LAB 3 ASSIGNMENT SOLUTIONS

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?

   There are no treatments imposed on the subjects in the study. Their fate in the ordeal (survival or non-survival), age, position, and all other variables are recorded for each subject. Therefore, the study is an example of an observational study.

   Since the 87 individuals were not randomly sampled from any well-defined population, inference to a broader population is not justified. Moreover, as the study is observational, there is no possibility of random assignment and no causal inferences about survival can be made based on the data. In other words, the results cannot be used as evidence that females were more likely to survive than males; some confounding variables could be responsible for difference if any.

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 in the data file. There are seven categorical variables: *Name*, *Gender*, *Family*, *Position*, *Child*, *Survival*, and *Alone*. There are two numerical variables: *Age*, and *Group Size*. *Order* is an ordinal variable. The identifier variable is *Name*.

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.)

   Frequency table results for Survival:
   Count = 87

Survival	Frequency	Relative Frequency
Died	40	0.45977011
Survived	47	0.54022989

   The overall survival rate was 54.023% (or 47 out of 87).

   (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.

   In order to answer the question, obtain a frequency table of *Survival*, grouping by *Family*. The families that proportionally lost most of their members are the Eddy family (75% died, 3 out of 4), the Keseberg family (66.7% died, 4 out of 6), and the Wolfinger family (66.7% died, 2 out of 3).

   (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.)

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

Survival	Frequency	Relative Frequency
Died	13	0.8125
Survived	3	0.1875

Frequency table results forSurvival:
Group: Alone=No
Count = 71

Survival	Frequency	Relative Frequency
Died	27	0.38028169
Survived	44	0.61971831

There is a large difference in the survival rates between people travelling alone (Teamsters and Passengers) and people who were not. The survival rate for people travelling alone is 18.75% versus 61.972% travelling not alone.

(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.

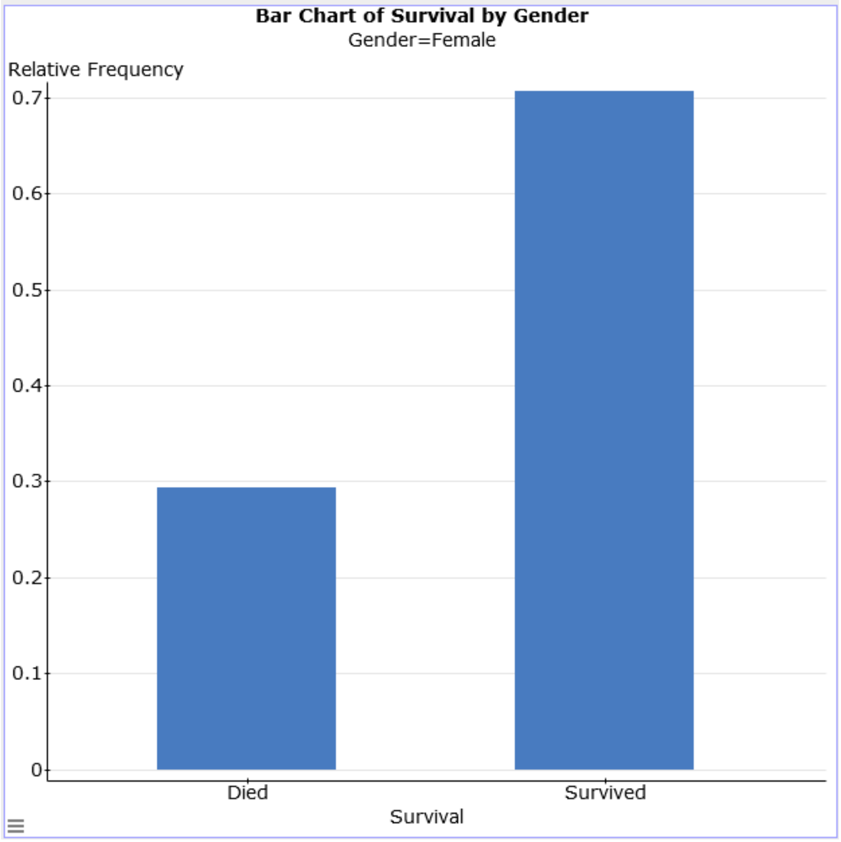
Frequency table results for Survival:
Group: Gender=Female
Count = 34

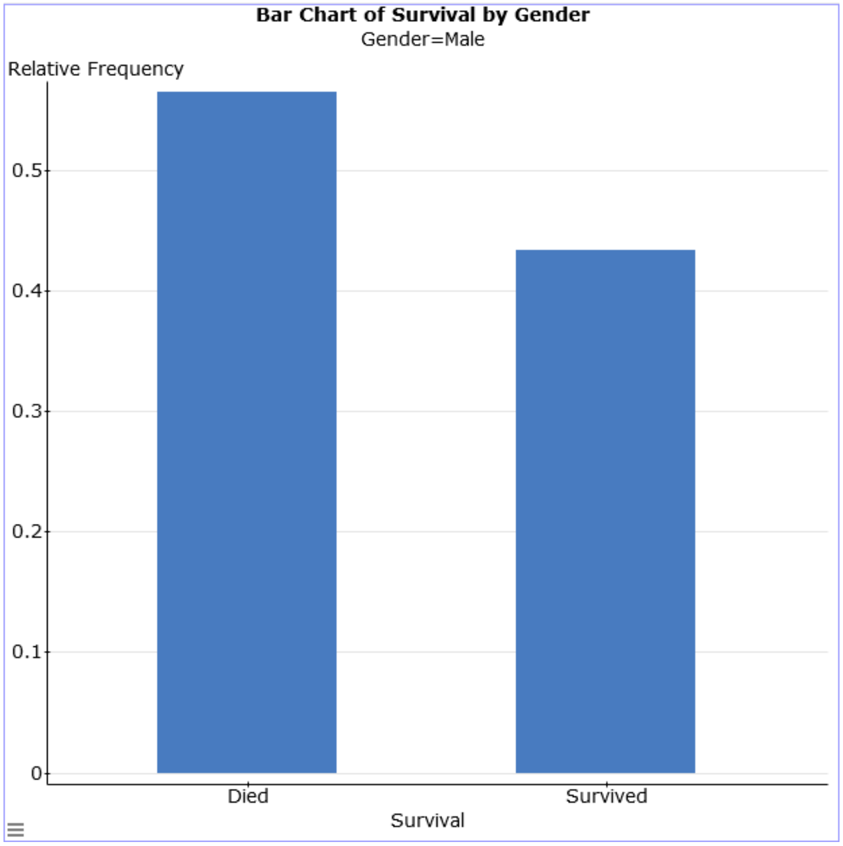
Survival	Frequency	Relative Frequency
Died	10	0.29411765
Survived	24	0.70588235

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

Survival	Frequency	Relative Frequency
Died	30	0.56603774
Survived	23	0.43396226

According to the output, 70.588% of females survived versus 43.396% of males. The survival rate for females was substantially higher than the survival rate for males. This is confirmed by the bar charts, where the survival and non-survival rates for males are much closer together than they are for females. It would appear that being female made it more likely to survive the ordeal. Under conditions marked by famine and extreme cold in the study, females fared better than their male counterparts. As the data are observational, the differences in the mortality rates between the two genders cannot be attributed to gender alone.


Bar Chart of Survival by Gender
Gender=Female
Relative Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Died
Survived
Survival


Bar Chart of Survival by Gender
Gender=Male
Relative Frequency
0.5
0.4
0.3
0.2
0.1
0
Died
Survived
Survival

(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.

Frequency table results for Survived:
Group: Child=Yes
Count = 46

Outcome	Frequency	Relative Frequency
Died	15	0.32608696
Survived	31	0.67391304

Frequency table results for Outcome:
Group: Child=No
Count = 41

Outcome	Frequency	Relative Frequency
Died	25	0.6097561
Survived	16	0.3902439

The survival rate for children was 67.391% versus 39.024% for adults. Children were much more likely to survive.

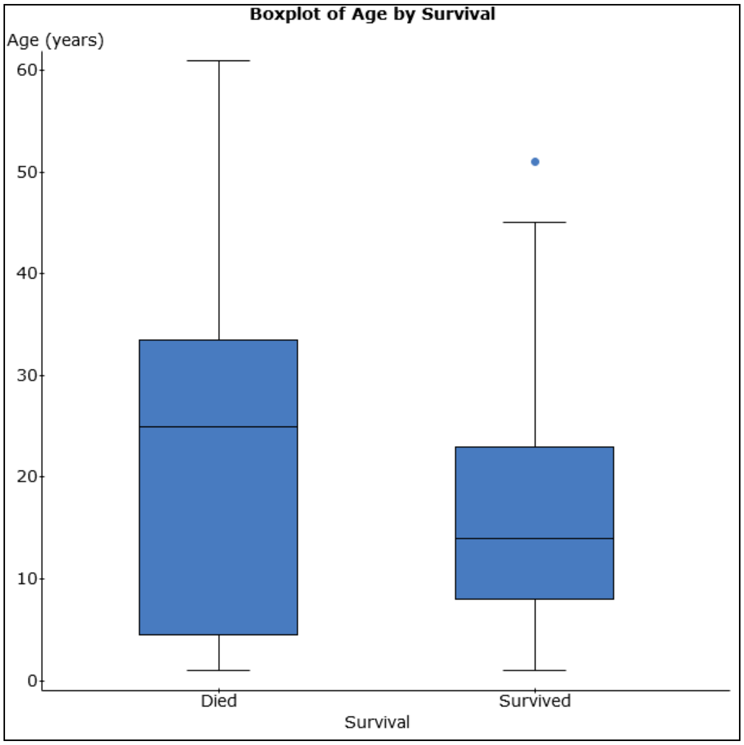
(f) Obtain the summary statistics of age (sample size, mean, median, standard deviation, and the interquartile range) for survivors and non-survivors. Paste the output into your report. Moreover, obtain the side-by-side boxplots of age for survivors and non-survivors. Use fences to identify outliers.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

Summary statistics for Age:
Group by: Survived

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

According to the summary statistics and the boxplots, those who did not survive tended to be older than those who survived. Survivors were, on average, 7.281 years younger than non-survivors. (This is consistent with the survival rate of children being higher.) The difference in the medians is even more extreme: The median of those who died was 25 years whereas the median of those who survived was 14 years. The spread of ages measured by standard deviation and IQR was substantially larger for those who died.

It is not possible to determine the shape of a distribution with certainty from its five summaries provided by the corresponding boxplot. Nevertheless, the positions of the quartiles and the lengths of whiskers in each boxplot are consistent with right skewness of each distribution.



(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?

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

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

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

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

The males who survived were an average of 6.888 years younger than those who died, while surviving females averaged 6.092 years younger than those who died.

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

(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.

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) Using $\alpha = 0.05$, 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.

Define the null and alternative hypotheses as follows:

H_0: The survival and gender variables are independent.
H_A: The survival and gender variables are dependent.

According to the output, the value of the test statistic (to three decimal places) is 6.166 and the statistic follows a chi-squared distribution with 1 degree of freedom. According to the output, the *P*-value of the test is 0.013. Since the *P*-value is smaller than $\alpha = 0.05$, we reject H_0 and conclude that there is evidence of a relationship between survival and gender.

(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?

There were 24 females among the 47 survivors, therefore 24/47 = 51.064% of the survivors were females. There were 24 surviving females among the 87 total members, thus 24/87 = 27.586% were female survivors.

(d) Using $\alpha = 0.05$, 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=Female
p_2 : Proportion of successes (Success = Survived) for Survival where Gender=Male
$p_1 - p_2$: 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
$p_1 - p_2$	24	34	23	53	0.27192009	0.10950701	2.4831296	0.013

According to the output, the value of the test statistic (to three decimal places) is 2.483 and the statistic follows a standard normal distribution (or *z*-distribution). According to the output, the *P*-value of the test is 0.013. Since the *P*-value is smaller than $\alpha = 0.05$, we reject H_0 and conclude that there is a difference in the survival rate for females and males.

(e) What is the relationship between the tests in parts (b) and (d)?

The two tests are equivalent. More precisely, the *z*-test about the difference in two proportions is equivalent to the chi-squared test for independence when each variable has only two categories. The two tests produce the same *P*-value and there is the following mathematical relationship between the test statistics: $\chi^2 = (z)^2$. Indeed, $6.1659327 \approx 2.4831296^2$.

(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)?

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	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	24	34	23	53	0.27192009	0.10363854	0.068792281	0.4750479

With 95% confidence, the difference in survival rates of females and males, to four decimal places, is between 0.0688 and 0.4750. The interval does not contain zero, so there is a difference between the proportions. Additionally, since both endpoints are positive, the survival rate among females is higher than among males. This is consistent with the outcome of the test in part (d).

5. In this question, you will explore the relationship between age and survival. First, divide *Age* into several non-overlapping 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.)

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

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

	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

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

The individuals in the 13-19 age interval are the most likely to have survived with a survival rate of 91.67% (followed by the 7-13 age interval with the survival rate of 84.62%). The 25-31 age interval are the least likely to have survived with a survival rate of 29.41% (ignoring the 43-49 age interval consisting of just four individuals). The second lowest survival rate of 42.86% is shared by the 1-7 and 31-37 age intervals.

Very young children and older individuals are more vulnerable to famine and hypothermia. This is likely why the mortality is high for young children and older party members.

(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: The survival and age category variables are independent.
H_A: The survival and age category variables are dependent.

According to the output, the value of the test statistic (to three decimal places) is 25.908 and the statistic follows a chi-squared distribution with 10 degrees of freedom. According to the output, the *P*-value of the test is 0.0039. Since the *P*-value is smaller than α = 0.01, we reject H_0 and conclude that there is evidence of a relationship between survival and age. However, the test is not trustworthy because over 20% of cells have an expected count less than 5.

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?

Contingency table results:
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

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

People travelling alone had the lowest survival rate of 18.75%. The survival rate was much higher for almost every other group size, especially the group size of 9 where everyone survived (group 7 has 100% survival rate as well, but there are curiously only 6 people listed). Note that all people travelling alone were males. It appears that group size (possible lack of social support) played a major role in the survival rate of single males. There is not enough evidence that survival rate increases as group size increases.

(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_0: The survival and group size variables are independent.
H_A: The survival and group size variables are dependent.

According to the output, the value of the test statistic (to three decimal places) is 22.073 and the statistic follows a chi-squared distribution with 8 degrees of freedom. According to the output, the *P*-value of the test is 0.0048. Since the *P*-value is smaller than α = 0.01, we reject H_0 and conclude that there is evidence of a relationship between survival and group size. However, the test is not trustworthy because over 20% of cells have an expected count less than 5.

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.

The chance of survival was apparently related to all three factors studied: gender, age, and group size. Age category and group size (in this particular order) seem to be the most important factors in survival since survival rate was much higher for younger members (though not the extremely young) and also those with social family support. The *P*-values from the corresponding tests are 0.0039 and 0.0048, respectively. Note that people travelling alone (all males) had a very small chance of survival (18.75%).

Gender is also an important factor in survival since survival rate was much higher for females than for males (70.588% vs. 43.396%). The *P*-value of the test relating survival to gender is not as small as the other two tests.

In summary, age, group size, and gender are all important predictors for survival, but the importance of the factors appear to be age first, followed by group size, and then gender.

LAB 3 ASSIGNMENT: MARKING SCHEMA

Proper header and appearance: 10 points (see cover page format on eClass; labs must be **typed**)

Question 1 (6)

Observational study or randomized experiment: 2 points
Generalizations to a broader population: 2 points
Causal inferences: 2 points

Question 2 (4)

Number of cases: 1 point
Categorical variables: 2 points
Numerical variables: 1 point

Question 3 (40)

(a) Frequency table of survival: 2 points
Overall survival rate: 3 points
(b) Three families that proportionally lost most of their members: 2 points
(c) Frequency table of survival rate by group (alone vs. not): 2 points
Survival rate of members travelling alone vs. not alone: 2 points
(d) Frequency table of survival by gender: 2 points
Bar chart of survival for females: 3 points
Bar chart of survival for males: 3 points
Comments: 2 points
(e) Frequency table of survival rate by group (child vs. adult): 2 points
Comments: 2 points
(f) Summaries of survivors and non-survivors: 2 points
Side-by-side boxplots: 3 points
Comments: 4 points
(g) Summary statistics of age of survivors and non-survivors for males: 2 points
Summary statistics of age of survivors and non-survivors for females: 2 points
Comments: 2 points

Question 4 (38)

(a) Contingency table of survival by gender: 4 points
(b) Hypotheses: 2 points
The value of the test statistic: 2 points
The distribution of the test statistic under the null hypothesis: 2 points
The *P*-value: 2 points
Conclusion: 2 points

(c) Percent of the survivors that were females : 2 points
Percent that were female survivors: 2 points

(d) Hypotheses: 2 points
The value of the test statistic: 2 points
The distribution of the test statistic under the null hypothesis: 2 points
The *P*-value: 2 points
Conclusion: 2 points

(e) The relationship between the tests in parts (b) and (d): 3 points

(f) 95% confidence interval: 3 points
Interpretation of the 95% confidence interval: 2 points
Comparison of the confidence interval with the outcome of the test: 2 points

Question 5 (18)

(a) Contingency table of survival by Bin(Age): 4 points
Age groups with two highest survival rates: 2 points
Age groups with two lowest survival rates: 2 points

(b) Hypotheses: 2 points
The value of the test statistic: 2 points
The distribution of the test statistic under the null hypothesis: 2 points
The *P*-value: 2 points
Conclusion: 2 points

Question 6 (18)

(a) Contingency table of survival by group size: 4 points
Comments on lowest/highest survival rates: 2 points
Answers to other questions: 2 points

(b) Hypotheses: 2 points
The value of the test statistic: 2 points
The distribution of the test statistic under the null hypothesis: 2 points
The *P*-value: 2 points
Conclusion: 2 points

Question 7 (5)

Summary: 5 points

TOTAL = 139

Created by: Henryk Kolacz;
Edited by: Paul Cartledge
University of Alberta
October 2017