STAT 151 E1

GROUP 95

LAB 1

ALVAREZ, Samuel

RAMOS, Janella

1) This is an observational study as the event had already happened and the data was collected after the event. The data cannot be generalized to a broader population since the data contains specified names, age, group size, gender, family and more. If females turned to be more fit to survive than male in this case, this study cannot not be used as proof that females in general can withstand harsh conditions than male because this is an observational study. This means that any data collected from this study is only for this study only. No additional information, guesses or inference must be made to other generalized matter that will be based on this specified observational study. Data collected in this observational study is independent from the generalized conditions that could be done with controlled experiments.

2) Overall, there are 87 cases in the file. Categorical Variables: Name, Gender (which is an identifier variable whether it is Male or Female), Family, Position, Child (yes or no), Survival (if they survived or not), Whether they are alone or not. Numerical Variables: Age, Order of Death (1st, 6th, last) and Group size which states the number of groups during the immigration.

3)

a. **Frequency table results for Survival:**
Count = 87

Survival	Frequency	Relative Frequency
Died	40	0.45977011
Survived	47	0.54022989

b. The families Donner, Graves and Murphy lost more members than the others, losing 14, 7 and 5 members. This is in comparison to the Breen Family (losing 1 member), Eddy (3 members), Keseberg (4 members), McCutchen (1 member), Reed (3 members) and Wolfinger (2 member).

Frequency table results for Survival:
Group: Family=Donner
Count = 24

Survival	Frequency	Relative Frequency
Died	14	0.58333333
Survived	10	0.41666667

Frequency table results for Survival:
Group: Family=Graves
Count = 14

Survival	Frequency	Relative Frequency
Died	7	0.5
Survived	7	0.5

Frequency table results for Survival:
Group: Family=Murphy
Count = 12

Survival	Frequency	Relative Frequency
Died	5	0.41666667
Survived	7	0.58333333

c. According to the data shown below, people who traveled alone has a lower survival rate than the people who didn't travel alone with the rate of 19%, surviving and a greater chance (81%) of not surviving.

Survival Rate of Traveling Alone:

Frequency table results for Survival:
Group: Alone=Yes
Count = 16

Survival	Frequency	Relative Frequency
Died	13	0.8125
Survived	3	0.1875

Survival Rate of People NOT Traveling Alone:

From this data, people have greater survival rate of 62% when you are not traveling alone, and a 38% rate of not surviving.

Frequency table results for Survival:
Group: Alone=No
Count = 71

Survival	Frequency	Relative Frequency
Died	27	0.38028169
Survived	44	0.61971831

D. The frequency tables below show the different survival frequencies between Male and Female genders. **(Frequency tables for each gender):**

Frequency table results for Survival:
Group: Gender=Female
Count = 34

Survival	Frequency	Relative Frequency
Died	10	0.29411765
Survived	24	0.70588235

Frequency table results for Survival:
Group: Gender=Male
Count = 53

Survival	Frequency	Relative Frequency
Died	30	0.56603774
Survived	23	0.43396226

Out of all the females, only 71% has a chance or rate of survival and 29.4% chance of not surviving during the travel. Meanwhile, out of the males, only 43% has a chance or rate of survival and a greater rate of 57% or not surviving. **(explaining the tables above)**

Explaining the data below: The Overall Survival Rate out of both genders, categorized as Male and Female, (as shown in the table below) Females has a higher frequency or rate of survival than men with a relative frequency of 51.1% while the men has 48.9% survival rate.

Also from the data below, alone, it is seen that more Female survived than Male during the immigration/travel. Specifically from this data, it is shown that 24 female and 23 male withstand harsh conditions during that travel out of 87 people. Also out of 40 people who died, 10 or (25%) were female and 30 or (75%) were male.

Frequency table results for Gender:
Where: Survival = Survived
Count = 47

Gender	Frequency	Relative Frequency
Female	24	0.5106383
Male	23	0.4893617

Frequency table results for Gender:
Group: Survival=Died
Count = 40

Gender	Frequency	Relative Frequency
Female	10	0.25
Male	30	0.75

3) e. **Comments about Survival of Child and Adults:** According to the data below, out of all the people who died, 63% or 25 of them are adults and only 38% or 15 were children. Meanwhile out of the people who survived the journey (47 people); 34% or 16 of them are adults and 66% or 31 of the 47 people who survived are children.

*(Children and Adult Comment):
The data of this observational study states that (out of 46 children) 31/46 children survived and (out of 41 adults) 16/41 adults survived the harsh conditions of the journey. While 15/26 children and 25/41 adults did not survive the travel.

Frequency table results for Child:
Group: Survival=Died
Count = 40

Child	Frequency	Relative Frequency
No	25	0.625
Yes	15	0.375

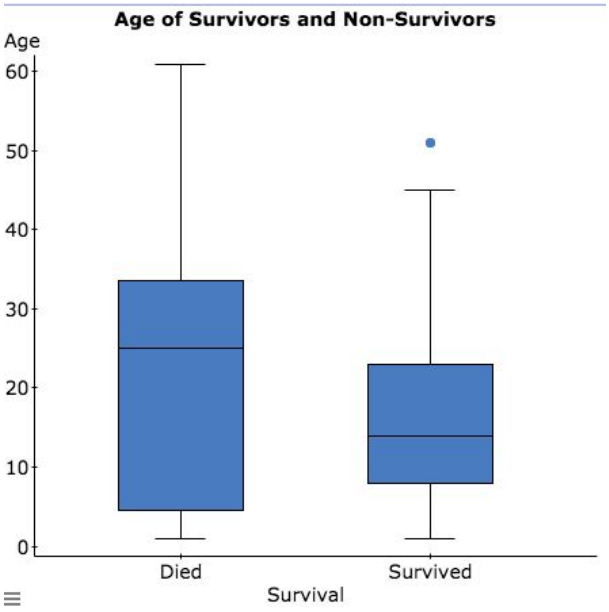
Frequency table results for Child:
Group: Survival=Survived
Count = 47

Child	Frequency	Relative Frequency
No	16	0.34042553
Yes	31	0.65957447

f. **Summary Stats for Age:**

Summary statistics for Age:
Group by: Survival

Survival	n	Mean	Std. dev.	Median	IQR
Died	40	23.6	18.005412	25	29
Survived	47	16.319149	11.62289	14	15



Comment: Based on the data above: For the frequency of age for non-survivors, the boxplot for that is a right skewed spread of distribution, although this should mean that the mean should usually be greater than the median, but for this case the mean is less than the median. This is because less old people (around 45-60 yrs old) died during the journey compared to the number of young people who died. The best measure of the spread of this data is the Q1 and Q3 and the IQR. The median for this non-survivor data is around the 25 years of age, which quite young for one to die due to harsh conditions.

Meanwhile for the survivors data, this also shows a right skewed distribution where the mean is greater than the median. The right skewed is also due to the uneven distribution regarding the age of who survived. While for this data, the median or the centre for this data is 14 years of age. According to the data, around 16 years old is the average age (mean) for survival, meaning more of the people who have ages around 16 years old survived and quite a few old people (age around 40-60) survived.

3) g. **Summary Stats for Age (Female)**

Summary statistics for Age:
Where: Gender = "Female"
Group by: Survival

Survival	n	Mean	Std. dev.	Median	IQR
Died	10	21.3	19.793658	18.5	43
Survived	24	15.208333	10.064916	13.5	13

Summary Stats for Age (Male)

Summary statistics for Age:
Where: Gender = "Male"
Group by: Survival

Survival	n	Mean	Std. dev.	Median	IQR
Died	30	24.366667	17.662513	25	25
Survived	23	17.478261	13.183563	14	20

According to the data above, the difference in average age between the male and female who survived is 17.478261 - 15.208333 = 2.269928. Meanwhile, the average age between the male and female who did not survive is 24.366667 - 21.3 = 3.066667.

Also it is understood from the data above, that more female survived with a frequency of 71% (out of all the female only) and only 43% (out of all the males only) survived.

4)

a) From the data below, the survival rates between males and females differ. When referencing the total number of females that did survive out of ALL of the females, we get a number of 70.58% or 24 out of 34 females survive. In terms of males, we get a percentage of 43.40% or 23 out of 53 males which shows that a higher percentage of females survived compared to males. When referencing the entire sample population, for females, we have a survival rate of 27.59% or 24 out of 87, for males, we get a value of 26.44% or 23 out of 87. Also, 24 out of the 47 survivors were females (51.06%). This means that regardless of which point of reference we use, male survival rates appear to be lower than females, therefore the survival rates differ in the fact that more females survive than males.

Contingency table results:
Rows: Survival
Columns: Gender

Cell format
Count (Row percent) (Column percent) (Percent of total)

	Female	Male	Total
Died	10 (25%) (29.41%) (11.49%)	30 (75%) (56.6%) (34.48%)	40 (100%) (45.98%) (45.98%)
Survived	24 (51.06%) (70.59%) (27.59%)	23 (48.94%) (43.4%) (26.44%)	47 (100%) (54.02%) (54.02%)
Total	34 (39.08%) (100%) (39.08%)	53 (60.92%) (100%) (60.92%)	87 (100%) (100%) (100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	6.1659327	0.013

b)The Chi-square for independence tests for the relationships between the two categorical values, in this case we are examining whether these two categorical values are independent of each other or not. The null hypothesis in this case is that gender and survival rates are independent of each other, this means that gender does not whatsoever influence a person's chance of survival. The alternative hypothesis however states that these two categorical values are not independent of each other, and therefore, gender may influence the outcome for survival indefinitely. Because the theoretical p-value (0.013) is much lower than the significance level (0.05) there is evidence that goes against the null hypothesis. Therefore, the concluding idea is that these two variables are not independent of each other, we can reject the null hypothesis.

Contingency table results:
Rows: Gender
Columns: Survival

	Died	Survived	Total
Female	10	24	34
Male	30	23	53
Total	40	47	87

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	6.1659327	0.013

c) Out of all the survivors, 24 of them or 51.1% were females, the rest were males. And, out of the entire sample, 24 out of the 87 which were females survived, this number accounts for approximately 0.2758 or 27.6% of the sample.

d)The null hypothesis in this case is that there is no difference between the survival rates for males and females. The alternative hypothesis in the other hand states that there *is* a difference between the survival rates for males and for females. In this case, since the P-value (0.013) is less than the significance value (0.05), we have enough evidence to point against the null hypothesis, and thus reject it. The concluding idea for this test then is that there is a difference between the survival rates for males and females.
P-value: 0.013
Test Statistic: 2.483
Distribution: Normally Distributed

Difference	Female Survivor	Female Total	Male Survivor	Male Total	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	24	34	23	53	0.27192009	0.10950701	2.4831296	0.013

e) The tests that were carried out in part a and part b differs in the sense that, one tested for the difference in the survival rates between the two sub-categorical values (Male or Female). While the other tested for whether there was a relationship (In terms of dependence) between two main categorical values; Gender and survival.

f)The table below confirms the findings from part d as the lower limit is greater than and NOT equal to zero.

Two sample proportion confidence interval:
p_1 : Proportion of successes (Success = Survived) for Survival where Gender = "Female"
p_2 : Proportion of successes (Success = Survived) for Survival where Gender = "Male"
$p_1 - p_2$: Difference in proportions

95% confidence interval results:

Difference	Female Survivor	Female Total	Male Survivor	Male Total	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	24	34	23	53	0.27192009	0.10363854	0.068792281	0.4750479

5.
a) According to the chart below, the two age groups that were favoured for survival are ages 7-13 and 13-19, with both having a survival percentage of 12.6%, or each make up 11 out of the 47 survivors (23.40%), with a combined survival rate of 46.81% (22/47). of The two least favoured age groups were ages 1-7 and 25-31 with each having a survival rate of 5.75%.

	1 to 7	7 to 13	13 to 19	19 to 25	25 to 31	31 to 37	37 to 43	43 to 49	49 to 55	55 to 61	61 to 67	Total
Died	12 (30%) (57.14%) (13.79%)	2 (5%) (15.38%) (2.3%)	1 (2.5%) (8.33%) (1.15%)	2 (5%) (28.57%) (2.3%)	12 (30%) (70.59%) (13.79%)	4 (10%) (57.14%) (4.6%)	0 (0%) (0%) (0%)	3 (7.5%) (75%) (3.45%)	0 (0%) (0%) (0%)	3 (7.5%) (100%) (3.45%)	1 (2.5%) (100%) (1.15%)	40 (100%) (45.98%) (45.98%)
Survived	9 (19.15%) (42.86%) (10.34%)	11 (23.4%) (84.62%) (12.64%)	11 (23.4%) (91.67%) (12.64%)	5 (10.64%) (71.43%) (5.75%)	5 (10.64%) (29.41%) (5.75%)	3 (6.38%) (42.86%) (3.45%)	1 (2.13%) (100%) (1.15%)	1 (2.13%) (25%) (1.15%)	1 (2.13%) (100%) (1.15%)	0 (0%) (0%) (0%)	0 (0%) (0%) (0%)	47 (100%) (54.02%) (54.02%)
Total	21 (24.14%) (100%) (24.14%)	13 (14.94%) (100%) (14.94%)	12 (13.79%) (100%) (13.79%)	7 (8.05%) (100%) (8.05%)	17 (19.54%) (100%) (19.54%)	7 (8.05%) (100%) (8.05%)	1 (1.15%) (100%) (1.15%)	4 (4.6%) (100%) (4.6%)	1 (1.15%) (100%) (1.15%)	3 (3.45%) (100%) (3.45%)	1 (1.15%) (100%) (1.15%)	87 (100%) (100%) (100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	10	25.908103	0.0039

b)The null hypothesis (Distribution of 10), which states that Age and survival rates are independent of each other, and the alternative hypothesis which states that these two categorical values are not independent of each other are tested here. Because the significance value of 0.01 is greater than the 0.0039 of the P-value, there is strong evidence that goes against the null

hypothesis, which means that the variables age(bin) and survival rates do have an affect on each other. The Distribution is 10, which comes from the degrees of freedom.

Chi-Square test:

Statistic	**DF**	**Value**	**P-value**
Chi-square	10	25.908103	0.0039

6.

a) From the chart below, group 7 and 9 had the highest survival rates. Group 7 with 6 individuals whom all 6 survived. Group 9 which had 9 individuals whom all 9 survived. The group which has the lowest survival rate is group 1, had a total of 16 members whom in which 13 died and only 3 survived giving the group a survival percentage of 18.75%. Group 1, which had the lowest survival rate was made up entirely of males. The survival rate does not increase when the group size increase, the groups with the highest number of individuals also suffered the highest mortality rate.

	1	2	3	4	7	9	12	13	16	Total
Died	13 (32.5 %) (81.25 %) (14.94 %)	2 (5%) (50%) (2.3 %)	1 (2.5%) (33.33 %) (1.15 %)	5 (12.5 %) (62.5 %) (5.75 %)	0 (0%) (0%) (0%)	0 (0%) (0%) (0%)	5 (12.5 %) (41.67 %) (5.75 %)	6 (15%) (46.15%) (6.9%)	8 (20%) (50%) (9.2%)	40 (100%) (45.98%) (45.98%)
Survived	3 (6.38 %) (18.75 %) (3.45 %)	2 (4.26 %) (50%) (2.3 %)	2 (4.26 %) (66.67 %) (2.3%)	3 (6.38 %) (37.5 %) (3.45 %)	6 (12.77 %) (100%) (6.9%)	9 (19.15 %) (100%) (10.34 %)	7 (14.89 %) (58.33 %) (8.05 %)	7 (14.89%) (53.85%) (8.05%)	8 (17.02 %) (50%) (9.2%)	47 (100%) (54.02%) (54.02%)

Total	16 (18.39 %) (100%) (18.39 %)	4 (4.6 %) (100 %) (4.6 %)	3 (3.45 %) (100%) (3.45 %)	8 (9.2 %) (100 %) (9.2 %)	6 (6.9%) (100%) (6.9%)	9 (10.34 %) (100%) (10.34 %)	12 (13.79 %) (100%) (13.79 %)	13 (14.94%) (100%) (14.94%)	16 (18.39 %) (100%) (18.39 %)	87 (100%) (100%) (100%)

Chi-Square test:

Statistic	**DF**	**Value**	**P-value**
Chi-square	8	22.073269	0.0048

b) The significance value of 0.01 is much greater than the theoretical P-value of 0.0048 which means that we have strong enough evidence to reject the null hypothesis. This means that there happens to be a relationship between the variables group size and survival rates, which is that the higher the number of individuals in a group, the less the survival rate is for that group as, as well, the lower the number of individuals in a group, the higher the survival rate is.

Null hypothesis: Group size and survival rates do not have a relationship, and thus is independent of each other. Distribution = 8.

Alternative hypothesis: Group size and survival rates do have a relationship, and thus influences each other's outcome.

7)In this study, a few factors determine whether an individual was favoured for survival or not. To start off, we must narrow down the variables which were not independent of survival; whether an individual travelled alone, gender, youth or adult, age, and group size. In this study, it was found that those who travelled alone suffered a higher mortality rate (lower survival) compared to those that travelled in groups. Particularly, those who travelled alone made up only 6.38% or 3 out of the 47 that survived while those that did not travel alone made up a total of 44 out of 47 or 93.62%, therefore, those who travelled with a group were favoured to survive.

Gender in this case also made a difference. It was found that females made up a total of 51.06% of those that survived while the males made up 48.94% of those that survived, hence, being a female slightly increases an individual's rate of survival.

Being a youth also increased the survival rate of an individual, from the calculations in question 3e and the chart in 5a, it was found that 66% of the those that survived were made up of individuals ages 1-19, clearly enough, this means that survival was favoured to those that were considered as “children”. Tying in with the idea of children being favoured, age logically would also play a major role. From the chart found in 5a, the age group(s) that had the highest rate of survival were age groups 7-13 and 13-19 with each making up 23.40% of the 47 individuals that survived, and combined making up a total of 46.80% of the 47.

Family name also fits into the picture, when relying on the data alone, family name does not really boast much of importance compared to the other factors in terms of determining survival but is still worthy of being mentioned. With the given data, it was found that the Breen family which had lost only 1 family member had the highest survival rate among all the families (90% survival). Breaking down the reason why the Breen family was so successful in surviving, we look at the data. 7 out of the 10 were children, which from data shows that higher survival rates were higher for children than adults. 9 out of 10 travelled in a group and not alone, which also agrees with the concluding idea that higher survival rates favoured those that travelled in groups. 2 out of the 10 were female, from the data in this study it was showed that being a female slightly increased the rate of survival. Lastly, the family travelled as a relatively small group (9 members), which agrees with the conclusion that smaller groups had higher rates of survival compared to larger groups like 16.

Last but not the least, group sizes also played a major role in determining the survival rates. From this study, it was found that survival favoured those that had smaller groups. From the chart in 6a, groups 1 and 16 which had the largest group (each having 16) both suffered a large mortality rate. Group 1 lost 13 out of the 16 members, which accounts for 81.25% of the group of 16, while group 2 lost 8 of their 16 membership, which accounts for 50% of their group.

In conclusion, the most important factors within this study that determines an individual’s survival are; whether they travel alone or not, their gender, their age, the group size that they travel with.