# Analysis of Titanic dataset

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

This code was written by Steve Simon and Leroy Wheeler on 2024-11-13 and is placed in the public domain.

## Load the tidyverse library

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

## 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…
```

## Replace numeric codes for survived

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

## Question 1: Create a new variable, third_class that indicates whether a passenger is in third class or not.

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

## 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 passengers by survival

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

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

## Get proportions for died/survived by third class status

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

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

## Interpretation of the output for question 2.

If you were a third class passenger, you had less than a 20% chance of survival, whereas if you were a first or second class passenger, you had more than a 50% chance of survival.

## 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.

```
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 for question 3.

The Chi-squared statistic is much larger than the degrees of freedom and the p-value is small. Therefore we will reject the null hypothesis and conclude that there is a statistically significant difference in the mortality rates between the third class passengers and the 1st/2nd class passengers. The 95% confidence interval for the difference in proportions is 0.27 to 0.37. This interval excludes the value of zero and indicates that the mortality rate is at least 27% higher and possibly as much as 37% higher for third class passengers compared to the rest of the 1st/2nd class passengers.