STATS 151 A1

GROUP #10

LAB 3

HUANG, Ye

SATO, Chris

Huang Ye 1530842
Sato Chris 1530225

1. Is it an observational study or a randomized experiment? Can the data be generalized to a broader population? If females in the study turned out to be more apt to survive than males, could this be used as proof that, in general, females are better able than males to withstand harsh conditions?

This is an observational study because when collecting the data, the researchers had no influence on any variables or outcomes. The data could not be applied to a broader population.
If the females had a higher survival rate, it would not be usable as proof for females surviving harsher conditions because the two events are independent.

2. Now discuss the data in the file. How many cases are there? Identify categorical and numerical variables in the data, noting which one is the identifier variable.

There are 87 cases. Name, gender, family, position, child, survival, and alone are all categorical variables. Age, order and group size are all numerical variables.

3. Now you will use frequency tables to summarize the family and gender composition of the group and obtain the proportions of group members who survived for each family and gender. Provide all values for parts (a) – (e) in percentage.
(a) What was the overall survival rate? Obtain the appropriate frequency table to answer the question and paste the table into your report. (Include frequency and relative frequency.)

Survival	Frequency	Relative Frequency
Died	40	0.45977011
Survived	47	0.54022989

The overall survival rate is 54.02 percent.
(b) Which three families proportionally lost more members than others? Obtain the appropriate frequency table to answer the question. Do not paste the output into your report.

The Eddies, the Kesebergs, and the Wolfinger proportionally lost more members.
Eddies mortality rate: 75%
Kesebergs mortality rate: 66.6%
Wolfingers mortality rate: 66.6%
(c) What was the survival rate of people travelling alone? What was the survival rate of people who were not? Obtain the appropriate frequency table to answer the question and paste the table into your report. (Include frequency and relative frequency.)
Traveled alone

Survival	Frequency	Relative Frequency
Died	13	0.8125
Survived	3	0.1875

In group

Survival	Frequency	Relative Frequency
Died	27	0.38028169
Survived	44	0.61971831

According to the relative frequency table, the survival rate of going alone is severely lower than the survival rate of going in groups.

(d) What percentage of each gender group survived the ordeal? Obtain the appropriate frequency table to answer the question and paste the table into your report. (Include frequency and relative frequency.) Moreover, obtain the corresponding relative frequency bar charts of survival by gender (separate graph for each gender). Paste the two charts into your report. Comment briefly.

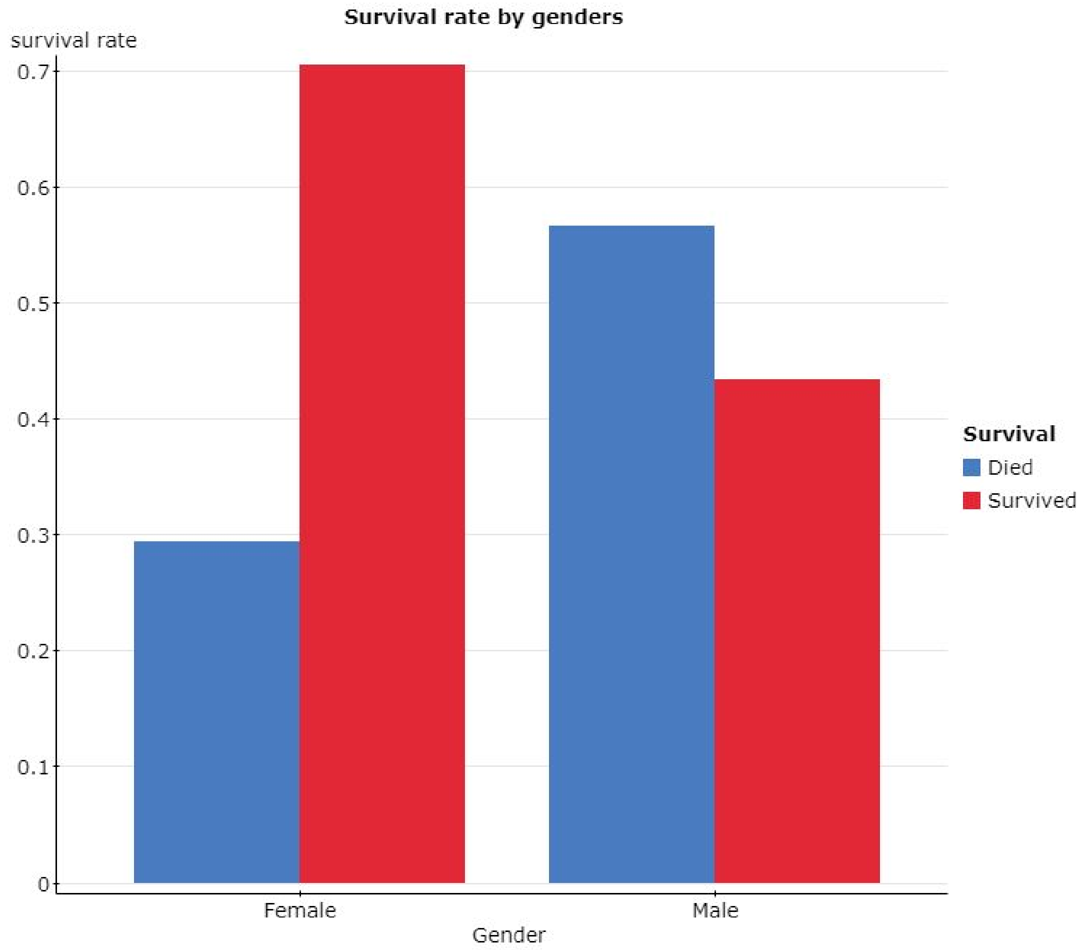
Female participants:

Survival	Frequency	Relative Frequency
Died	10	0.29411765
Survived	24	0.70588235

Male participants:

Survival	Frequency	Relative Frequency
Died	30	0.56603774
Survived	23	0.43396226

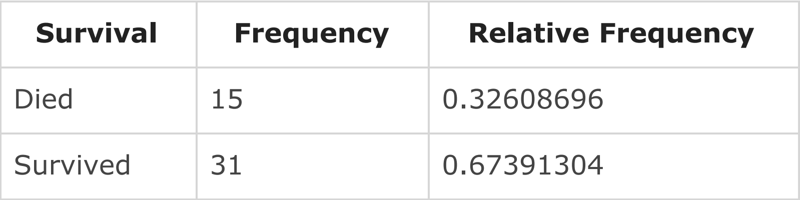
Huang Ye 1530842
Sato Chris 1530225



There is a noticeable difference in survival rate between the two genders, females have a significantly larger survival rate than males.

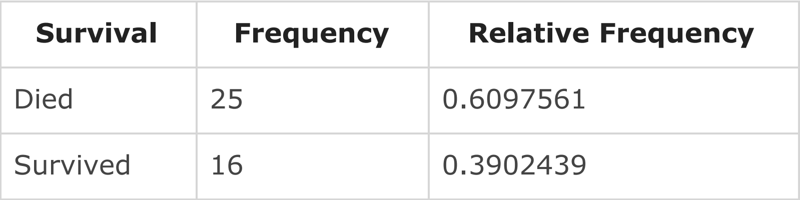
(e) What was the survival rate of children? The survival rate of adults? Obtain the appropriate frequency table to answer the question and paste the table into your report. (Include frequency and relative frequency.) Comment briefly.

Children:



Survival	Frequency	Relative Frequency
Died	15	0.32608696
Survived	31	0.67391304

Adults:



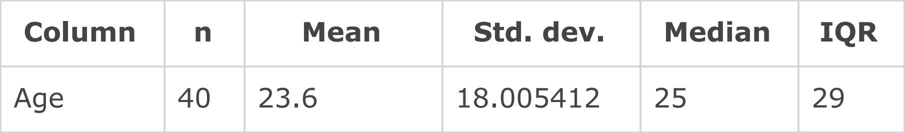
Survival	Frequency	Relative Frequency
Died	25	0.6097561
Survived	16	0.3902439

The survival rate of Children is higher compares the the survival rate of adults.

Huang Ye 1530842
Sato Chris 1530225

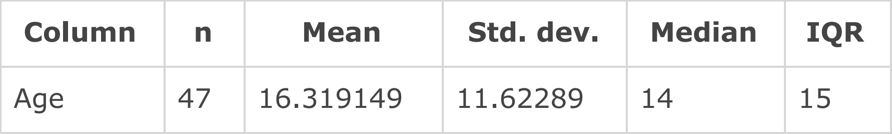
(f) Obtain the summary statistics of age (sample size, mean, median, standard deviation, and the interquartile range) for survivors and non-survivors. Paste the output into your report. Moreover, obtain the side-by-side boxplots of age for survivors and non-survivors. Use fences to identify outliers. Paste the boxplots into your report. Comment about the centers, spreads, and shapes of the two distributions.

Died

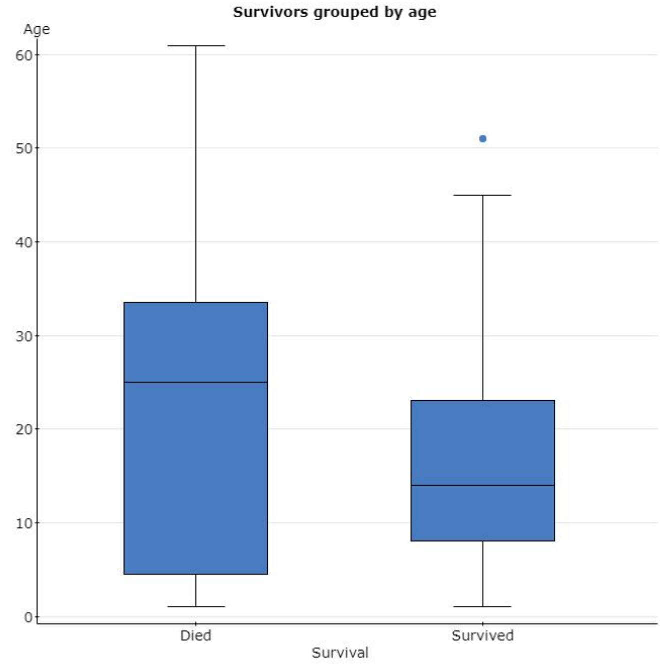


Column	n	Mean	Std. dev.	Median	IQR
Age	40	23.6	18.005412	25	29

Survived



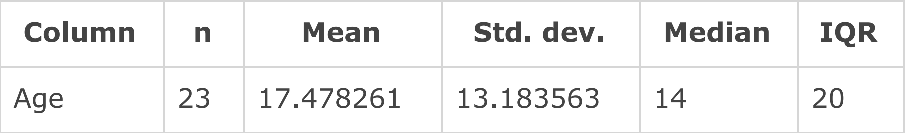
Column	n	Mean	Std. dev.	Median	IQR
Age	47	16.319149	11.62289	14	15



The non-survivor group has a larger median than the survivor group and is left skewed, and the standard deviation of the non-survivor group is larger than the survivor group.

(g) Obtain the summary statistics of age (sample size, mean, median, standard deviation, and the interquartile range) for survivors and non-survivors for each gender. Paste the output into your report. What was the difference in average age of those who survived and not survived for each gender?

Male survivor:



Column	n	Mean	Std. dev.	Median	IQR
Age	23	17.478261	13.183563	14	20

Male non-survivor:

Huang Ye 1530842
Sato Chris 1530225

Column	n	Mean	Std. dev.	Median	IQR
Age	30	24.366667	17.662513	25	25

Female survivor:

Column	n	Mean	Std. dev.	Median	IQR
Age	24	15.208333	10.064916	13.5	13

Female non-survivor:

Column	n	Mean	Std. dev.	Median	IQR
Age	10	21.3	19.793658	18.5	43

The difference in average is 18.7280855-16.343297 = 2.3847885.

4. In this question, you will examine the relationship between survival and gender.

Contingency table results:

Rows: Gender
Columns: Survival

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

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

Huang Ye 1530842
Sato Chris 1530225

Chi-Square test:

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

(a) Were the chances of survival different for females than for males? In order to answer the question, obtain the contingency table of survival by gender. Make sure that Row percent, Column percent, and Percent of Total as well as Chi-Square test for independence are selected. Paste the table into your report.

The survival rate was different between females and males. Females had a 70.59% chance of survival, where Males only had a 43.40% chance of survival.

(b) Using α = 0.05, test that there was no relationship between survival and gender. State the null and alternative hypotheses. Report the value of the appropriate test statistic, the distribution of the test statistic under the null hypothesis, and the P-value of the test to answer the question. State your conclusion.

H_0: There is no relationship between survival and gender.

H_a: There is not no relationship between survival and gender.

SLA: α=0.05 > P-Value=0.013, this gives us strong evidence to reject H_0.

JA: 0.01 < P-Value=0.013 < 0.05, this gives us moderate to strong evidence to reject H_0.

(c) Refer to the output in part (a) to answer the following questions: What percent of the survivors were females? What percent were female survivors?

Of the survivors, 51.06% were females. Female survivors made up 27.59% of the people.

(d) Using α = 0.05, is there evidence that there was a difference in the survival rate for females and males? Carry out the appropriate two-sample proportion test. State the null and alternative hypotheses. Report the value of the appropriate test statistic, the distribution of the test statistic under the null hypothesis, and the P-value of the test to answer the question. State your conclusion.

Two sample proportion hypothesis test:

p_1 : Proportion of successes (Success = Survived) for Survival where Gender="Male"
p_2 : Proportion of successes (Success = Survived) for Survival where Gender="Female"
$p_1 - p_2$: Difference in proportions
$H_0 : p_1 - p_2 = 0$
$H_A : p_1 - p_2 \neq 0$

Hypothesis test results:

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	23	53	24	34	-0.27192009	0.10950701	-2.4831296	0.013

Huang Ye 1530842
Sato Chris 1530225

H_0: There is a difference in survival between females and males.
H_a: There is not a difference in survival between females and males
SLA: α=0.05 > P-Value=0.013
JA: 0.01 < P-value=0.013 < 0.05, Moderate to strong evidence to reject H_0.
Because of the evidence provided by the SLA and JA, we reject H_0.
(e) What is the relationship between the tests in parts (a) and (b)?
The test in (a) simply states the percentage of people that survived of each gender, while (b) looks for a relationship between survival and gender.

(f) Obtain and interpret a 95% confidence interval for the difference in survival rates of females and males. Paste the output into your report. What do you conclude? Does it confirm your result in part (d)?

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

5. In this question, you will explore the relationship between age and survival. First, divide Age into several nonoverlapping intervals so that the age of each member falls into exactly one of those age categories. In order to do it, obtain a bin column, Bin(Age), for the Age variable with the bins starting at 1 and a binwidth of 6 (see Introductory Lab Lab Instructions, pages 15-16). Make sure that the left endpoint of each class interval is included (and that the right endpoint is excluded).
(a) Obtain a contingency table to study the relationship between survival and Bin(Age). Make sure that Row percent, Column percent, and Percent of Total as well as Chi-Square test for independence are selected. Paste the table into your report. What age intervals represent the two highest and two lowest survival rates? (Ignore age intervals with less than five total members.)

Huang Ye 1530842
Sato Chris 1530225

Contingency table results:
Rows: Bin(Age)
Columns: Survival

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

	Died	Survived	Total
1 to 7	12 (57.14%) (30%) (13.79%)	9 (42.86%) (19.15%) (10.34%)	21 (100%) (24.14%) (24.14%)
7 to 13	2 (15.38%) (5%) (2.3%)	11 (84.62%) (23.4%) (12.64%)	13 (100%) (14.94%) (14.94%)
13 to 19	1 (8.33%) (2.5%) (1.15%)	11 (91.67%) (23.4%) (12.64%)	12 (100%) (13.79%) (13.79%)
19 to 25	2 (28.57%) (5%) (2.3%)	5 (71.43%) (10.64%) (5.75%)	7 (100%) (8.05%) (8.05%)
25 to 31	12 (70.59%) (30%) (13.79%)	5 (29.41%) (10.64%) (5.75%)	17 (100%) (19.54%) (19.54%)
31 to 37	4 (57.14%) (10%) (4.6%)	3 (42.86%) (6.38%) (3.45%)	7 (100%) (8.05%) (8.05%)

37 to 43	0 (0%) (0%) (0%)	1 (100%) (2.13%) (1.15%)	1 (100%) (1.15%) (1.15%)
43 to 49	3 (75%) (7.5%) (3.45%)	1 (25%) (2.13%) (1.15%)	4 (100%) (4.6%) (4.6%)
49 to 55	0 (0%) (0%) (0%)	1 (100%) (2.13%) (1.15%)	1 (100%) (1.15%) (1.15%)
55 to 61	3 (100%) (7.5%) (3.45%)	0 (0%) (0%) (0%)	3 (100%) (3.45%) (3.45%)
61 to 67	1 (100%) (2.5%) (1.15%)	0 (0%) (0%) (0%)	1 (100%) (1.15%) (1.15%)
Total	40 (45.98%) (100%) (45.98%)	47 (54.02%) (100%) (54.02%)	87 (100%) (100%) (100%)

Chi-Square test:

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

Warning: over 20% of cells have an expected count less than 5.
Chi-Square suspect.

Ignoring intervals with less than 5 members, groups from ages 13 to 19, and 7 to 13 have the highest survival rates at 91.67% and 84.62% respectively.
The ranges from 25 to 31, and 31 to 37 had the lowest survival rates with 29.41% and 42.86%.

(b) Using α = 0.01, test that there was no relationship between survival and age category. Refer to the output in part (a). State the null and alternative hypotheses. Report the value of the appropriate test statistic, the distribution of the test statistic under the null hypothesis, and the P-value of the test to answer the question. State your conclusion.

H_0 : Survival and Age category are independent.
H_a : Survival and Age category are not independent.
SLA: α=0.01 > P-value=0.0039, strong evidence to reject H_0.
JA: 0 < P-Value=0.0039< 0.01, strong evidence to reject H_0.
Since SLA and JA both provide strong evidence, we reject H_0.

6. In this question, you will examine the relationship between survival and group size.
(a) Obtain the contingency table of survival by group size. Make sure that Row percent, Column percent, and Percent of Total as well as Chi-Square test for independence are selected. Paste the table into your report. Comment briefly on lowest/highest survival rates. (Ignore group sizes with less than five total members.) The group size with the lowest survival rate consists of what gender? Does survival rate increase with group size?

Rows: Survival
Columns: Group Size

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

	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

According to the table, group size of 7 and 9 have the highest survival rates with all group members survived, while group size of 1 has the lowest survival rate with only 3 out of the 16 groups survived.

All groups of group size 1 are males.

The survivability does increase with group size with the exception from group size 9 to group size of 12.

(b) Using α = 0.01, test that there was no relationship between survival and group size. State the null and alternative hypotheses. Report the value of the appropriate test statistic, the distribution of the test statistic under the null hypothesis, and the P-value of the test to answer the question. State your conclusion.

H-null: there was no relationship between survival and group size.

H-alternative:there was relationship between survival and group size

SLA: α > p-value, strong evidence, reject H-null.

JA: p-value < 0.01, strong evidence, reject H-null.

Therefore, because both SLA and JA proves strong evidence, I reject H-null.

7. Briefly summarize the study. In particular, answer the following question: Which factors were the most important predictors of survival? Refer to the statistics obtained in Questions 1-6.

Genders, age and group size play significant roles in the event. By being a female instead of a male, the survival rate increases from 40% to 70%. And by being inside a large group, the survival rate increases gradually as the group size increases till. And finally, by being a child (age under 19), the survival rate increases to 67% instead of being an adult which only has 60%. From the data, because it is an observational study, therefore we cannot make any causal inference.