STAT 151 X01

Group #33

Lab 1

**Dhaliwal Jaskaranjit** Robert Joseph

Zheng Jingyi

1.*First discuss the data in the file. How many cases are there? Identify categorical and numerical variables in the data.*

Categorical Variables: Pclass, Survived, Sex

Numerical Variables: Age

2. *Now you will use frequency tables to summarize the gender composition aboard Titanic and obtain the proportions of passengers who survived for each gender.*

**Frequency table results for sex:**

Count = 1309

|  |  |  |
| --- | --- | --- |
| **sex** | **Frequency** | **Relative Frequency** |
| female | 466 | 0.35599694 |
| male | 843 | 0.64400306 |

64.4% men and 35.6% women boarded Titanic.

(b) *What percentage of the passengers survived the disaster? Obtain the appropriate relative frequency table to answer the question. Paste the table into your report. Briefly describe the output.*

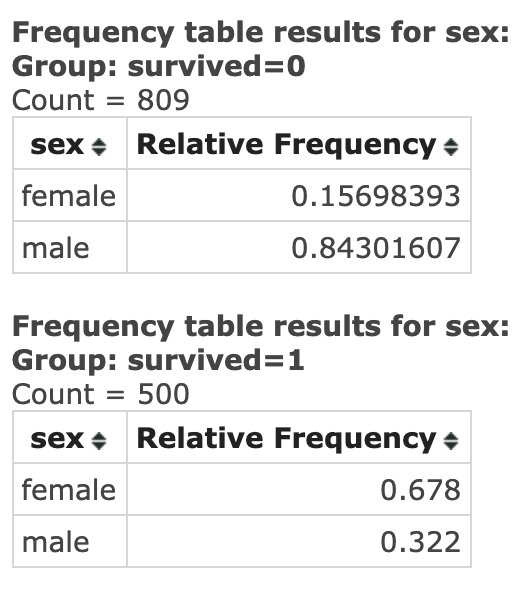
**Frequency table results for survived:**

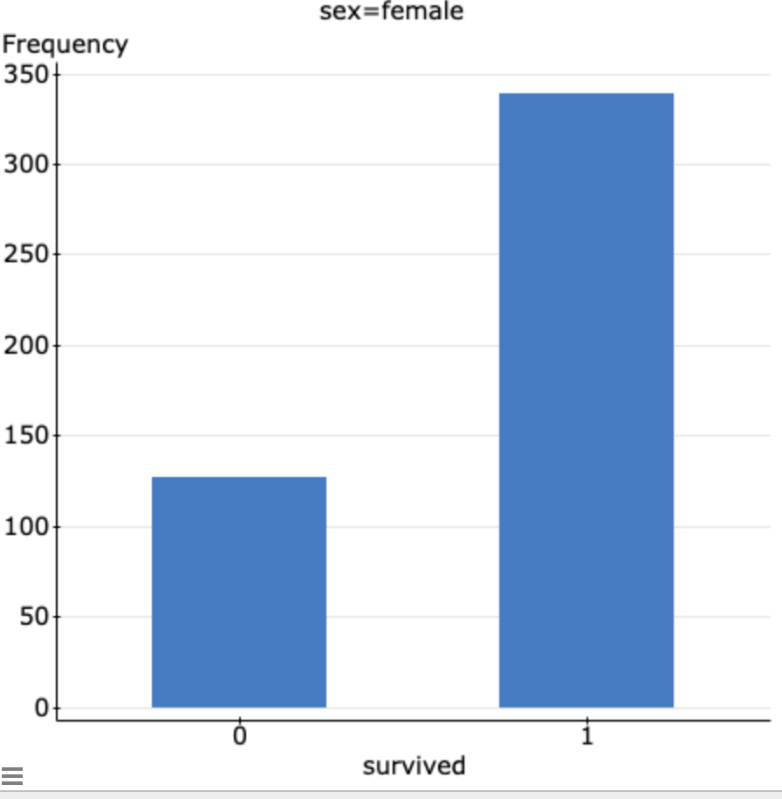
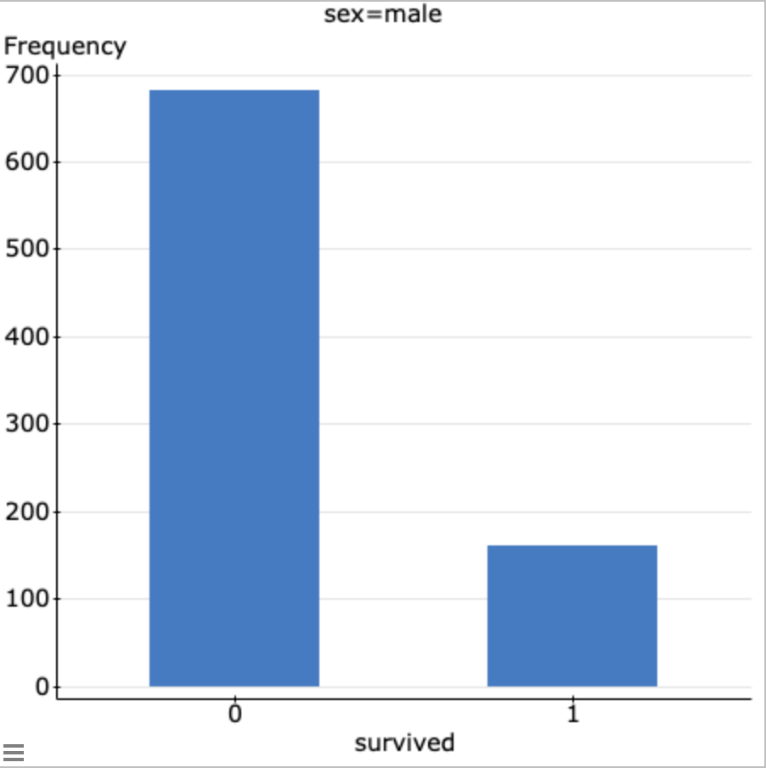
Count = 1309

|  |  |  |
| --- | --- | --- |
| **survived** | **Frequency** | **Relative Frequency** |
| 0 | 809 | 0.61802903 |
| 1 | 500 | 0.38197097 |

Only 38.2% passengers survived the disaster. That means only 500 out of 809 survived. While 61.8% lost their lives on the Titanic.

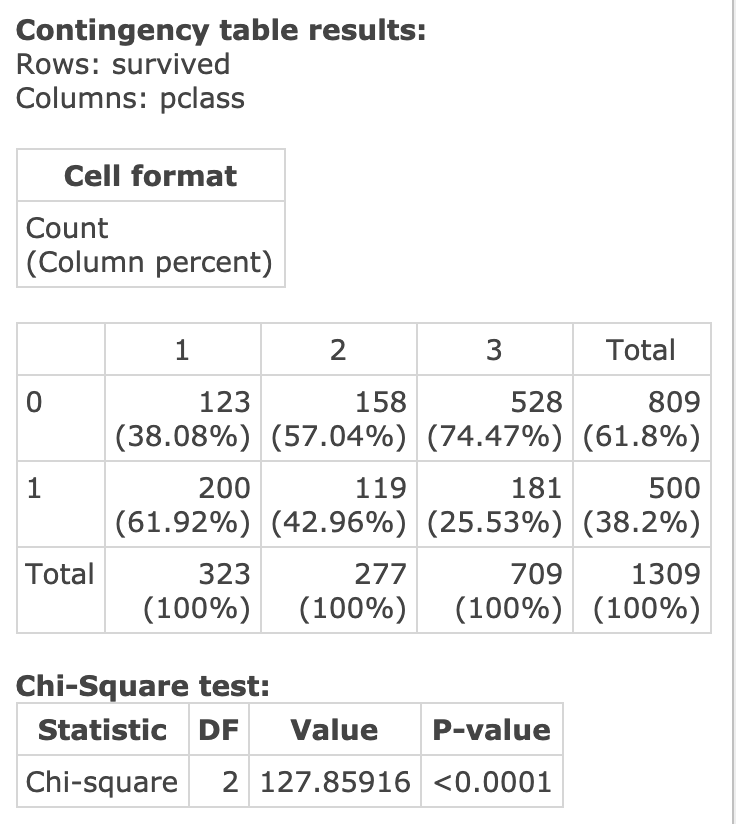
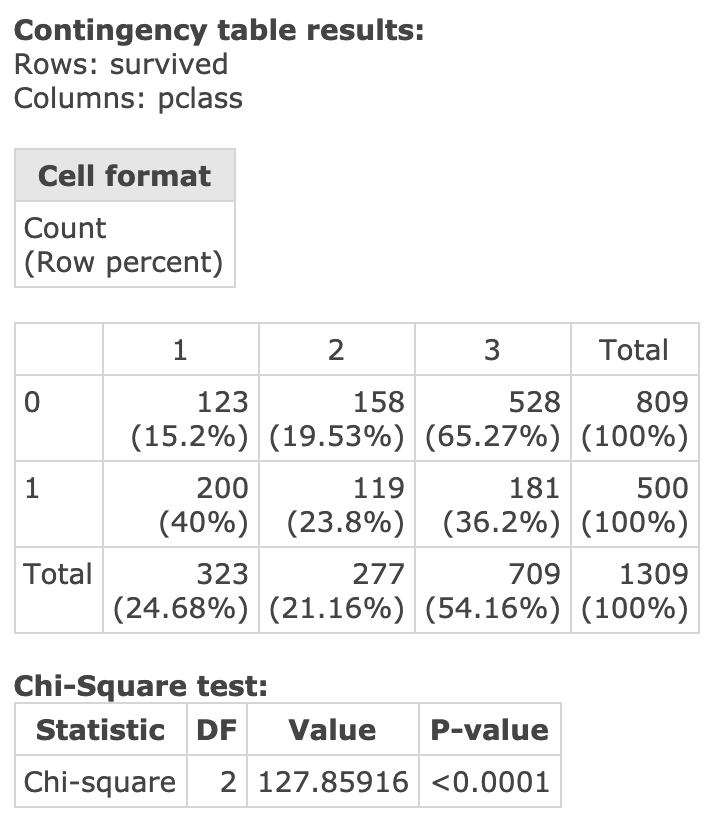
(c) *Obtain the frequency table displaying the proportions of males and females who survived. Paste the tables into your report. Moreover, obtain the corresponding bar chart of survival by gender (separate graph for each gender). Paste the two charts into your report. Does gender affect survival?*



  
  
  
  
  
  
Therefore gender and survival are both independent quantities which implies gender does not affect survival.

**3.** *In this question, you will examine the relationship between survival and passenger class with contingency tables.*

**(a)** *First you will examine the relationship between survival and passenger class (ignore gender or age). Obtain the contingency table of survival by passenger class. Make sure that Row percent, Column percent, and Chi-Square options are checked in each contingency table dialog box. Paste the table into your report. In which of the three passenger classes were passengers most likely to have survived? In which class were the passengers least likely to have survived?*



Looking at the table, Class 1 passengers have a higher survival rate than Class 2 and 3. But Class 2 has the lowest survival rate compared to the other two classes which were only 119 in number.

**(b)** *Refer to the output in part (a) to answer the following questions: What percent of the survivors were in third class? What percent of all passengers were third-class passengers who survived? What percent of the third-class passengers survived?*

13.8% survivors were from the third class. While only 36.2% passengers survived out of all the third-class passengers. And 25.53% of the third-class passengers survived.

**(c)** *Using α = 0.01, test that there was no relationship between survival and passenger class. 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.*

**Chi-Square test:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **DF** | **Value** | **P-value** |
| Chi-square | 2 | 127.85916 | <0.0001 |

H0: The two variables of survival and passenger classes are independent of each other.

HA: The two variables of survival and passenger classes are dependent on each other.

Test statistics = 127.9 according to the Chi-squared distribution.

P(value) < 0.0001

As the P(value) from the data is less than 0.0001 which is less than the α = 0.01. So, we will reject the null hypothesis that survivor of classes are independent of each other. There is enough evidence to prove that the survival rate is dependent on the class.

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

**(a)** *Obtain the contingency table of survival by gender. Make sure that Row percent, Column percent, and Chi-Square options are checked in each contingency table dialog box. Paste the table into your report. Using α = 0.01, 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.*

**Contingency table results:**

**Rows: survived**

**Columns: sex**

|  |
| --- |
| **Cell format** |
| **Count**  **(Row percent)**  **(Column percent)** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **female** | **male** | **Total** |
| **0** | **127**  **(15.7%)**  **(27.25%)** | **682**  **(84.3%)**  **(80.9%)** | **809**  **(100%)**  **(61.8%)** |
| **1** | **339**  **(67.8%)**  **(72.75%)** | **161**  **(32.2%)**  **(19.1%)** | **500**  **(100%)**  **(38.2%)** |
| **Total** | **466**  **(35.6%)**  **(100%)** | **843**  **(64.4%)**  **(100%)** | **1309**  **(100%)**  **(100%)** |

**Chi-Square test:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **DF** | **Value** | **P-value** |
| **Chi-square** | **1** | **365.88695** | **<0.0001** |

H0: The two variables of survival and passenger classes are independent of each other.

HA: The two variables of survival and passenger classes are dependent on each other.

Test statistics = 365.9

We are using the Chi-Squared distribution table to prove that survival rate of women is the same as the men.

P(value) < 0.0001 which is less than *α = 0.01.* So, we reject the null hypothesis.

 There is enough evidence which proves that the female survivors are not the same as the male survivors.

**(b)** *Using α = 0.01, is there evidence that there was a difference in the survival rate for women and men? 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 tests:**

p1 : Proportion of successes (Success = 1) for survived where sex="female"

p2 : Proportion of successes (Success = 1) for survived where sex = "male"

p1 - p2 : Difference in proportions

H0 : p1 - p2 = 0

HA : p1 - p2 ≠ 0

**Hypothesis test results:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Difference** | **Count1** | **Total1** | **Count2** | **Total2** | **Sample Diff.** | **Std. Err.** | **Z-Stat** | **P-value** |
| p1 - p2 | 339 | 466 | 161 | 843 | 0.53648323 | 0.02804676 | 19.128172 | <0.0001 |

Test Statistics = 19.1 calculated based on the Z- distribution tables to prove the null hypothesis.

The P(value) < 0.0001 which is less than the given value of *α = 0.01*. That’s why, we reject the null hypothesis and there is enough evidence which proves that there was a difference in the survival rate for women and men.

**(c)** *What is the relationship between the tests in parts (a) and (b)?*

Both tests from part (a) and (b) rejected their null hypothesis which stated that the number of female survivors were same as the male survivors. Both tests agreed that there is a difference in the survival rate of females compared to that of males.

**5.** *In this question, you will examine the relationship between survival and passenger class for children aboard Titanic.*

**(a)** *Obtain the appropriate contingency table to examine the relationship between survival and passenger class for children (males and females combined; under 12 years of age). Make sure that Row percent, Column percent, and Chi-Square options are checked in each contingency table dialog box. Paste the table into your report. Using the same process as Question 3 part (a), do the summaries indicate that the survival rate for child passengers depended on passenger class?*

**Contingency table results:**

Where: age < 12

Rows: survived

Columns: pclass

|  |
| --- |
| **Cell format** |
| Count  (Row percent)  (Column percent) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | Total |
| 0 | 1  (2.5%)  (20%) | 0  (0%)  (0%) | 39  (97.5%)  (60.94%) | 40  (100%)  (43.96%) |
| 1 | 4  (7.84%)  (80%) | 22  (43.14%)  (100%) | 25  (49.02%)  (39.06%) | 51  (100%)  (56.04%) |
| Total | 5  (5.49%)  (100%) | 22  (24.18%)  (100%) | 64  (70.33%)  (100%) | 91  (100%)  (100%) |

**Chi-Square test:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **DF** | **Value** | **P-value** |
| Chi-square | 2 | 25.911441 | <0.0001 |

Warning: over 20% of cells have an expected count less than 5.

Chi-Square suspect.

*The p value is less than 0.0001 which is also less than* α = 0.01. Therefore, we reject the null hypothesis. There is enough evidence which proves that the survival rate does not depend on the passenger class.

*(b) The emergency protocol “women and children first” was followed during evacuation to ensure that women and children were the first to board lifeboats. Was the rule effective in saving lives of women and children aboard the Titanic? Using α = 0.05, carry out the appropriate test to answer the question. Paste the output into your report. 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:**

p1 : Proportion of successes (Success = 1) for survived where sex = "female" or age < 12

p2 : Proportion of successes (Success = 1) for survived where sex = "male" and age >= 12

p1 - p2 : Difference in proportions

H0 : p1 - p2 = 0

HA : p1 - p2 > 0

***Hypothesis test results:***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Difference*** | ***Count1*** | ***Total1*** | ***Count2*** | ***Total2*** | ***Sample Diff.*** | ***Std. Err.*** | ***Z-Stat*** | ***P-value*** |
| *p1 - p2* | *365* | *515* | *109* | *609* | *0.52975593* | *0.029563131* | *17.91948* | *<0.0001* |

*(c) Obtain a 95% confidence interval for the difference in proportions of survivors of women and children compared to adult males. Paste the output into your report. What do you conclude? Does it confirm your result in part (b)?*

**Two sample proportion confidence interval:**

p1 : Proportion of successes (Success = 1) for survived where sex = "female" and age < 12

p2 : Proportion of successes (Success = 1) for survived where sex = "male" and age >= 12

p1 - p2 : Difference in proportions

**95% confidence interval results:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Difference** | **Count1** | **Total1** | **Count2** | **Total2** | **Sample Diff.** | **Std. Err.** | **L. Limit** | **U. Limit** |
| p1 - p2 | 25 | 42 | 109 | 609 | 0.41625616 | 0.077315681 | 0.26472021 | 0.56779211 |

The 95% confidence interval does not contain the value 0 and both upper and lower bounds are positive therefore it confirms with the result in b).

6. *In this question, you will explore the relationship between age and survival. First you will divide age into several non-overlapping intervals so that the age of each passenger 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 0 and a binwidth of 12 (see “Introductory Lab”, pages 15-16). Make sure that 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 Chi-Square options are checked in each contingency table dialog box. Paste the table into your report. Using the same process as Question 3 part (a), which age groups were the most likely to have survived and which the least likely to have survived? (Ignore age groups with less than five passengers.)*

Question 6 (14) (a) Contingency table of survival by age category: 4 points Highest/lowest survival rates: 2 points

*(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.*

 (b) 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

*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.*

Question 7 (5) Summary: 5 points