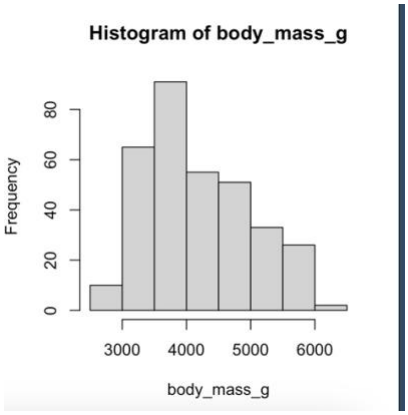
Question 1: In the Penguins dataset, a one-way ANOVA was conducted (Is the average weight different for each species of penguin?)

(a) [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).

Step 0: Assumptions

Assume that the distribution of the average weight for each species of penguin is independent, randomly sampled, continuous, and normally distributed.



Step 1: Null Hypothesis

Ho: The average weight for each species of penguin is normally distributed.

Step 2: Alternative Hypothesis

Ha: The average weight for each species of penguin is not normally distributed.

Step 3: Test Statistic

W = 0.95801

```
        Shapiro-Wilk normality test

data:  body_mass_g
W = 0.95801, p-value = 3.568e-08
```

Step 4: Rejection Region

Determine if graph from R follows a normal or not normal distribution.

$p - \text{value} < 0.95801$

$p - \text{value} = 3.568 \times 10^{-8}$

$(3.568 \times 10^{-8}) < 0.95801$

Reject Ho

Step 5: Conclusion

At $\alpha = 0.05$, we reject Ho and conclude that the average weight for each species of penguin is not normally distributed.

(b) Regardless of if the assumptions are violated, conduct a Kruskal-Wallis H-test to answer the question above for the Penguins dataset.

Step 0: Assumptions

Assume that the 3 samples are random and independent from a continuous population. Each sample has at least 5 observations.

Step 1: Null Hypothesis

Ho: The 3 distributions of average weight are identical.

Step 2: Alternative Hypothesis

Ha: At least 2 distributions of average weight differ.

Step 3: Test Statistic

$H = 212.09$

```
        Kruskal-Wallis rank sum test

data:  body_mass_g by species
Kruskal-Wallis chi-squared = 212.09, df = 2, p-value < 2.2e-16
```

Step 4: Rejection Region

$X^2_{df=2} > X^2_{0.05}$

$H > 2.2 \times 10^{16}$

$212.09 > 2.2 \times 10^{16}$

Reject Ho

Step 5: Conclusion

At $\alpha = 0.05$, we reject Ho and conclude that the distributions of average weight for each species are different.

Question 2: In the *Titanic*, a famous ship that sank in 1912, many unnecessary lives were lost due to many factors. Of the people on the board, did a disproportionate number of those in first class survive compared to those in the other classes?

Of the approximate 1,317 passengers on board, 324 were first class passengers, 284 were second class, and 709 were third class. This means that about 24.6% of passengers were first class, 21.6% were second class, and 53.8% were third class.

From a random sample of 100 passengers who did not survive the *Titanic* (original data source: https://www.kaggle.com/c/titanic/data), 17 were first class, 16 were second class, and 67 were third class.

Is there enough evidence (based on the sample of 100 passengers) 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.

Step 0: Assumptions

Assume that the distribution of the class of those who did not survive the *Titanic* is multinomial and all expected counts are at least 5.

Step 1: Null Hypothesis

Ho: $p_1 = 0.246$, $p_2 = 0.216$, $p_3 = 0.538$

Step 2: Alternative Hypothesis

Ha: At least 2 proportions differ from their expected values.

Step 3: Test Statistic

$X^2 = 7.0385$

```
        Chi-squared test for given probabilities

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

Step 4: Rejection Region

$X^2_{df=2} > X^2_{0.05}$

$X^2 > 0.02962$

$7.0385 > 0.02962$

Reject Ho

Step 5: Conclusion

At $\alpha = 0.05$, we reject Ho and conclude that there is significant difference between the observe counts and expected counts of the distribution of the class of those who did not survive.

[30 pts] Question 3: In 2020, the vaccine for the COVID-19 virus was being developed, and the CDC wanted to ensure that the vaccines were effective at reducing the risk for infection, hospitalization, and death from the virus. The CDC pooled many studies together to evaluate the strength of the evidence (this is called meta-analysis). A link to this summary is shown here: https://www.cdc.gov/vaccines/acip/recs/grade/covid-19-pfizer-biontech-vaccine.html

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

(a) Are the assumptions of a chi-squared test valid? Explain why or why not.

The assumptions of a chi-squared test are valid because the samples of both treatment and COVID status are representative of the population. The expected counts of each cell are also at least 5.

```
         Infection No Infection
Vaccine     454.654     19256.35
Placebo     455.346     19285.65
```

(b) Conduct a chi-square test for independence to determine if infection status and treatment are dependent.

Step 0: Assumptions

Assume that the counts of infection status and treatment are multinomial, and all expected counts are at least 5.

Step 1: Null Hypothesis

Ho: Infection status and treatment are independent.

Step 2: Alternative Hypothesis

Ha: Infection status and treatment are dependent.

Step 3: Test Statistic

$X^2 = 640.02$

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

data:  treatment
X-squared = 640.02, df = 1, p-value < 2.2e-16
```

Step 4: Rejection Region

$X^2 > (2.2 \times 10^{16})$

$640.02 > (2.2 \times 10^{16})$

Reject Ho

Step 5: Conclusion

At $\alpha = 0.05$, we reject Ho and conclude that infection status and treatment are dependent.