% - the 3rd quartile), half of the passengers' age was between 21 and 39 (3rd quartile – 1st quartile), which indicates that the population of TITANIC was quite young.



The range of the age was pretty wide (79.833), since the youngest baby was 0.167 and the oldest was 80. Therefore, even though the majority of passengers were young, each category of ages from 0 to 80 had some representatives.

General equilibrium of the age (mean) is greater than the median, which means that the whole tendency of the age goes rightward, resulting in positive skewness (rightward skewed

distribution). As we can see in bar chart below, there is a high frequency in ages 20 and 30, resulting in an asymmetry distribution. Age 80 in this case is an outlier; its value is too high with respect to the other ages. There was only one person who was 80, out of 1309 passengers.

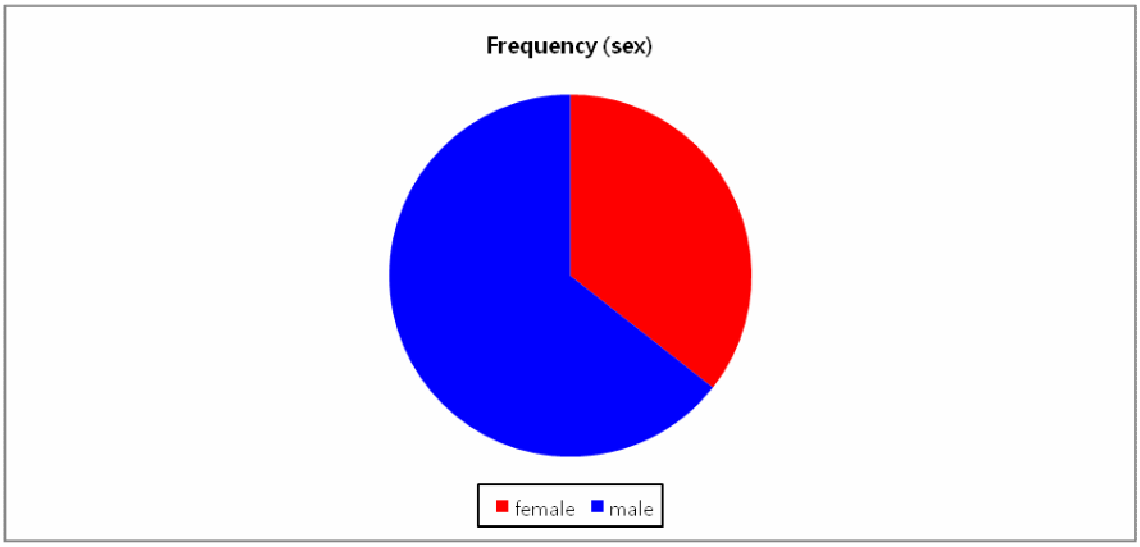


The kurtosis index is low (0.14) suggesting a hypo-normal (more flat) distribution relative to a normal distribution. In other words, the data have a flat top near the mean rather than a sharp peak.

1.2. Gender

The data inferred that there were more males than females on TITANIC. More than half of the population of TITANIC was male compared to only about 40% of the passengers who were female.

Category	Rel. frequency per category (%)
female	35.600
male	64.400

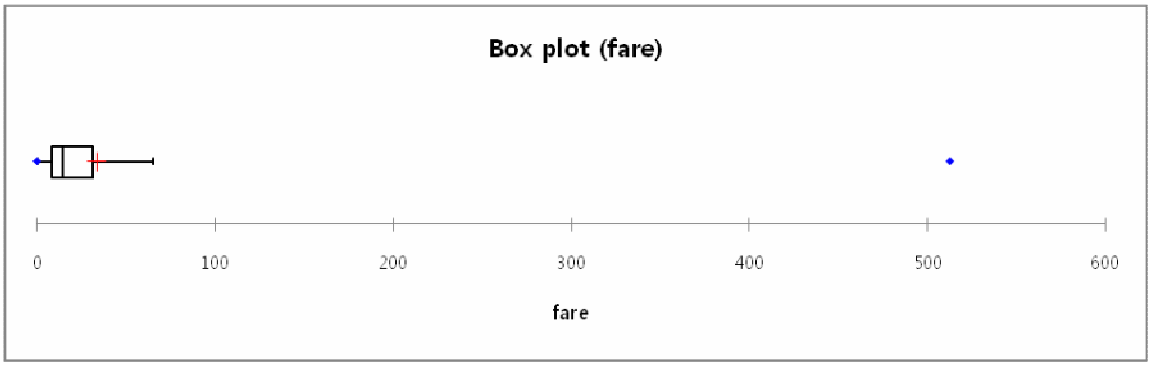


1.3. Ticket Fare

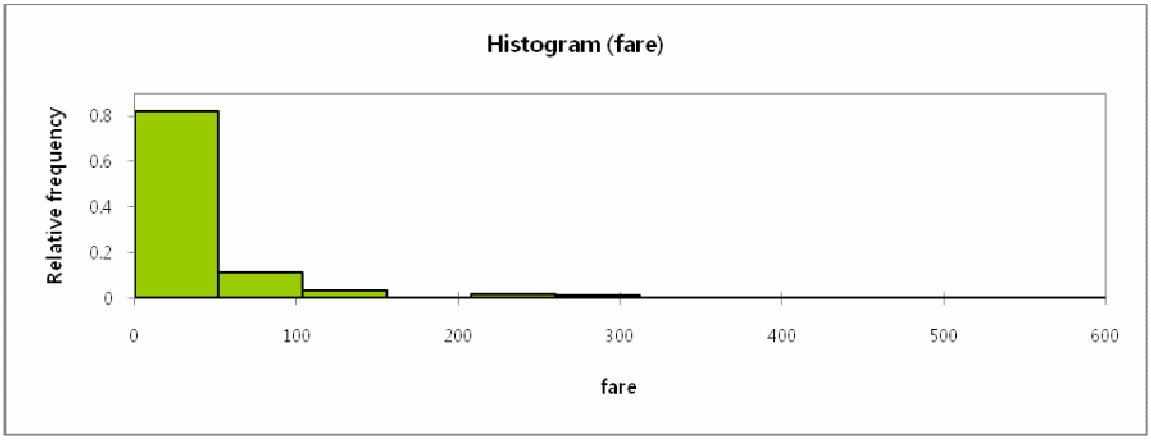
There were considerable variances in ticket fare on TITANIC.

Statistic	fare
Minimum	0.000
Maximum	512.329
Range	512.329
1st Quartile	7.896
Median	14.454
3rd Quartile	31.275
Mean	33.295

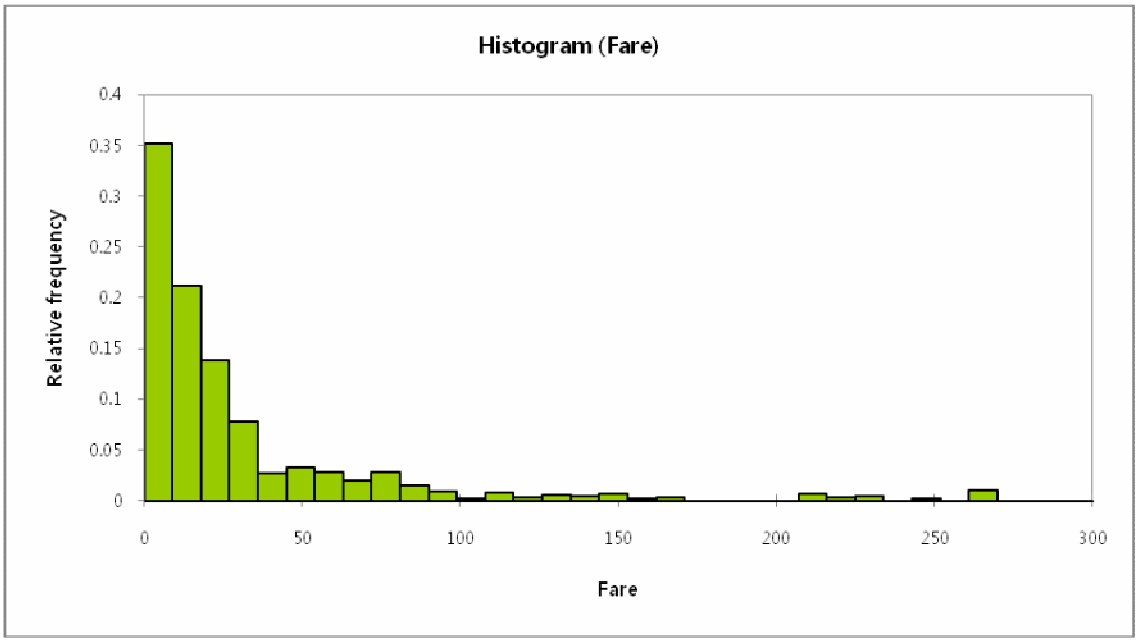
The minimum ticket price was 0 which means that some passengers did not pay anything for the trip. Average ticket price was about 33 dollars which is quite low considering the most expensive fare was 512 dollars. That is because 75% of the passengers (3rd quartile) bought the ticket below the price of 31.27 dollars , 25% of them (the rest) paid from 32 dollars to 512 dollars. One of the main reasons the mean is not an appropriate measur e is its sensitivity to extreme values. This is certainly the case with our data which comprises some extreme prices.



Therefore, we can see the outlier (512) in the box plot which is way higher compared to the mean of the ticket fare. This means that a very select clientele paid extremely much, perhaps being offered the appropriate luxury. Additionally the distribution is asymmetric, rightward (positive) skewed because the mean is greater than the median; general equilibrium is much higher the middle number. Upper limit is quite wide since, 3rd quartile is high, respectively 1st quartile is only 7 dollars.



As we can see in Histogram above, the majority of the passengers bought their ticket under the 50 dollars threshold. Between 32 dollars and 512 dollars ticket, many of them sold by price about 200-250 dollars. In fact only 4 people got the 512 dollar ticket out of 1309 passengers. As a result, most of the passengers on TITANIC had comparatively cheap tickets, even though a select few enjoyed the luxury travel that an expensive ticket bought them. Below, here's a more representative Histogram of fares, one in which we have purposely omitted these luxury ticket prices, and we have increased the number of classes:



Other Statistics:

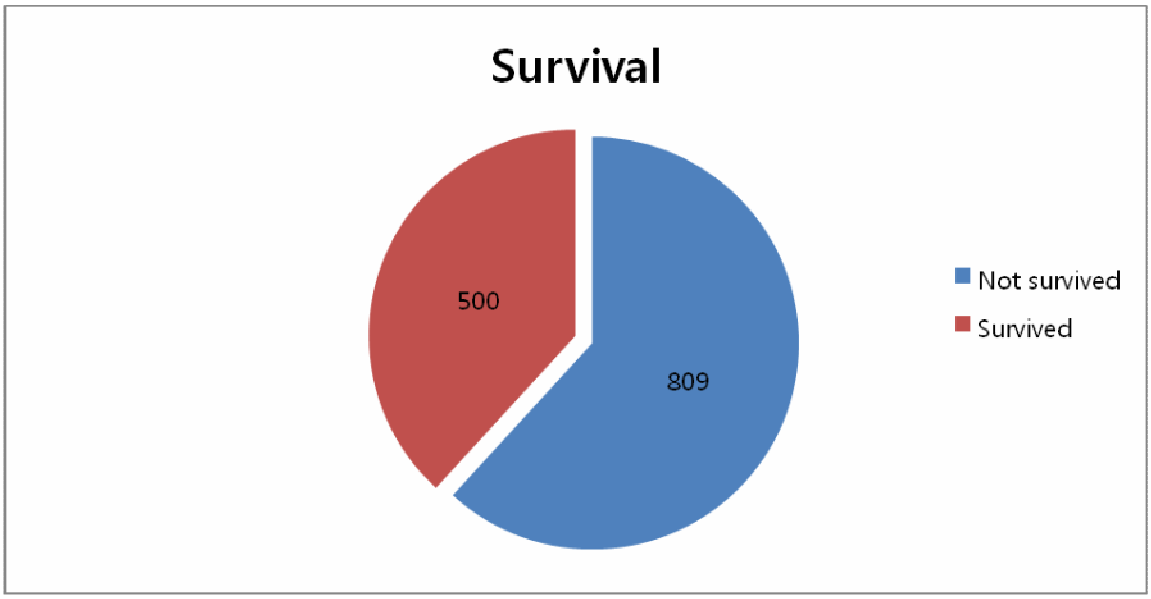
Statistic	fare
No. of observations	1309
No. of missing values	1
Sum of weights	1308
Freq. of minimum	18
Freq. of maximum	4
Variance (n-1)	2678.960
Standard deviation (n-1)	51.759
Variation coefficient	1.554
Skewness (Fisher)	4.368
Kurtosis (Fisher)	27.028

The high kurtosis (**27.028**) tells us that we are dealing with a hyper -normal distribution, which means that the TITANIC data is very peaked relative to a normal distribution. More specifically, the data have a distinct peak near the mean and then decline rapidl y, having heavy tails.

1.4. Survival ratio

The most important aspect of our analysis of the TITANIC data concerned the survival ratio and if there was a statistical link between this ratio and other variables discussed above.

Analyzing this parameter we treated the non-survival event with the value 0 and the survival event with the value 1 ; average ratio of the survival was only 0.382. In other words, only 500 people survived out of the total of 1309.



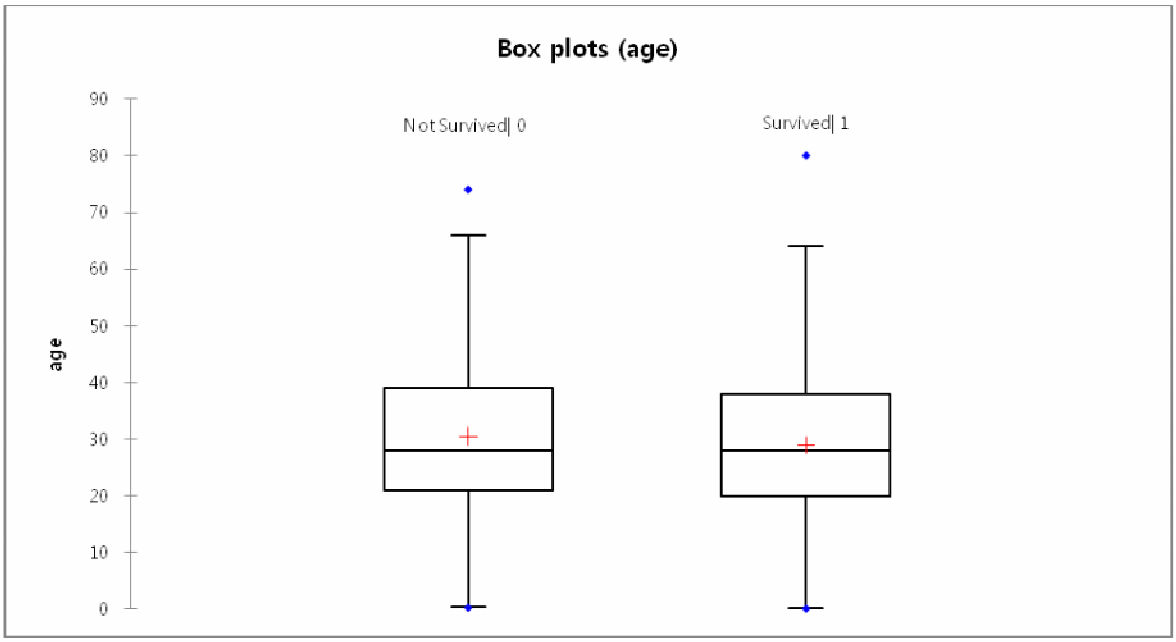
2. Comparison

2.1.1. Age and Survival

Statistic	age | 0	age | 1
No. of observations	809	500
No. of missing values	190	73
Range	73.667	79.833
1st Quartile	21.000	20.000
Median	28.000	28.000
3rd Quartile	39.000	38.000
Mean	30.545	28.918
Variance (n)	193.524	226.317
Variance (n-1)	193.837	226.848
Standard deviation (n)	13.911	15.044
Standard deviation (n-1)	13.923	15.061
Variation coefficient	0.455	0.520
Lower bound on mean (95%)	29.446	27.486
Upper bound on mean (95%)	31.644	30.351

We have compared the age variation of those who survived with the age var iation of those who did not to see if and how survival depends on the age of the population of TITANIC. We have found that survival does not depend on the gender of the people as the statistics are extremely similar in both cases. The average age of the su rvival group was about 29, with a

variance of 226.3, while the deceased group had, on average, about 30 years, with a variance of 193.5. Standard deviation was 15 and 13.9 respectively.



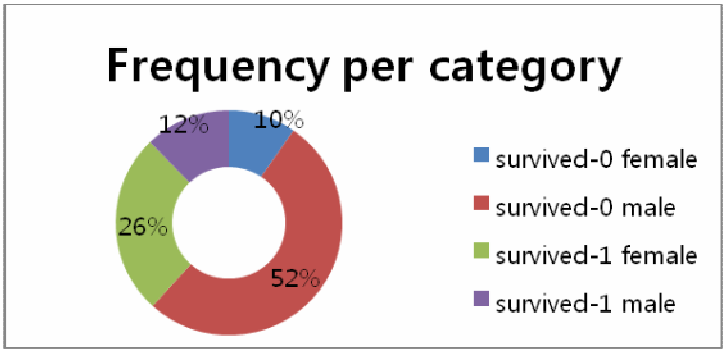
2.1.2. Gender and Survival

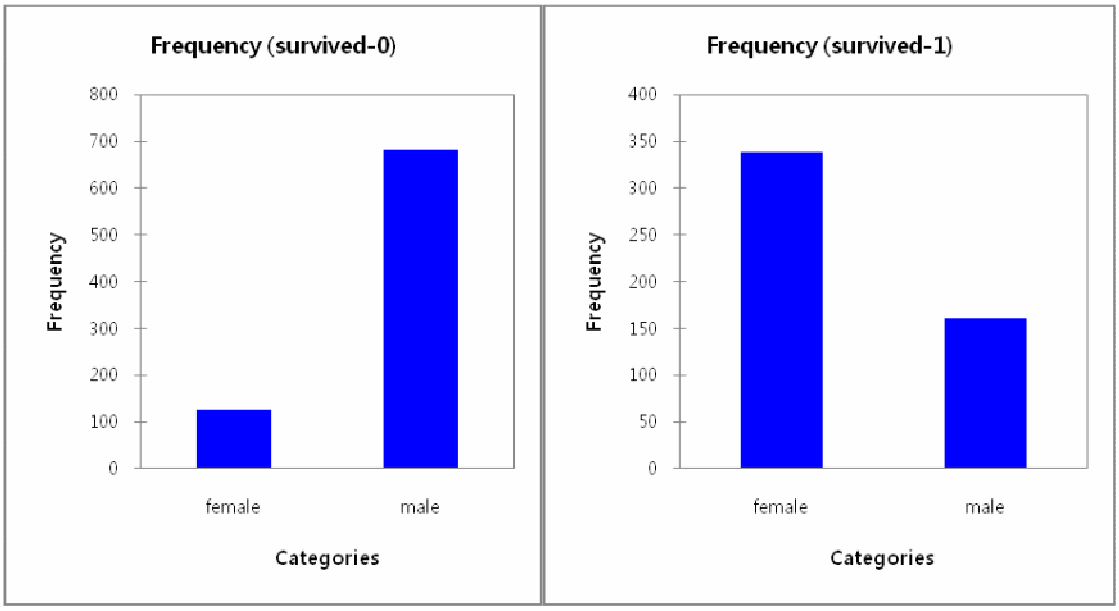
We also compared the relative frequency of gende r and survival, to see which gender group formed which category. Here is what we found:

Descriptive statistics (Qualitative data):

Sample	No. of observations	missing values	Sum of weights	No. of categories	Mode	Mode frequency	Category	Frequency per category	Rel. frequency per category (%)
sex | survived-0	1309	500	809	2	male	682	female	127.000	15.698
							male	682.000	84.302
sex | survived-1	1309	809	500	2	female	339	female	339.000	67.800
							male	161.000	32.200

Basically, 68% of the survivals were females while 32% were males. On the other hand, 16% of the passengers that did not survive were female, while 84% were males. Ho wever, from this statistic we cannot say in particular that survival depended on one gender or another.





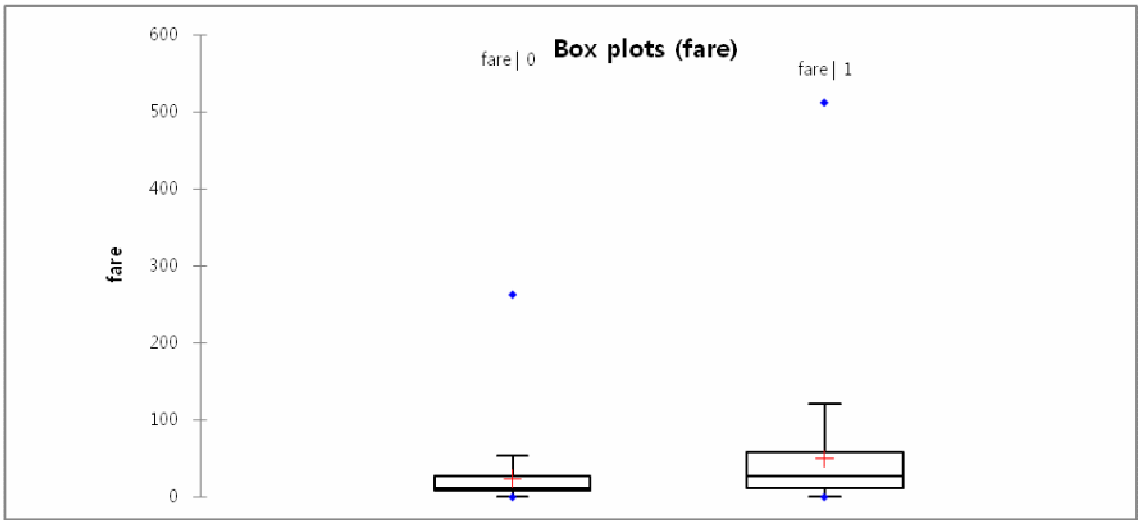
2.1.3. Ticket Price and Survival

Last, we compared ticket price paid and survival, to see if there is any dependency between the two, or more importantly to answer the question: Did the passengers of TITANIC buy a ticket to "survival"?

Statistic	fare | 0	fare | 1
No. of observations	809	500
Minimum	0.000	0.000
Maximum	263.000	512.329
Freq. of minimum	516	811
Freq. of maximum	2	4
Range	263.000	512.329
1st Quartile	7.854	11.215
Median	10.500	26.000
3rd Quartile	26.000	57.750
Sum	18869.895	24680.592
Mean	23.354	49.361
Variance (n)	1164.445	4703.232
Variance (n-1)	1165.888	4712.657
Standard deviation (n)	34.124	68.580
Variation coefficient	1.461	1.389

We found out that there was a difference in the prices of those who survived, namely they paid more for their tickets. However, we cannot infer that survival depended on the ticket price because of the high variability of those who survived. The ma ximum price paid by a

survivor was 512 while the maximum paid by a non -survivor was 263. This doesn't tell us much except that all 4 people who paid the extreme amount of 512 survived.
On average, the survivors paid more than double for their prices but a gain, the average was influenced heavily by the outliers.



3. Conclusion

By making a simple analysis of the general information of TITANIC, we were able to describe the population of the ship in terms of several variables. To summarize, the population was relatively young, more than half of them were males who did not pay too much for their tickets. This fits with the stereotype of the poor young man, in his 20s, looking for a new life in the land of the dreams. Unfortunately, survival ratio from TITANIC was not very high, only slightly less than 40% of the passengers were able to further pursue their dreams.