Analysis of Titanic dataset

AUTHOR
Michael Dang

PUBLISHED
November 20, 2024

This program reads data on survival of passengers on the Titanic. Find more information in the data dictionary.

Load the tidyverse library

```
library(broom)
library(epitools)
library(tidyverse)
```

Comments on the code

For most of your programs, you should load the tidyverse library. The messages and warnings are suppressed.

In previous programs, I put a label for each chunk inside the curly braces ({}). It is recommended instead to put the label on a separate line inside the program chunk. It is a bit more work to provide a unique label for each chunk, but it helps quite a bit to isolate where to look when your code produces an error.

Read the data and view a brief summary

```
ti <- read_tsv(
  file="../data/titanic.txt",
  col_names=TRUE,
  col_types="ccncn",
  na="NA")
names(ti) <- tolower(names(ti))
glimpse(ti)
```

```
Rows: 1,313
Columns: 5
$ name     <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen Loraine"…
$ pclass   <chr> "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st"…
$ age      <dbl> 29.00, 2.00, 30.00, 25.00, 0.92, 47.00, 63.00, 39.00, 58.00, …
$ sex      <chr> "female", "female", "male", "female", "male", "male", "female…
$ survived <dbl> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1…
```

Comments on the code

Use read_tsv from the readr package to read this file. Use col_names=TRUE because the column names are included as the first row of the file. The col_types="ccncn" specifies the first second and fourth columns as strings and the third and fifth as numeric. There are missing values in this dataset, designated by the letters “NA”.

Replace numeric codes for survived

```
        ti$survived <-
            factor(
                ti$survived,
                level=1:0,
                labels=c("yes", "no"))
```

Comments on the code

The factor function places the levels of a categorical variable in a specific order and (optionally) attaches labels to each level. In this code, the number codes are reordered so that 1 appears first followed by 0. The labels “yes” and “no” are attached to these two codes.

Question 1

Create a new variable, third_class that indicates whether a passenger is in third class or not. The code would look something like this.

```
        ti$third_class <-
          case_when(
            ti$pclass == "1st" ~ "no",
            ti$pclass == "2nd" ~ "no",
            ti$pclass == "3rd" ~ "yes")
```

How many passengers were in the thrid class?

```
        sum(ti$third_class == "yes", na.rm = TRUE)
```

```
[1] 711
```

Question 2

What are the probabilities of survival for third class passengers. How does this compare to the probability of survival for the other passengers.

Get counts of third class by survival

```
        table1 <-xtabs(~third_class+survived, data=ti)
```

```
table1
```

```
           survived
third_class yes  no
        no  312 290
        yes 138 573
```

Interpretation of the output

There were 138 third class passengers survive and 573 third class passengers died.

Get proportions for died/survived by third class

```
table1 |>
  proportions("third_class")
```

```
           survived
third_class       yes        no
        no  0.5182724 0.4817276
        yes 0.1940928 0.8059072
```

Interpretation of the output

The proportion of first/second class passenger who died is 48%. The proportion of third class passenger who died is much higher at 80%

Question 3

Test the hypothesis that the survival probability is different for third class passengers and the other passengers. Interpret the p-value and confidence interval.

- Null hypothesis (H_0): The survival probabilities for third-class passengers and other passengers are the same.
- Alternative hypothesis (H_A): The survival probabilities for third-class passengers and other passengers are different.

```
prop.test(table1, correct=FALSE)
```

```

	2-sample test for equality of proportions without continuity correction

data:  table1
X-squared = 152.08, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2748006 0.3735586
```

```
sample estimates:
   prop 1    prop 2 
0.5182724 0.1940928 
```

Interpretation of the output

- Since the p-value is almost 0, which is less than 0.5; hence, we can reject the null hypothesis and conclude that the survival probabilities for third-class passengers and other passengers are different.
- 95% confidence interval for the difference in survival probabilities is between 27.48% and 37.36%. Since the interval does not contain 0, hence there is a significant difference in survival probabilities for third-class passengers and other passengers.