Vincent Lazo
STA 4163
Mini Project 2, Dataset 1

# Question 1: Wage dataset. Is the average wage different for those at different education levels?

**[10 pts] Determine if the assumption of normality is violated. You may either construct graph(s) to visually assess this, or you may conduct a hypothesis test for normality (Shapiro-Wilk test).**

In order to determine if the assumption of normality is violated, I conducted the Shapiro-Wilk test in my code, on line 17. The result of this test yielded $p < 0.05$, therefore proving the population's normality:

```
        Shapiro-Wilk normality test

data:  wage_col
W = 0.87957, p-value < 2.2e-16
```

**[30 pts] Regardless of if the assumptions are violated, conduct a Kruskal-Wallis H-test to answer the question above for your selected dataset.**

Steps:

0) Assume:
   - Both the wage and education samples are random and independent.
   - There are at least 5 measurements in each sample.
   - The population distributions are continuous.

1/2) Null Hypothesis ($H_0$) : The two distributions are identical.
Alternative Hypothesis ($H_A$): The two distributions are different.

3/4) Test statistic: output from code, line 20:

```
        Kruskal-Wallis rank sum test

data:  wage_col by factor(edu_col)
Kruskal-Wallis chi-squared = 767.05, df = 4, p-value < 2.2e-16
```

5) At alpha = 0.05, we reject $H_0$. We see that 2.2e-16 < 0.05. There is enough evidence to conclude that the average wage is different for those at different education levels.

Vincent Lazo
STA 4163
Mini Project 2, Dataset 1

## Question 2: Is there enough evidence to say that the distribution of the class of those who did not survive the Titanic differ from the distribution of class of everyone on the titanic? Conduct a hypothesis test to determine this.

A Chi-Square Goodness-of-Fit Test will be conducted to determine this.

Steps:

0) Assume the distribution of the class of those who did not survive the Titanic and the distribution of class of everyone on the Titanic are multinomial and all the counts are greater than or equal to 5.

1/2) Null Hypothesis (H_0):

- p_first = 24.6%
- p_second = 21.6%
- p_third = 53.8%

Alternative Hypothesis (H_A):

- At least two proportions differ from their expected value.

3/4) Test statistic: output from code, lines 23-27:

```
        Chi-squared test for given probabilities

data:  counts
X-squared = 7.0385, df = 2, p-value = 0.02962
```

5) At alpha = 0.05, we reject H_0. We see that 0.02 < 0.05. There is sufficient evidence to show that the distribution of the class of those who did not survive the Titanic differs from the distribution of class of everyone on the Titanic.

Vincent Lazo
STA 4163
Mini Project 2, Dataset 1

Question 3: A summary of their results is shown below (note that this is just the pooled data from randomized studies). The two variables of interest are the treatment (vaccine or placebo) and whether or not a study member got infected by COVID.

| | COVID status | | |
|---|---|---|---|
| Treatment | Infection | No Infection | Total |
| Vaccine | 77 | 19,634 | 19,711 |
| Placebo | 833 | 18,908 | 19,741 |
| Total | 910 | 38,542 | 39,452 |

**[5 pts] Are the assumptions of a chi-squared test valid? Explain why or why not.**

They are valid since given that we assume the distribution is multinomial, it allows for data to be categorized. Also, the expected counts needing to be at least five denotes that the data is sufficiently large enough to do analysis on. These two assumptions validate a chi-square test for independence.

**[25 pts] Conduct a chi-square test for independence to determine if infection status and treatment are dependent.**

Steps:

0) Assume:
- The distributions of infection status and treatments are multinomial with r*c possible outcomes
- All expected counts are at least five.

1/2) Null Hypothesis (H_0): Infection status and treatment are independent.
Alternative Hypothesis (H_A): Infection status and treatment are dependent.

(continued on next page)

Vincent Lazo
STA 4163
Mini Project 2, Dataset 1

3/4) Test statistic: output from code, result of lines 30-36:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  status
X-squared = 640.21, df = 1, p-value < 2.2e-16
```

5) At alpha = 0.05, we reject H_0. We see that 2.2e-16 < 0.05. There is sufficient evidence to show that infection status and treatment are dependent.