**Female Speaking Parts in Films: A Script-Based Analysis**

**By Jonathan Shepard**

**Introduction**

There exists an inherent gender bias in the movie business. Anecdotally, for some time it has been clear to industry insiders that men are more central to the plot of most Hollywood films than women. However, quantifying the extent of this bias has proven to be somewhat difficult.

In 2014, the staff of fivethirtyeight.com conducted an analysis of the relationship between women’s centrality (or lack thereof) in movies, and the box office revenues - and profitability - of those films. The project involved using a proxy variable for ‘female involvement’ in films – namely, whether a given film passed or failed the famous “Bechdel test”. The test was initially conceived by cartoonist Alison Bechdel in 1985. A film is said to “pass” the test if it contains at least one scene in which two female characters converse about something other than a man.

[Fivethirtyeight](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/) found that about half of all recent films pass the Bechdel test (though, interestingly, that number drops off significantly for earlier films). More importantly, the researchers found that films that passed the test tended to enjoy better domestic and international box office receipts, and higher profit margins.

The purpose of the present study is to take the fivethirtyeight analysis a step further by analyzing the dialogue contents of the films themselves. That is, rather than classifying a film as either a categorical variable (‘pass’, ‘fewer than two women’, ‘women don’t talk to each other’, etc.), this paper attempts to use a continuous variable to better explore the relationship between women in film and the commercial success (or failure) of a film. That variable is **the percentage of lines of dialogue spoken by women in the film’s original screenplay (**hereafter referred to as %F).

**Research Questions**

The purpose of this study is NOT to generate a generally predictive model for box office revenue. Such an analysis (while certainly interesting) would require a vast array of feature inputs (production budget, marketing budget, genre, celebrity stars, intellectual property, release date, etc. etc.). Rather, the goal is to more fully understand the nature of the relationship between female participation in movies (proxied by %F) and box office revenues.

There are thus several important research questions of interest:

1. In general, what is the distribution of male/female lines of dialogue across movie screenplays? Is there a significant difference in %F between ‘blockbuster’/’tentpole’ films and other types of films? Should these datasets be treated as categorically distinct?
2. Is %F changing over time? How?
3. In general, what is the relationship between production budget and total box office take?
4. Controlling for production budget, what is the relationship between %F and total box office revenues? Domestic revenues? International? Profits?

**Data Sources and Processing**

All financial data for the films in the dataset were gleaned from <http://www.the-numbers.com/>. All movie scripts were scraped from the html on <http://www.imsdb.com/>. The code used to scrape the data and assign a male/female “line count” to each script is available here: [https://github.com/jsshep/Project/blob/master/Script Line Counts.py](https://github.com/jsshep/Project/blob/master/Script%20Line%20Counts.py). The process of assigning the line count was as follows:

1. Identify screenplay online. Confirm that the screenplay is formatted such that the code will parse the text correctly.
2. Run the first part of the program to generate a list of ‘character candidates’ – essentially, a line of bold-faced text in the script.
3. Manually assign a “M” (male), “F” (female), or “X” (neither/unknown) to each “character candidate” in the screenplay.
4. Run the second part of the program to assign a “line count” for all male characters and all female characters.
5. Repeat.

**Caveats**

Before the results can be discussed, a few major caveats should be noted:

1. Unfortunately, only a relatively limited number of screenplays are publicly available. As a result, there may be some unknown amount of sampling bias. In particular, it should be noted that commercially unsuccessful films appear to less commonly published than commercially successful films.
2. The selection of scripts in the “blockbuster” category was decidedly non-random – essentially, the top fifty highest-grossing available scripts were analyzed.
3. The assignment of ‘character candidates’ into M/F/X had to be conducted manually, and is thus subject to human error.
4. It is possible that improper formatting of some screenplays led to ‘false positives’ being counted as lines of dialogue, when they are really descriptions of action.
5. The screenplays posted online, while close to the actual shooting draft of the film, does not represent the actual spoken dialogue of any given film with 100 percent accuracy.

**Data**

The dataset contains the top 49 available “blockbuster” films (total box office gross over $450 million) and 31 randomly-selected “non-blockbuster” films (total box office gross under $450 million).

**SUMMARY DATA:**

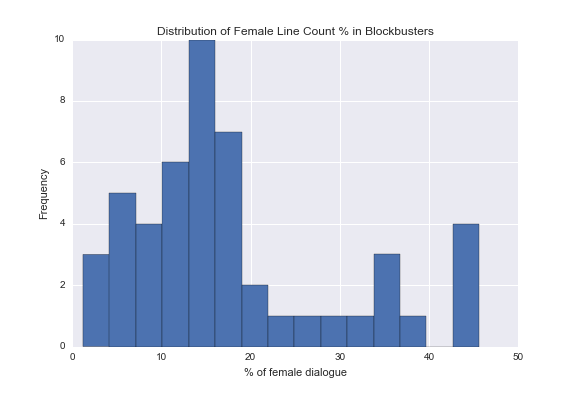
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean Release Year** | **Mean Production Budget** | **Mean Domestic Box Office** | **Mean International Box Office** | **Mean Total Box Office** | **Mean Profit** | **Female Dialogue %** |
| **Blockbusters** | 2002 (9.67) | $115.5 million (72.8 million) | $279.2 million ($118.8 million) | $449.2 million ($280.2 million) | $728.4 million ($380.0 million) | $613.0 million ($339.1 million) | 18.04 (11.97) |
| **Non-blockbusters** | 1997 (10.4) | $27.8 million  ($27.1 million) | $53.7 million ($48.3 million) | $57.5 million ($74.6 million) | $111.2 million ($116.2 million) | $83.4 million ($111.2 million) | 31.00 (28.01) |
| **Total** | 2000  (10.23) | $81.2 million ($73.1 million) | $191.8 million ($147.2 million) | $297.5 million ($294.5 million) | $489.3 million ($429.4 million) | $407.7 million ($376.8 million) | 21.91  (14.27) |

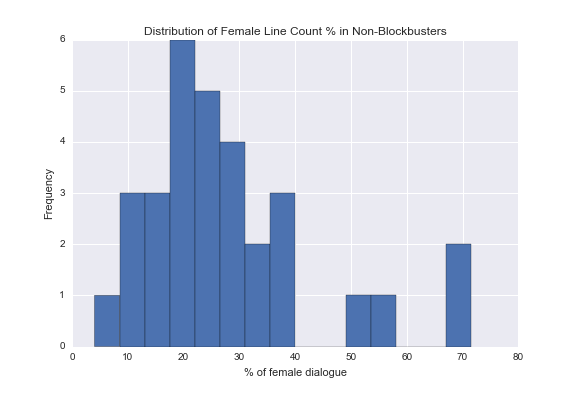
\*Items in parentheses indicate standard deviation

Lumping all movies (blockbusters and non-blockbusters) together in the same dataset results in a fairly high degree of variability across all variables of interest. For the purposes of analysis, therefore, blockbusters and non-blockbusters are split into distinct datasets.

**Analysis, Visualizations, and Findings**

First, let us examine the distribution of F% for both blockbuster and non-blockbusters.

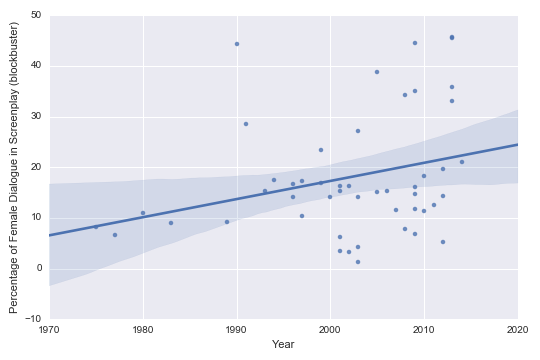




As we might expect, %F is (roughly) normally distributed in both cases. For both blockbusters and non-blockbusters, the vast majority of observations fall below a certain threshold, with a small number of outliers exceeding that threshold. Notably, no blockbuster features a value of %F above 50%. Only four non-blockbusters feature a value of %F above 50%.

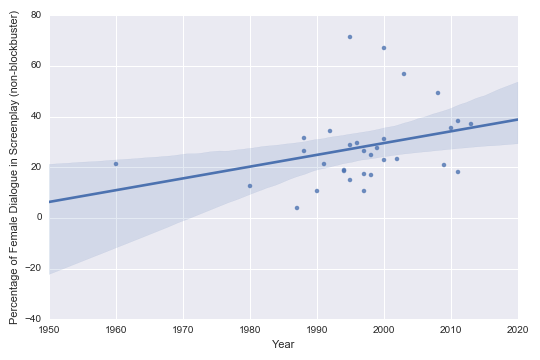
Now let us examine whether F% is changing over time.

**Blockbusters**



Correlation Coefficient = .358

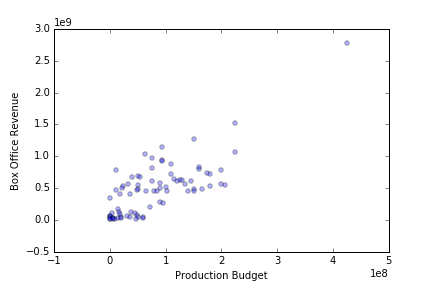
**Non-Blockbusters**



Correlation Coefficient =.465

For both blockbusters and non-blockbusters, F% is going up (albeit gradually) over time.

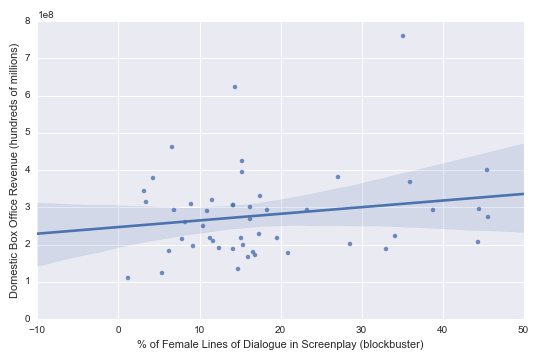
Now let’s take a look at the financial data. First, let’s examine the relationship between production budget and box office revenues. We’d expect this to be a strong positive correlation, and indeed, that’s exactly what we find:



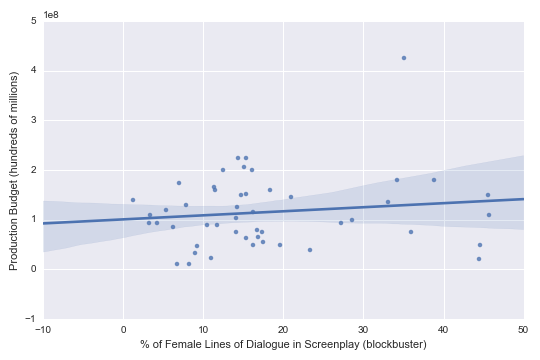
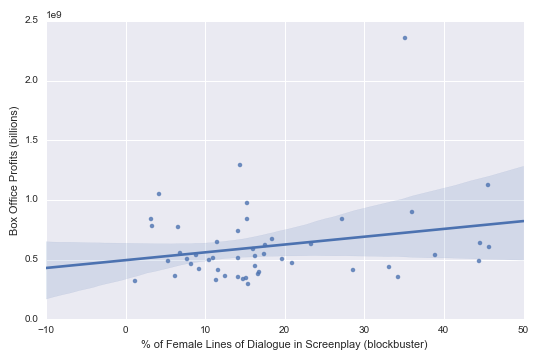
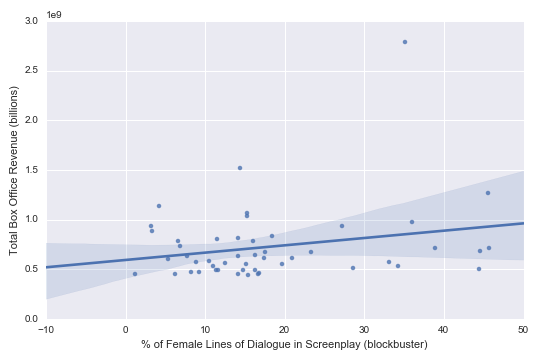
Next, we’ll examine the relationship between F% and financials, starting with blockbusters.

**Blockbusters**

For the ‘blockbuster’ dataset, we observe a positive correlation between F% and: (a) production budget, (b) domestic box office revenue, (c) international box office revenue, and (d) profit:







Correlation coefficient = 814,636. A 1-percentage point increase in F% is associated with a $814,636 increase in production budget.

Building a model that includes both F% and production budget as features (in recognition of the fact that production budgets are crude but significant predictors of box office outcomes), we find the following for **blockbusters:**

|  |  |  |
| --- | --- | --- |
| **Outcome Variable of Interest** | **Correlation Coef.** | **A 1-percentage point increase in F% is associated with…** |
| Domestic box office revenue | $1.3 million | …a $1.3 million increase in domestic box office revenue |
| International box office revenue | $3.5 million | …a $3.5 million increase in international box office revenue |
| Total box office revenue | $4.8 million | …a $4.8 million increase in total box office revenue |
| Profit | $4.8 million | …a $4.8 million increase in total box office profit |

**Non-Blockbusters**

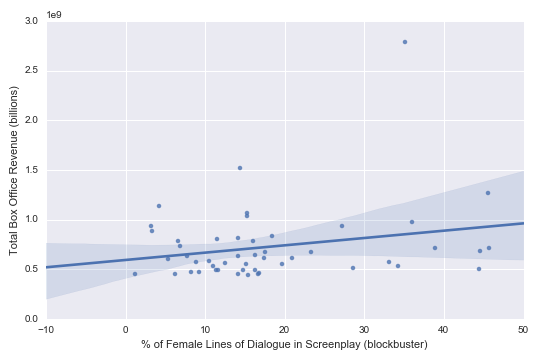
For the ‘non-blockbuster’ dataset, we observe a slight negative relationship between F% and production budgets:

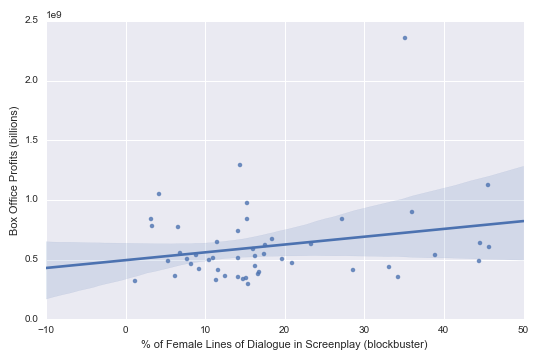


But similar results to the ‘blockbuster’ dataset for everything else:









Building a model that includes both F% and production budget as features (in recognition of the fact that production budgets are crude but significant predictors of box office outcomes), we find the following for **non-blockbusters:**

|  |  |  |
| --- | --- | --- |
| **Outcome Variable of Interest** | **Correlation Coef.** | **A 1-percentage point increase in F% is associated with…** |
| Domestic box office revenue | $745,000 | …a $745,000 increase in domestic box office revenue |
| International box office revenue | $3.5 million | …a $739,000 increase in international box office revenue |
| Total box office revenue | $1.5 million | …a $1.5 million increase in total box office profit |
| Profit | $1.5 million | …a $1.5 million increase in total box office profit |

Conclusion

As noted in the caveats section above, there do exist some sampling challenges with this particular dataset. Insofar as we are essentially ‘starting’ with a dataset of screenplays for the minority of films that are commercially successful, we should be cautious about making undue generalizations about the entirety of the movie business. In addition, with such a labor-intensive manual process of “sex assignment” with this data, we unfortunately had to restrict ourselves to a relatively small dataset (n=49 and 31 for blockbusters and non-blockbusters, respectively). Expanding the dataset would likely reduce variance and reveal other interesting trends.

Two clear patterns emerge from the data:

(1) F% remains abysmally low on average, particularly for high-budget, high-revenue ‘tentpole’ movies.

(2) Both internationally and domestically, there exists a significant relationship between F% and box office revenues.

Taken together, these two facts should induce a greater emphasis on increasing speaking roles for women in Hollywood films.