In 2007, Warner Bros. President of Production, Jeff Robinov, decreed, “[we are no longer doing movies with women in the lead](https://deadline.com/2007/10/warners-robinoff-gets-in-catfight-with-girls-3362/)” stating their underperformance at the box office as the reason. **In my final capstone, I want to examine whether this sexist statement has any statistical basis.**

My [dataset](https://www.kaggle.com/mathurinache/women-in-movies) comes from Kaggle. It lists 1,794 films from 1970 to 2013 on whether they pass the Bechdel test or not. The Bechdel test was developed by cartoonist Alison Bechdel in 1985 and is one of the most enduring tools to measure Hollywood’s gender bias. Three criteria must be met to pass the Bechdel test: there must be two named women characters, they must speak to each other at some point, and the conversation is not about a man.

Using this dataset, I intend to analyze the relationship between the presence of women in a film and that film’s budget and gross revenue. I will conduct a **z-test on binary variables** (Pass or Fail) to determine if there is a significant difference between the **budget, domestic gross, international gross,** and **profit** of films that Pass the Bechdel Test and films that Fail the Bechdel test.

This analysis could be used by filmmakers and studios to determine whether investing time and money in films that feature women in prominent speaking roles is a wise financial decision.

My hypotheses are as follows:

**Budget**

H₀: µ₁ - µ₂ = 0   
Hₐ: µ₁ - µ₂ ≠ 0

**Domestic Gross**

H₀: µ₁ - µ₂ = 0   
Hₐ: µ₁ - µ₂ ≠ 0

**International Gross**

H₀: µ₁ - µ₂ = 0   
Hₐ: µ₁ - µ₂ ≠ 0

**Profit**

H₀: µ₁ - µ₂ = 0   
Hₐ: µ₁ - µ₂ ≠ 0

**Budget**: My null hypothesis is there is no significant difference between the budget of films that pass or fail the Bechdel test. My alternative hypothesis is that there is a significant difference between the budgets of films that pass or fail the Bechdel test.

**Domestic gross**: My null hypothesis is there is no significant difference between the domestic gross of films that pass or fail the Bechdel test. My alternative hypothesis is that there is a significant difference between the domestic gross of films that pass or fail the Bechdel test.

**International gross**: My null hypothesis is there is no significant difference between the international gross of films that pass or fail the Bechdel test. My alternative hypothesis is that there is a significant difference between the international gross of films that pass or fail the Bechdel test.

**Profit**: My null hypothesis is there is no significant difference between the profit of films that pass or fail the Bechdel test. My alternative hypothesis is that there is a significant difference between the profit of films that pass or fail the Bechdel test.

My [dataset](https://www.kaggle.com/mathurinache/women-in-movies) features 1794 rows and 15 columns.

The columns I will be using for my analysis:

* year (integer): The year the movie was released
* imdb (object): Unique code for each title that is used on IMDB.com
* title (object): Title of the movie
* test (object): Classification of why the film failed or passed with disagreements. This column will be dropped.
* clean\_test (object): Classification of why the film failed or passed
* binary (object): Whether the movie Passed or Failed the Bechdel test
* budget (integer): Budget of movie. This column will be dropped.
* domgross (float): Domestic gross of a movie. This column will be dropped.
* intgross (float): International gross of a movie. This column will be dropped.
* code (object): Contains year of release and binary (pass/fail) for film
* budget\_2013 (integer): Budget of movie adjusted for 2013 dollars
* domgross\_2013 (float): Domestic gross of movie adjusted for 2013 dollars
* intgross\_2013 (float): International gross of movie adjusted for 2013 dollars
* period code (float): Unsure of what data this column stores, therefore this column will be dropped.
* decade code (float): Unsure of what data this column stores, therefore this column will be dropped.

I will be using the **year** column to show how many films per year are released and how many of those films either pass or fail the Bechdel test. **Clean\_test** will be used to categorize how films fail. **Binary** is whether the movie has passed or failed the Bechdel test. **Budget\_2013**, **domgross\_2013**, and **intgross\_2013** will be used for testing of my hypotheses.

<https://www.kaggle.com/mathurinache/women-in-movies>