Inference of Population Proportions Categorical Variables

Arjun Mannem

2023-11-07

#### Installing necessary packages

# Load necessary packages  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(janitor)

## Warning: package 'janitor' was built under R version 4.3.2

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(broom)

#### We can set default theme settings for plots, and load some functions to simplify table customization and creation using the code below.

# Set ggplot theme for visualizations  
theme\_set(ggthemes::theme\_few())  
  
# Set options for flextables  
set\_flextable\_defaults(na\_str = "NA")  
  
# Load function for printing tables nicely  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

# Chi2-Test of Independence

**What are the response and explanatory variables in each scenario?**

1. **Response Variable**: Parkinson’s disease

* **Explanatory Variable**: Mutation in the SNCA gene

1. **Response Variable**: Preferred smartphone brand

* **Explanatory Variable**: Age group of customers (young, middle-aged, and senior)

1. **Response Variable**: Visitor to the site makes a purchase or not

* **Explanatory Variable**: Layout type of a website (layout A, B, or C)

1. **Response Variable**: Risk of a skin reaction when taking abacavir, a drug for HIV infections

* **Explanatory Variable**: Mutation in the HLA-B gene

#### Let’s import data from the website GitHub to use for this activity, and print 7 randomly selected rows from the data.

# Importing data from course GitHub page  
titanic <- read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/titanic.csv")

## Rows: 887 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): survived, sex, passenger\_class, name  
## dbl (4): age, fare, n\_siblings\_spouses\_aboard, n\_parents\_children\_aboard  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

set.seed(1994)  
  
# Printing sample of 7 rows from data set  
titanic %>%   
 dplyr::sample\_n(size = 7) %>%   
 make\_flex(caption = "Randomly selected people who were aboard the Titanic.")

Table 1: Randomly selected people who were aboard the Titanic.

| survived | sex | passenger\_class | name | age | fare | n\_siblings\_spouses\_aboard | n\_parents\_children\_aboard |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Yes | Female | Second | Miss. Amelia Brown | 24.00 | 13.00 | 0.00 | 0.00 |
| Yes | Female | Third | Miss. Anna Kristine Salkjelsvik | 21.00 | 7.65 | 0.00 | 0.00 |
| No | Male | Third | Mr. Nils Johan Goransson Olsson | 28.00 | 7.85 | 0.00 | 0.00 |
| Yes | Female | First | Mrs. William Thompson (Edith Junkins) Graham | 58.00 | 153.46 | 0.00 | 1.00 |
| No | Male | Third | Mr. Ole Martin Olsen | 27.00 | 7.31 | 0.00 | 0.00 |
| No | Male | Second | Mr. Thomas Charles Mudd | 16.00 | 10.50 | 0.00 | 0.00 |
| Yes | Female | Second | Miss. Alice Herman | 24.00 | 65.00 | 1.00 | 2.00 |

**What are the response and explanatory variables in this scenario?**

**Response Variable**: Survival Status (Yes/No)

**Explanatory Variable**: Passenger class (Third, Second, or First Class)

**Formally state the hypotheses for our question of interest.**

H0: There is not a relationship between a passenger’s class and whether or not they survived the sinking of the Titanic.

Ha: There is a relationship between a passenger’s class and whether or not they survived the sinking of the Titanic.

#### Let’s obtain some output relevant to our question of interest.

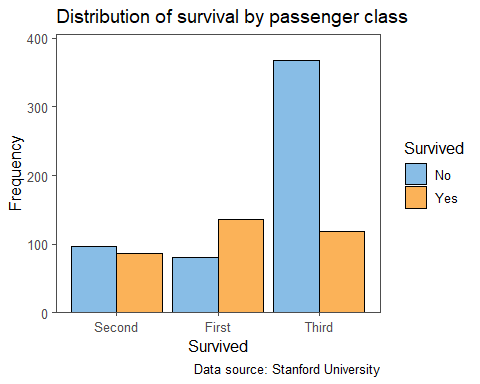
# Creating contingency table  
contingencyTable <- titanic %>%   
 janitor::tabyl(var1 = passenger\_class, var2 = survived)  
  
# Printing table  
contingencyTable %>%   
 make\_flex(caption = "Observed counts for survival status by passenger class", ndigits = 0)

Table 2: Observed counts for survival status by passenger class

| passenger\_class | No | Yes |
| --- | --- | --- |
| First | 80 | 136 |
| Second | 97 | 87 |
| Third | 368 | 119 |

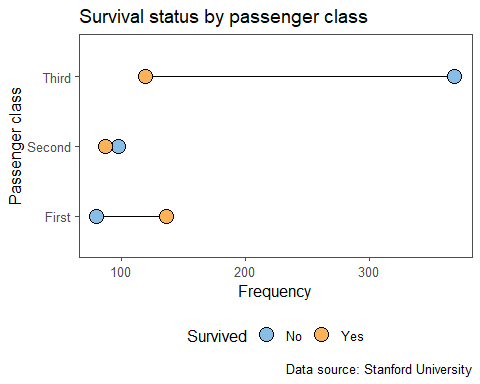
#### We can create a clustered bar chart to visualize the same information conveyed in the contingency table.

# Creating a clustered bar chart  
titanic %>%   
 dplyr::count(passenger\_class, survived, .drop = FALSE) %>%   
 dplyr::filter(!is.na(passenger\_class), !is.na(survived)) %>%   
 mutate(passenger\_class = fct\_reorder(passenger\_class, n)) %>%   
 ggplot(aes(x = passenger\_class, y = n,  
 fill = survived)) +   
 geom\_col(position="dodge", color = "black") +  
 scale\_fill\_few() +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of survival by passenger class",  
 y = "Frequency",  
 x = "Survived",  
 caption = "Data source: Stanford University",  
 fill = "Survived")



#### A dumbbell chart can also be used to create the relationship between two categorical variables.

# Creating dumbbell chart of survival status by passenger class  
titanic %>%   
 dplyr::count(survived, passenger\_class, .drop = FALSE) %>%   
 dplyr::filter(!is.na(survived), !is.na(passenger\_class)) %>%   
 ggplot(aes(x = n, y = passenger\_class,  
 color = survived, fill = survived)) +   
 geom\_line(aes(group = passenger\_class), color = "black") +  
 geom\_point(pch = 21, color = "black", size = 5) +  
 scale\_fill\_few() +  
 labs(title = "Survival status by passenger class",  
 x = "Frequency",  
 y = "Passenger class",  
 fill = "Survived",  
 caption = "Data source: Stanford University") +  
 theme(legend.position = "bottom",  
 strip.background.y = element\_rect(linetype = "solid", color = "black"))



**Based on the clustered bar chart and dumbbell plot, does there appear to be a difference in the likelihood of survival by the class of the passengers?**

Yes, there are clear differences in the likelihood of survival based on the class of the passengers. First class passengers has the greatest survival chances and the third class passengers has the lowest chances of survival.

#### Implement the Chi2-test of independence.

# Implementing a chi-square test of independence  
chi2Res <- chisq.test(contingencyTable, tabyl\_results = TRUE)  
# Printing table of expected counts  
chi2Res$expected %>%   
 make\_flex(caption = "Expected counts for survival status by passenger class", ndigits = 1)

Table 3: Expected counts for survival status by passenger class

| passenger\_class | No | Yes |
| --- | --- | --- |
| First | 132.7 | 83.3 |
| Second | 113.1 | 70.9 |
| Third | 299.2 | 187.8 |

#### Results of the test

# Printing model output  
chi2Res %>%   
 broom::tidy() %>%   
 make\_flex(caption = "Results of the chi-square test of independence",  
 ndigits = 2)

Table 4: Results of the chi-square test of independence

| statistic | p.value | parameter | method |
| --- | --- | --- | --- |
| 101.22 | <2e-16 | 2 | Pearson's Chi-squared test |

**State each assumption for the 2-test of independence, and indicate whether or not each assumption is met, citing specific evidence from the output obtained.**

Some conditions must be met to conduct a valid 2-test of independence:

* Independent observations (commonly violated when observations consist of repeated measurements across time)

*Since these were all separate passengers, we will consider this sufficiently met.*

* Both variables being studied must be categorical

*Both survival status and passenger class are categorical, so this is met.*

* All expected counts under the null hypothesis should be at least 5

*The smallest expected count is 70.9 which is larger than 5, so this is met.*

**What are the degrees of freedom associated with the 2-test here?**

N-1 = 2

**State our test statistic, p-value, decision, and conclusion in the context of the problem testing at the 5% significance level, citing specific evidence from the obtained output.**

*test statistic*: 101.22

*p-value*: <2e-16

*decision*: Reject the H0 (null hypothesis), since the p-value is less than 0.05

*interpretation of the result in context*: We have sufficient evidence that there is a relationship between a passenger’s class and whether or not they survived the sinking of the Titanic at the 5% significance level.

## Inference of population proportions

#### Confidence interval for p

**What does 1-pHat represent in general?**

1 - pHat represents the proportion of failures.

**Provide and interpret a 95% confidence interval for the survival of third-class passengers based on the Titanic data using the code below.**

Calculating 95% confidence interval for p

# Calculating confidence interval for p  
third\_class\_counts <- contingencyTable %>%   
 dplyr::filter(passenger\_class == "Third") %>%   
 dplyr::select(c("Yes", "No"))  
  
third\_surv\_CI <- prop.test(x = third\_class\_counts$Yes,  
 n = sum(third\_class\_counts),  
 conf.level = 0.95,  
 correct = FALSE)  
  
# Printing table of results  
third\_surv\_CI %>%   
 broom::tidy() %>%   
 make\_flex(caption = "Confidence interval for proportion of 3rd-class passengers who survived on the Titanic",  
 ndigits = 3)

Table 5: Confidence interval for proportion of 3rd-class passengers who survived on the Titanic

| estimate | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0.244 | 127.312 | <2e-16 | 1 | 0.208 | 0.284 | 1-sample proportions test without continuity correction | two.sided |

The sample proportion, pHat, is 0.244, which is the proportion of third class passengers who survived the sinking of the Titanic.

Lower bound: 0.208

Higher bound: 0.284

We are 95% confident that the proportion of third class passengers who survive luxury cruise ships sinking in the Atlantic is between 0.208 and 0.284.

**Is this a valid confidence interval? Provide specific evidence to support your conclusion.**

Some conditions must be met to construct a valid Wald confidence interval of a population proportion: - Independent observations (commonly violated when observations consist of repeated measurements across time)

*Since these were all saparate passengers, we will consider this sufficiently met.*

* Variable being studied must be categorical

*The survival status (Yes/No) is a binary categorical variable, so this is met.*

* Number of successes should be at least 5 and number of failures should be at least 5

*Since the number of failures, 368, and the number of successes 119, are both greater than 5, this is met.*

**Provide and interpret a 95% confidence interval for the survival of first-class passengers based on the Titanic data, reproducing the output below.**

# Calculating confidence interval for p  
first\_class\_counts <- contingencyTable %>%   
 dplyr::filter(passenger\_class == "First") %>%   
 dplyr::select(c("Yes", "No"))  
  
first\_surv\_CI <- prop.test(x = first\_class\_counts$Yes,  
 n = sum(first\_class\_counts),  
 conf.level = 0.95,  
 correct = FALSE)  
  
# Printing table of results  
first\_surv\_CI %>%   
 broom::tidy() %>%   
 make\_flex(caption = "Confidence interval for proportion of 1st-class passengers who survived on the Titanic",  
 ndigits = 3)

Table 6: Confidence interval for proportion of 1st-class passengers who survived on the Titanic

| estimate | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0.630 | 14.519 | 0.000139 | 1 | 0.563 | 0.691 | 1-sample proportions test without continuity correction | two.sided |

The sample proportion, pHat, is 0.630, which is the proportion of first-class passengers who survived the sinking of the Titanic.

Lower bound: 0.563

Higher bound: 0.691

**Provide and interpret a 95% confidence interval for the survival of first-class passengers based on the Titanic data, reproducing the output below.**

We are 95% confident that the proportion of first class passengers who survive luxury cruise ships sinking in the Atlantic is between 0.563 and 0.691.

**Is this a valid confidence interval? Provide specific evidence to support your conclusion.**

Some conditions must be met to construct a valid Wald confidence interval of a population proportion: - Independent observations (commonly violated when observations consist of repeated measurements across time)

*Since these were all separate passengers, we will consider this sufficiently met.*

* The variable being studied must be categorical

*The survival status (Yes/No) is a binary categorical variable, so this is met.*

* The number of successes should be at least 5 and the number of failures should be at least 5

*Since the number of failures, 80, and the number of successes 136, are both greater than 5, this is met.*

**What is the margin of error for this interval?**

Margin of error = (upper bound - lower bound)/2 = (0.691 - 0.563)/2 = 0.064

## Confidence interval for p1−p2

#### Provide and interpret a 95% confidence interval for the difference in the survival rates of first-class and third-class passengers based on the Titanic data using the code below.

# Calculating confidence interval for p1 - p2  
class\_counts <- contingencyTable %>%   
 dplyr::filter(passenger\_class %in% c("First", "Third")) %>%   
 dplyr::select(c("Yes", "No"))  
  
diff\_surv\_CI <- prop.test(x = as.matrix(class\_counts),  
 n = colSums(class\_counts),  
 conf.level = 0.95,  
 correct = FALSE)  
  
# Printing table of results  
diff\_surv\_CI %>%   
 broom::tidy() %>%   
 make\_flex(caption = "Confidence interval for difference in the survival rates of 1st-class and 3rd-class passengers",  
 ndigits = 3)

Table 7: Confidence interval for difference in the survival rates of 1st-class and 3rd-class passengers

| estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.630 | 0.244 | 96.087 | <2e-16 | 1.000 | 0.310 | 0.460 | 2-sample test for equality of proportions without continuity correction | two.sided |

* confidence interval limits : (0.310, 0.460)

We are 95% confident that first-class passengers on sinking cruise ships in the Atlantic have a survival rate that is between 31% and 46% greater than that of third-class passengers.

**Is this a valid confidence interval? Provide specific evidence to support your conclusion.**

First-class passengers: 80 died and 136 survived

Third-class passengers: 368 died and 119 survived

Since all four of these numbers are at least 5, we have a sufficient number of successes and failures for the confidence interval to be valid.

**Do the results of the 2-test and 95% confidence interval of the difference align? Why or why not?**

Yes, since the survival rates differed between first and third-class passengers (0 is outside the interval 0.310 and 0.460), this aligns with the statistically significant result of the chi-squared test.