.

The statistics of the Oscars:

What type of nominee will win?

Jacqueline R. Carlton

Florida Southern College

Thesis Advisor:

Dr. Susan Serrano

Florida Southern College

Abstract

The trends and correlations in the Academy Awards have been in the public eye for years. These trends may lead to the ability to understand the differences between winners and nominees in each category. This study will analyze the Oscars over the years 2000-2019, and identify the different variables that may be significant in predicting a winner from a nominee in each category. This study strives to understand trends in the Oscars, and find correlations between a winning entity and different variables (examples include genre, gender, number of previous nominations, etc). The study will be using statistical methods such as ANOVA analysis, and binary logistic regression along with other statistical tests to calculate the variables that will be the most effective in comparing and understanding the differences between the winners and nominees in each category. Its goal is to explain the different factors that go into an Oscar nominee becoming an Oscar winner.

Keywords: Academy Awards, Oscars, Logistic Regression, Statistics, ANOVA

Introduction

The Academy Awards, also known as the Oscars, is an award show that signifies great achievement in film by a group of film industry experts. The Oscars began in 1929 and it has been an annual award event ever since. A film becomes eligible for an Oscar once it is publicly shown in a commercial theater in Los Angeles for paid admission format for at least a week between January 1 and December 31 of that year (Britannica, T. Editors of Encyclopaedia, 2020). Categories in the Oscars include Actor/Actress in a Leading Role, Actor/Actress in a Supporting Role, Best Picture, Directing, etc.

The information in this study is applicable for career paths in the film industry, such as producers, investors, or even the every-day family. Producers in the film industry are looking to increase the chance of their film becoming popular and making money, and one way to do so is by being recognized by a well-respected, national award (Simonoff & Sparrow, 2000, 15-24). Knowing the different factors that go into a film winning an Oscar could assist them when pitching ideas to get funding for different movies. Being able to claim that their movie has high odds to win an Oscar in a handful of categories would make the project more marketable. This research can assist them in making decisions on who to cast, the length of the movie, and the different number of people they would like on their teams in the film crew. Investors will know what movies they can fund with a high chance of return through popularity of the film. Finally, the every-day family can use this during Oscars parties and can make a more educated prediction of what nominee will win each year.

This research is intended for any person, mathematician or non-mathematician. There should be no background knowledge required to understand anything discussed throughout the

study. The movie industry is something that everyone is involved in, one way or another. Not everyone follows the Oscars, but everyone can distinguish a good movie from a bad one.

A. Current conversation in the field

There have been several researchers that have examined the trends in Oscars, a prevalent name being D.K. Simonton. Simonton has written numerous papers revolving around studies of the Oscars, including but not limited to "The 'Best Actress' Paradox: Outstanding Feature Films Versus Exceptional Womens Performances" (2004), "Collaborative Aesthetics in the Feature Film: Cinematic Components Predicting the Differential Impact of 2,323 Oscar-Nominated Movies" (2002), "Film Awards as Indicators of Cinematic Creativity and Achievement: A Quantitative Comparison of the Oscars and Six Alternatives" (2004), and "Applying Discrete Choice Models to Predict Academy Award Winners" (2008) along with Ian Pardoe.

In "Applying discrete choice models to predict Academy Award winners," Pardoe and Simonton give "The Internet Movie Database" (us.imdb.com) as their main source for their data (2008). This paper did an excellent job at using information that would be available before the award show to predict the winners. For example, they would use information about their genre, their lead actors and actresses, and if the movie had won a Golden Globe, which is held before the Oscars (Pardoe & Simonton, 2008).

Another paper done by Simonton is titled "Collaborative Aesthetics in the Feature Film: Cinematic Components Predicting the Differential Impact of 2,323 Oscar-Nominated Movies" (2002). This paper focuses on the different roles in a movie, such as the director, costumer, actor, special effects, etc., and aims to see which role is the most influential to the Academy Awards

(Simonton, 2002). This gets a bit complicated in the different categories of the Oscars, but for the Best Picture category, this type of comparison would be helpful. From this paper, this study will explore the idea of determining if the number of people on a team would influence the likelihood of a nominee winning in categories like Sound Mixing and Sound Editing, which often have more than one person working in that role.

B. Different holes in the research

Many of these studies had similar holes in the research. The main gap in prior research that tried to predict the Oscar winners is that they only analyzed the 'top four' categories: Best Picture, Directing, Actor in a Leading Role, and Actress in a Leading Role. This study expands on these works by adding each category with multiple nominees in the Oscars to the analysis.

In addition, many of these studies did not include animated films, foreign films, or documentaries. These films are included in this study's calculations, and are exemplified through categories such as: Foreign Language Film, Animated Feature Film, and Documentary (Feature).

The largest hole in the studies previously conducted on the Oscars is that they are over 10 years old. This study adds current data to the conversation, as it includes data up to the 2019 Oscar Awards. Many of the previously conducted Oscars research studies were unable to have access to data that we have now, as they were published over a decade ago.

C. Aims

This study is intended to address several questions, which can be grouped into questions solely about actors and actresses, questions regarding teams, and then questions that deal with all categories in the Oscars. The first of these investigates if there is a difference between the age of actors and actresses that were nominated or winners overall. The age difference among actors and actresses in the Oscars awards is something that has been a topic of conversation and has been proven to be true in many of the previously referenced studies. This study will conduct an inquiry to determine if the difference is also present in the past 20 years through the current data in the Oscars. The next aim in the study is similar, but solely refers to the Oscar winners. It explicitly analyzes if there is a difference in age between actors and actresses that were winners.

Though the main component to the previous question is a difference in age between genders, there is also the idea that a lead actress typically has a different age than her supporting actress counterparts. This raises the question, is there a statistically significant difference between the age of a nominated or winning supporting actress and a nominated or winning leading actress? We follow up this question with one of the same structure that instead investigates the difference between the age of a nominated or winning supporting actor and a nominated or winning leading actor.

In addition to the previous questions regarding age in actors and actresses' chances of winning, this study looks into the relationship of an actor or actress being nominated previously and if it affects their likelihood of winning. There are classic cases of actors and actresses that have been nominated numerous times and have only recently won an Oscar, like Leonardo DiCaprio (*Leonardo DiCaprio - Awards*). Therefore, this study will determine if an actor or

actress is more likely to win their current nomination if the number of their previous nominations is larger.

Aside from inquiries revolving around the actor and actress categories, there is also a curiosity behind team based categories. Team based categories, or group categories, are a subset of the Oscar categories that typically have more than one person in a team for the field of nomination. These categories include Documentary (Feature), Documentary (Short Subject), Film Editing, Sound, Sound Editing, and Sound Mixing. Therefore, this study will evaluate if a nominee with a larger team size is more likely to win in each individual group category.

Another hot button topic surrounding the Oscars is the distribution of different races that are nominated compared to that of the acting population. The main concern is a potential bias in the award show to award the efforts of white actors and actresses more often than actors and actresses of color. This study determines if there is a difference between the frequency of white actors and actresses versus actors and actresses that are people of color being nominated for an Oscar compared to the population distribution of actors and actresses overall.

Finally, this study will investigate the relationship between winning to the runtime of the film in each category. The curiosity for this aim comes from the idea that more time in a film gives the critics more time to view an actor or actress's ability, more time to view set design, more time to hear sound editing, etc. Specifically, this research will test if the odds of winning increases as runtime of the film increases by testing each category distinctly.

Methods

D. Data

This study focuses on analyzing Oscar data from 2000 - 2019 in order to predict winners from nominees of the Oscars in each category. The data in the study excludes honorary awards in the analysis, as there are no nominees for honorary awards. Data was retrieved from datahub.io (Pollock, 2018). The variables used in this study are shown in Table 1.

Table 1

Variable titles and descriptions

Title	Description
Year	The year of the nomination
Category	The category of the nomination
Winner	If the entity won = 1, lost = 0
Entity	The entity of the nomination
Movie title	The movie title of the nomination
Genre	The genre of the movie nominated
Age	The age of the person nominated, only for actor/actress categories
Runtime (Min)	The runtime in minutes of the movie nominated
Gender	The gender of the person nominated, only for actor/actress categories
Race	The race of the person nominated, only for actor/actress categories
Previous Noms	The previous number of nominations of the person nominated, only for actor/actress categories
Team #	The number of people on the team for the movie nominated, only for group categories

The Oscars have a wide variety of categories. This data includes all categories except for the honorary awards, as they are appointed and there are no nominations. A description of each category included in the data set can be found in Table 2.

Table 2

Oscar categories, team category status, and type of nomination

Oscar Categories	Team Category	Type of nomination
Actor in a Leading Role	No	Actor
Actor in a Supporting Role	No	Actor
Actress in a Leading Role	No	Actor
Actress in a Supporting Role	No	Actor
Animated Feature Film	No	Film
Art Direction	No	Film
Best Picture	No	Film
Cinematography	No	Film
Costume Design	No	Film
Directing	No	Film
Documentary (Feature)	**Yes**	Film
Documentary (Short Subject)	**Yes**	Film
Film Editing	**Yes**	Film
Foreign Language Film	No	Film
International Feature Film	No	Film
Makeup	No	Film
Makeup and Hairstyling	No	Film
Music (Original Score)	No	Film
Music (Original Song)	No	Film
Production Design	No	Film
Short Film (Animated)	No	Film
Short Film (Live Action)	No	Film
Sound	**Yes**	Film
Sound Editing	**Yes**	Film
Sound Mixing	Yes	Film
Visual Effects	No	Film
Writing (Adapted Screenplay)	No	Film
Writing (Original Screenplay)	No	Film
Writing (Screenplay Based on Material Previously Produced or Published)	No	Film
Writing (Screenplay Written Directly for the Screen)	No	Film

E. Statistical Testing Methods

Throughout this study, the program used for analysis will be Minitab. Each aim will have a different statistical method used to report results of the inquiry. Some of the main statistical testing methods that will be used include the two-sample *t*-test, two-proportion fisher's exact test, ANOVA, and binary logistic regression.

The two-sample *t*-test determines if two unknown population means are equal based off of their samples. In order to conduct a two-sample *t*-test, the model must have constant parameters (μ_1, $\mu_2$andσ), and the responses from the model must be the sum of the parameters

and error terms. There are also specific assumptions regarding the error terms, including that the error terms are independent and identically distributed, that they follow a normal distribution, that they have a mean of zero, and that the population variance (σ^2) is the same for both groups (Kuiper & Sklar, 2013, 50-52). The typical hypotheses for a two-sample *t*-test are:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0.$$

The two-proportion fisher's exact test is used for any sample size to compare two proportions. The main assumption that must be met for the fisher's exact test to be appropriate is the conditional test of independence, which states that the totals used for the test must be fixed before the study was conducted. The fisher's exact test can be used for categorical data, as the proportions can act as a count in a 2x2 table (Kuiper & Sklar, 2013, 180 - 182). The typical hypotheses for a two-proportion fisher's exact test are:

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2.$$

The analysis of variance, also abbreviated as ANOVA, is a statistical testing method that compares the means of data for more than two samples. For an ANOVA test to be appropriate, the parameters must be constant, each term in the ANOVA model is additive, the error terms must be independent and identically distributed, the error terms must follow a normal probability distribution, the error terms must have a mean of zero, and finally the population variances

within each factor level must be equal (Kuiper & Sklar, 2013, 39 - 40). The typical hypotheses for an ANOVA test are:

$H_{0:}\mu_1 = \mu_2 = \ldots = \mu_n$ where n is the number of samples

H_A:At least one μis different.

Finally, binary logistic regression is a statistical testing method that predicts the relationship between a binary dependent variable and independent variables of all types. A binary variable is a variable that only has two possibilities (Kuiper & Sklar, 2013, 230 - 235). In the case of this study, the binary response variable is whether or not a nominee was a winner, with 0 being no and 1 being yes.

The result of a binary logistic regression model is an equation in the form of $\hat{y} = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \ldots + \beta_n(x_n)$ where $\hat{y}$ is the odds of the response variable being 1, which in this study represents winning an Oscar. Odds describes the "ratio of the probability that an event will occur divided by the probability that the event will not occur" (Mertler & Vannatta, 2010, 289-305). The variable β_0 represents the odds of winning an Oscar when all of the predictors, the dependent variables x_n in the equation, equal zero. Finally, the variable β_n represents the change in odds for each unit increase in the dependent variable x_n. When β_n is positive, then the odds of the response variable, in this study the odds of winning, increases by β_n. When β_n is negative, the odds of the response variable decreases by β_n.

Results

This result section is separated by each question of the study. Each section will describe the sections of data used to answer the question, the test used to answer the question, and the result of the test.

F. Is there a difference between the average age of actors and actresses that were nominees or winners overall?

To determine whether or not there is a difference between the average age of actors and actresses that were nominated or winners in the Oscars between the years of 2000-2019, we used a two-sample *t*-test. For this question, we tested a subset of data that only included the categories that were relevant to actors, which include the following categories: Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role.

As required by the two-sample *t*-test, there are several assumptions that were verified before running the test. By the nature of the dataset, we are able to conclude that the model does have constant parameters, that the responses from the model are the sum of the parameters and error terms, and that the error terms are independent and identically distributed. We are able to use the residual graphs to verify normality of the error terms (Kuiper & Sklar, 2013).

We verified the equal variance assumption by using a Levene test for two variances, with results shown in Figure 1, and it resulted with a p-value of 0.874. The Levene test explains if variances are equal, where a p-value less than or equal to 0.05 means that the variances are statistically different, and a p-value greater than 0.05 means that the variances are not statistically

different (Kuiper & Sklar, 2013). Therefore, with a p-value of 0.874, we concluded that the variances are not statistically different and thus pass the equal variance assumption.

Figure 1

Test for equal variances for use in the two-sample t-test

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	0.12	1		0.726
Levene	0.03	1	398	0.874

In this particular question, we are testing the following hypotheses:

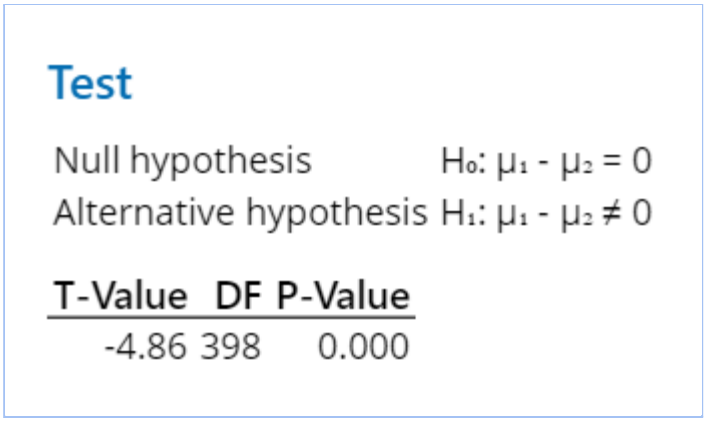
$$H_0: \mu_1 - \mu_2 = 0$$

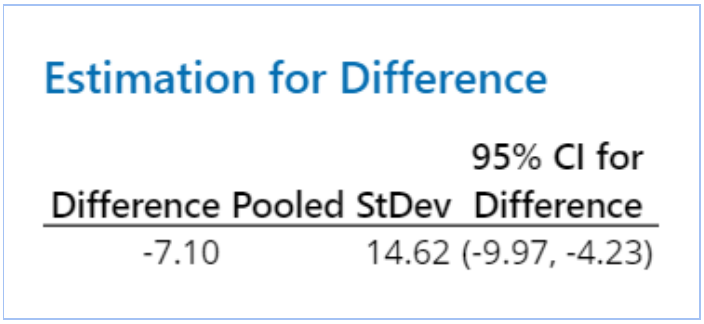
$$H_A: \mu_1 - \mu_2 \neq 0.$$

These hypotheses are written with the assignment of μ_1 representing the average age of actresses and μ_2 representing the average age of actors. The null hypothesis can also be interpreted as the average age of actors and actresses are the same, whereas the alternative hypothesis can be interpreted as the average age of actors and actresses are different. After running the two-sample *t*-test through Minitab, we received the results shown in Figure 2.

Figure 2

Two-sample t-test with confidence interval





Note: This particular two-sample *t*-test is answering the question: Is there a difference between the average age of actors and actresses that were nominees or winners overall?

After receiving such a small p-value of 0.000, we reject the null hypothesis and conclude that there is a significant difference in age between actors and actresses that were nominated or won an Oscar between the years 2000-2019. In addition, we are 95% confident that the true difference when comparing the age of actors and actresses is that actresses are between 9.97 and 4.23 years younger than their actor counterparts. The extremely small p-value led to a new

question, in which we ask, are the actresses younger than the actors? The new hypotheses are as follows:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 < 0$$

These hypotheses are written with the assignment of μ_1 representing the average age of actresses and μ_2 representing the average age of actors. After running the two-sample *t*-test through Minitab, we received the results shown in Figure 3.

Figure 3

Two-sample t-test

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 < 0$

T-Value	DF	P-Value
-4.86	398	0.000

Note: This particular two-sample *t*-test is answering the question: Is the average age of actresses lower than the average age of actors?

This two-sample *t*-test also resulted in a p-value of 0.000, and therefore we reject the null hypothesis and conclude that actresses are statistically significantly younger than actors when

comparing actors and actresses that were nominated or won an Oscar between the years 2000-2019.

G. Is the average age of winning actresses lower than the average age winning of actors?

We now investigate the difference in age between Oscar winning actors and actresses. Due to the result that overall, Oscar nominated or Oscar winning actresses have a lower average age than Oscar nominated or Oscar winning actors, we will specifically be testing to see if this holds true for winners only. We will also test this question using a two-sample *t*-test, and the same assumptions and applicable criteria apply. The main difference is the subset of data we use for this question, as it only involves the actors and actresses that won each year from 2000-2019 in the following categories: Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role.

We verified the equal variance assumption by using a Levene test for two variances, with results shown in Figure 4. As shown below, the test resulted in a p-value of 0.767, and therefore we conclude that the variances are not statistically different and thus pass the equal variance assumption.

Figure 4

Test for two variances between age and gender of the winners in Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

	Test Statistic	DF1	DF2	P-Value
Bonett	0.03	1		0.853
Levene	0.09	1	78	0.767

With all of the assumptions of the two-sample *t*-test passing, we are able to begin the test. The hypotheses to determine if Oscar winning actors are older than Oscar winning actresses are:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 < 0.$$

These hypotheses are written with the assignment of μ_1 representing the average age of actresses that won in their category and μ_2 representing the average age of actors that won in their category. The results of the two-sample *t*-test are shown in Figure 5.

Figure 5

Two-sample t-test with confidence interval

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 < 0$

T-Value	DF	P-Value
-3.13	78	0.001

Estimation for Difference

Difference	Pooled StDev	90% Upper Bound for Difference
-7.80	11.15	-4.58

Note: This particular two-sample *t*-test is comparing the average age of winning actresses to the average age of winning actors.

This two-sample *t*-test resulted in a p-value of 0.001, and therefore we reject the null hypothesis and conclude that Oscar winning actresses are statistically significantly younger than Oscar winning actors when comparing them between the years 2000-2019. In addition, we are 90% confident that the true difference when comparing the age of Oscar winning actors and Oscar winning actresses is that Oscar winning actresses are 4.58 years younger than their Oscar winning actor counterparts.

H. Is there a difference in age between the nominees and winners of actress in a leading role versus actress in a supporting role?

This question comes from the idea that the leading actresses may have a different average age than their supporting actress counterparts. This study analyzes this question with a two-sample *t*-test. For the two-sample *t*-test, it is necessary to check that the assumptions of the test are met before running the test for a particular subset of data. Once again, we are able to conclude that the model does have constant parameters, that the responses from the model are the sum of the parameters and error terms, and that the error terms are independent and identically

distributed by the nature of the dataset. Once again, we are able to use the residual graphs to verify normality of the error terms (Kuiper & Sklar, 2013).

We verify the equal variance assumption by using a Levene test for two variances as we have in the other two-sample *t*-tests. For this test, σ_1 represents the variance of the age in the Actress in a Leading Role category, and σ_2 represents the variance of the age in the Actress in a Supporting Role category. With the resulting p-value of 0.097 shown in Figure 6, we conclude that the variances are not statistically different and thus pass the equal variance assumption.

Figure 6

Levene test for two variances, Actress in a Leading Role and Actress in a Supporting Role

Test

Null hypothesis	H_0: $\sigma_1 / \sigma_2 = 1$
Alternative hypothesis	H_1: $\sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	1.94	1		0.163
Levene	2.78	1	198	0.097

After checking each of the assumptions of the two-sample *t*-test, we are able to begin the test. The hypotheses to determine if there is a difference in age between people nominated in the Actress in a Leading Role and Actress in a Supporting Role categories are:

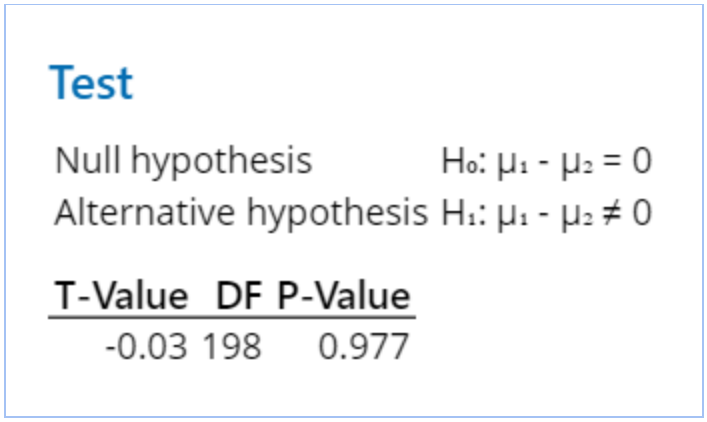
$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0.$$

These hypotheses are written with the assignment of μ_1 representing the average age of actresses in a leading role and μ_2 representing the average age of actresses in a supporting role. The results of the two-sample *t*-test are shown in Figure 7.

Figure 7

Two-sample t-test, actress in a leading role versus actress in a supporting role



This two-sample *t*-test resulted in a p-value of 0.977, and therefore we do not reject the null hypothesis and conclude that there is no statistically significant difference in age between actresses in a leading role and actresses in a supporting role when comparing them between the years 2000-2019.

I. Is the average age of an actor in a leading role lower than the average age of an actor in a supporting role?

This question is parallel to the previous question, but instead investigates the difference in age between actors. We will use a two-sample *t*-test with the subset of data in the Actor in a Leading Role and Actor in a Supporting Role categories. For the two-sample *t*-test assumptions, we are able to conclude that the model does have constant parameters, that the responses from the model are the sum of the parameters and error terms, and that the error terms are independent and identically distributed by the nature of the dataset. We are able to pass normality of the error terms by investigating the residual graphs, just as we have done in all previous two-sample *t*-tests.

We verify the equal variance assumption by using a Levene test for two variances, where σ_1 represents the variance of the age in the Actors in a Lead Role category, and σ_2 represents the variance of the age in the Actors in a Supporting Role category. With the resulting p-value of 0.020 shown in Figure 8, we conclude that the variances are statistically different and thus do not pass the equal variance assumption. This means that we must instead use the two-sample *t*-test for unequal variances.

Figure 8

Levene test for two variances, actor in a leading role versus actor in a supporting role

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
Bonett	7.18	1		0.007
Levene	5.48	1	198	0.020

For the two-sample *t*-test for unequal variances, we find the results shown in Figure 9. We tested the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 < 0.$$

This test's null hypothesis is that the average ages of actors are the same in the categories Actor in a Leading Role and Actor in a Supporting Role, whereas the alternative hypothesis is that the average age between actors in the different categories are different. The results of the test are shown in Figure 9.

Figure 9

Two-sample t-test for unequal variances between actors in a leading role and actors in a supporting role

Method

μ_1: population mean of Age when Category = ACTOR IN A LEADING ROLE
μ_2: population mean of Age when Category = ACTOR IN A SUPPORTING ROLE
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 < 0$

T-Value	DF	P-Value
-2.26	187	0.013

Estimation for Difference

Difference	Pooled StDev	90% Upper Bound for Difference
-4.56	14.29	-1.96

This two-sample *t*-test results in a p-value of 0.013, and therefore we reject the null hypothesis and conclude that there is a difference in average age among actors in the lead role category versus the supporting role category. In addition, we are 90% confident that the true difference in age between lead actors and supporting actors is that lead actors are 1.96 years younger than their supporting actor counterparts.

J. If a person has been nominated before, are they more likely to win in the future?

It is no question that it is common for the actors and actresses that are nominated for an Oscar to be nominated again in the future. When a person has been nominated for an Oscar multiple times without winning, people tend to think that the actor or actress's time is coming, or

that they're due for a win. This study considers and tests the idea that if a person has been nominated before, the person is more likely to win in the future. We test this using a binary logistic regression model with the relationship of winning status and the number of previous nominations represented in the Previous Noms variable.

This test is used only in relation to the categories Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role, as these are the only categories that nominate individuals rather than films. For this test, we used an alpha of 0.10, meaning that any p-value less than 0.10 is statistically significant in the regression equation for calculating the odds of winning using the number of previous nominations.

After running the binary logistic regression model for each category relating to actors and actresses, the only category that resulted in a statistically significant model was in the Actress in a Supporting Role category. As shown in Figure 10, this model resulted in a p-value of 0.077 and a β_n of -0.497, meaning that each additional previous nomination decreases the odds of winning an Oscar in the Actress in a Supporting Role category by 49.7%.

Figure 10

Binary Logistic Regression Model, winner versus previous nominations

ACTRESS SUPPORTING

Binary Logistic Regression: Winner? versus Previous noms

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -0.990 - 0.497 Previous noms

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-0.990	0.301	-3.29	0.001	
Previous noms	-0.497	0.281	-1.77	0.077	1.00

K. Does race play a factor in determining whether a nominee becomes a winner?

The next aim of the study is to determine if a person's race plays a factor in determining whether an Oscar nominee becomes an Oscar winner. For this question, we use an ANOVA test with the subset of data that involves actors and actresses in the categories Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role. Due to the disproportionate sizes of the specific race categories for actors nominated, this study categorizes all races that are represented in the Oscars as either a white person or a person of color.

In order to run an ANOVA test, it is imperative to check that the assumptions of the test are met. The assumptions run parallel to those of the two-sample *t*-test, and so many of them have been addressed in previous sections. The only difference is the assumption that the population variances within each factor level are equal. We ran the fisher's exact test to test if there was a difference in the proportion in race categories, however the results did not support an equal variance assumption, shown in Figure 11.

Figure 11

Fisher's exact test for two proportions

Test

Null hypothesis	$H_0: p_1 - p_2 = 0$
Alternative hypothesis	$H_1: p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	2.51	0.012
Fisher's exact		0.024

Note: Used to investigate the equal variance of levels assumption for ANOVA, which resulted in a non-supportive conclusion.

With a p-value of 0.024, we conclude that the race levels do not have an equal variance and therefore the data does not pass the assumption to use ANOVA with equal variances. We are still able to use the ANOVA testing method, but instead not assuming equal variances throughout the analysis.

Figure 12

ANOVA without assumption of equal variances

Welch's Test

Source	DF Num	DF Den	F-Value	P-Value
Recoded Ethinicity	2	37.7366	1.37	0.267

With the Welch's test resulting in a p-value of 0.267, we cannot conclude that race plays a factor in turning an Oscar nominee into a winner in the years 2000-2019. However, this brings up a flaw in the nature of our dataset in relation to this question - we do not have data on the overall population of people that are eligible for nomination from 2000-2019 and therefore we cannot test to see if race is a significant factor in being nominated in the first place. The distribution of actors and actress's race is heavily skewed, in that there are a disproportionate number of white people being nominated as compared to the number of people of color being nominated for an Oscar award. We explore this concept further in the next section.

L. Is there a difference between the frequency of white people and people of color being nominated for an Oscar compared to the population distribution of actors?

As we saw in the previous section, statistically, race does not play a factor in determining whether an Oscar nominee becomes a winner. However, there still seems to be an uneven distribution of people of color being nominated in the first place. In this study, we investigate if there is a difference between the frequency of white people and people of color being nominated for an Oscar compared to the population distribution of actors through a two-proportion fisher's exact test.

We used data from the Screen Actors Guild, which is an association of actors in America (SAG-AFTRA, n.d.), and data from UCLA Hollywood Diversity Report (Hunt et al., 2019, 14) to compare the populations. The population distribution of actors is shown in Figure 13. The total population for white actors was 128,320 people, and the total population for actors of color was 31,680 people. The Oscar nominated white people amounted to 339, and the Oscar nominated people of color amounted to 61.

Figure 13

Descriptive Statistics for two-proportion fisher's exact test

Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	128320	339	0.002642
Sample 2	31680	61	0.001926

All values used in the two-proportion fisher's exact test pass the conditional test of independence, as the totals used for the test were all fixed before the study was conducted. The hypotheses for this test are:

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2.$$

In this test, p_1 represents the proportion of white actors and actresses that were nominated or won an Oscar to the SAG population of white actors and actresses, and p_2 represents the proportion of actors and actresses of color that were nominated or won an Oscar to the SAG population of actors and actresses of color. In Figure 14, we explore the results of the two-proportion fisher's exact test.

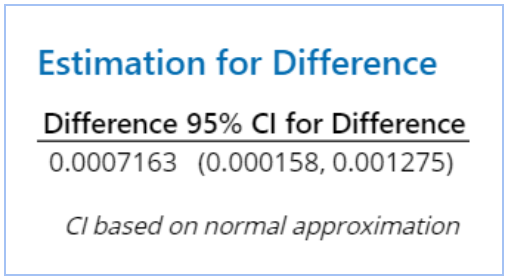
Figure 14

Two-proportion fisher's exact test with confidence interval

Test

Null hypothesis	$H_0: p_1 - p_2 = 0$
Alternative hypothesis	$H_1: p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	2.51	0.012
Fisher's exact		0.024



This two-proportion fisher's exact test resulted in a p-value of 0.024 which is less than the significance level of $\alpha = 0.05$, and therefore we reject the null hypothesis and conclude that there is a difference in the proportions of white actors and actresses and actors and actresses of color in the Oscars relative to the SAG population. In addition, we are able to conclude with 95% confidence that the difference in proportions of the proportions of white actors and actresses and actors and actresses of color in the Oscars relative to the SAG population is between 0.000158 and 0.001275.

M. Is a larger team more likely to win in team categories?

We shift our attention to a new subset of the data: team categories. As shown in Figure 2, the team categories include Documentary (Feature), Documentary (Short Subject), Film Editing, Sound, Sound Editing, and Sound Mixing. We tested each category distinctly using a binary logistic regression model to see if Team #, the independent variable that measures the number of people in the team, is an effective predictor in a nominee's winner status. After testing each category independently, the results shown in Table 3 were obtained.

Table 3

Binary logistic regression model p-values per team category

Category	P-value from Model
Documentary (Feature)	0.223
Documentary (Short Subject)	0.528
Film Editing	0.844
Sound	0.892
Sound Editing	0.250
Sound Mixing	0.283

The p-value for the Team # variable was greater than 0.05 in each instance, meaning that the number of people on the team is not statistically significant when predicting winners in each category. This means that the binary logistic regression model that is populated in each of these categories based on the number of people on the team would not be an effective method of predicting the odds of winning.

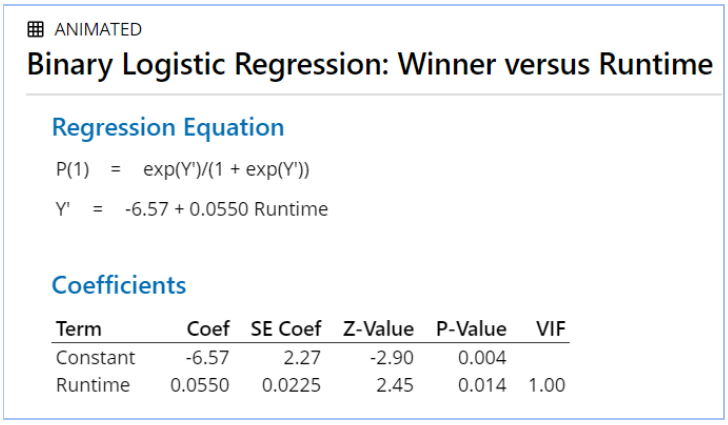
N. As the run time increases, does the probability of winning increase?

To understand if an increase in runtime increases the probability of winning an Oscar, we separated the data in order to test each category individually with a binary logistic model. This study conducted the binary logistic regression models with an alpha, or statistical significance level, of 0.05. Therefore, an increase in runtime would have a statistically significant effect on any model that comes back with a p-value less than or equal to 0.05.

After testing each category individually, only 4 categories returned a statistically significant result. These categories are Animation, Documentary (Feature), Documentary (Short Subject), and Short (Live Action). The results of each of these binary logistic regression models are shown in Figure 15.

Figure 15

Binary Logistic Regression Models testing winner vs runtime for Animation, Documentary (Feature), Documentary (Short Subject), and Short (Live Action)



ANIMATED

Binary Logistic Regression: Winner versus Runtime

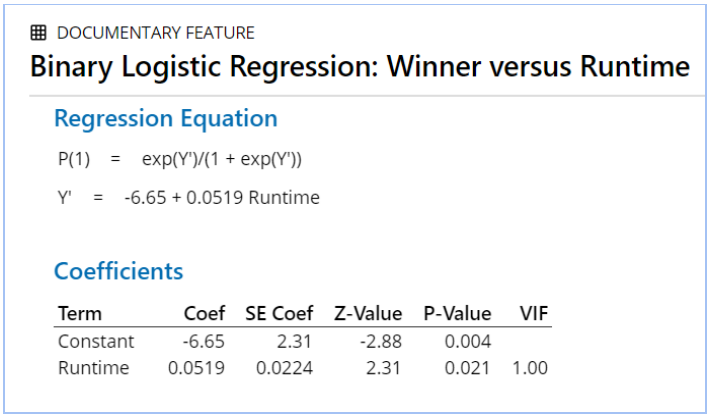
Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -6.57 + 0.0550 Runtime

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-6.57	2.27	-2.90	0.004	
Runtime	0.0550	0.0225	2.45	0.014	1.00



DOCUMENTARY FEATURE

Binary Logistic Regression: Winner versus Runtime

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -6.65 + 0.0519 Runtime

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-6.65	2.31	-2.88	0.004	
Runtime	0.0519	0.0224	2.31	0.021	1.00

DOCUMENTARY - SHORT

Binary Logistic Regression: Winner versus Runtime

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -3.58 + 0.0642 Runtime

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-3.58	1.10	-3.26	0.001	
Runtime	0.0642	0.0292	2.20	0.028	1.00

SHORT FILM (LIVE ACTION)

Binary Logistic Regression: Winner versus Runtime

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -2.878 + 0.0633 Runtime

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-2.878	0.775	-3.71	0.000	
Runtime	0.0633	0.0316	2.01	0.045	1.00

The regression equation is in the form $\hat{y} = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \ldots + \beta_n(x_n)$ where $\hat{y}$ is the odds of the response variable being 1, β_0 represents the odds of winning an Oscar when all of the predictors in the equation equal zero and β_n is the change in odds for each unit increase in the dependent variable x_n (Mertler & Vannatta, 2010). Therefore, in the Animation category, each additional minute of runtime increases the odds of winning by 5.50%, in the Documentary (Feature) category, each additional minute of runtime increases the odds of winning by 5.19%, in the Documentary (Short Subject) category, each additional minute of

runtime increases the odds of winning by 6.42% and finally in the Short (Live Action) category, each additional minute of runtime increases the odds of winning by 6.33%.

A summary of these results are shown in Table 4, which displays a breakdown of each category with the p-value and odds increase per minute.

Table 4

Binary Logistic Regression Model, categories with a p-value less than or equal to 0.05

Category	P-Value	Odds Increase per minute
Animation	0.014	5.50%
Documentary (Feature)	0.021	5.19%
Documentary (Short Subject)	0.028	6.42%
Short (Live Action)	0.045	6.33%

O. Is there a relationship between genre and winning in the top 6 categories?

To determine if there is a relationship between the type of genre of a film and whether or not it wins in Actor in a Leading Role, Actor in a Supporting Role, Actress in a Leading Role, Actress in a Supporting Role, Best Picture, and Directing, we use a binary logistic regression model with each subset of data separately. The individual data values for genre were put into 4 categories: Action, Comedy, Drama, and Other. These categories are broad enough to take into account the particular varieties that different films have in their genre. The distribution of genre in the films in these 6 categories are shown in Figure 16.

Figure 16

Distribution of genre in films nominated in Actor in a Leading Role, Actor in a Supporting Role, Actress in a Leading Role, Actress in a Supporting Role, Best Picture, and Directing

Tally

Genre -recoded	Count	Percent
Action	60	9.33
Comedy	71	11.04
Drama	435	67.65
Other	77	11.98
N=	643	

It is important to note the uneven distribution of the different categories. The data shows an overwhelming amount of films nominated for an Oscar between 2000-2019 being in the Drama genre. In the binary logistic regression model for this question, genre was not statistically significant for any of the categories tested. However, it is important to note that there were genres in each category that were comparably better than other genres in increasing the odds of winning in a category, even if they were not statistically significant in predicting a winner in the category.

One of the biggest differences we found were the odds ratios of the Actor in a Leading Role compared to the Actor in a Supporting Role. The odds ratios for the genre are shown in Figure 17.

Figure 17

Odds Ratios for Categorical Predictors, Actor in a Leading Role vs Actor in a Supporting Role

ACTOR LEAD

Binary Logistic Regression: Winner versus Genre -recoded

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Genre -recoded			
Comedy	Action	0.0000	(0.0000, 6.76495E+231)
Drama	Action	1.0818	(0.2051, 5.7063)
Other	Action	0.4375	(0.0323, 5.9257)
Drama	Comedy	654710.1748	(0.0000, 2.67513E+243)
Other	Comedy	264772.4972	(0.0000, 1.08584E+243)
Other	Drama	0.4044	(0.0472, 3.4677)

Odds ratio for level A relative to level B

ACTOR SUPP

Binary Logistic Regression: Winner versus Genre -recoded

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Genre -recoded			
Comedy	Action	2.1000	(0.2507, 17.5941)
Drama	Action	0.7903	(0.1480, 4.2197)
Other	Action	0.5833	(0.0418, 8.1459)
Drama	Comedy	0.3763	(0.0803, 1.7632)
Other	Comedy	0.2778	(0.0216, 3.5771)
Other	Drama	0.7381	(0.0822, 6.6281)

Odds ratio for level A relative to level B

The odds ratio by categorical predictors results show the comparative odds of winning in Level A versus Level B (Mertler & Vannatta, 2010). For example, in the Actor in a Leading Role category, the Drama (Level A) genre is 1.0818 times more likely to produce a winner than the Action (Level B) genre. When comparing the odds ratio by categorical predictors of the Actor in a Leading Role category to the Actor in a Supporting Role category, it is interesting to note the discrepancies between the two.

In the Actor in a Leading Role category, an actor is 1.0818 times more likely to win with a Drama than an Action movie, 654,710 times more likely to win with a Drama than a Comedy,

and 264,772 times more likely to win with an Other genre than with a Comedy. The discrepancy now shows up while looking at the Actor in a Supporting Role category, where an actor is 2.1 times more likely to win with a Comedy than an Action movie, while the rest of the ratios are less than one. Meaning, for a supporting actor the genre most likely to result in an Oscar would be in a Comedy film, though that is the worst genre for a leading actor's chances of winning an Oscar.

Discussion

The results of this study are consistent with the findings of previously conducted studies around the Oscars pertaining to actor and actress age comparisons. We found that, among actors and actresses that won or were nominated for an Oscar between the years of 2000 - 2019, actors are older on average than the actresses. When looking only at the subset of actors and actresses that were winners, the results were consistent and Oscar nominated actors were statistically older than Oscar nominated actresses on average between 2000-2019. Something unique to this study was the comparison of average age in Oscar nominated lead actresses to Oscar nominated supporting actresses, and the comparison of average age in Oscar nominated lead actors to Oscar nominated supporting actors. More simply, instead of testing between genders, we switched to testing within the gender group between their role type.

When comparing the average age of lead actresses that won or were nominated for an Oscar to the average age of supporting actresses that won or were nominated for an Oscar between the years of 2000-2019, this study concludes that there is no statistically significant difference. However, when the same question was tested with actors instead of actresses, the

result was statistically significant and we concluded that the average age of lead actors is different from the average age of supporting actors from 2000-2019.

Another factor we hypothesized would affect a nominee becoming a winner was a person's number of previous nominations. Although, the only category that returned with the number of previous nominations being a significant factor was Actress in a Supporting Role.

The last factor specific to individual actors and actresses that this study investigates is the relationship between race and winning an Oscar between the years 2000-2019. First, we determined that race is not a contributing factor in predicting whether or not a person will win an Oscar. After some deeper understanding, we attribute the lack of significance in race predicting a winner in the Oscars to being disproportionately represented in the nomination pool overall. Meaning, race does not affect an actor or actress's chance of winning, but there is no data to support whether or not race affects an actor or actress from being nominated in the first place. Therefore, this study investigated the proportions of the distribution of race groups in the Oscar nominations, and compared it to the distribution of race groups of actors in America through the use of the Screen Actors Guild (SAG-AFTRA, n.d.) population data and proportions of race described in the UCLA Hollywood Diversity Report (Hunt et al., 2019, 14).

We found that there is a difference in the proportions of white actors and actresses and actors and actresses of color being nominated or winning in the Oscars relative to the SAG population. Therefore, people of color are not getting nominated proportionately to their population in the SAG, and that could be affecting the result of predicting a winner in different categories based on race.

Aside from the actor and actress categories, this study also investigates if team size is a significant indicator of winning an Oscar between 2000-2019. After running the binary logistic regression model, the results show that the number of people on a team is not effective when trying to predict the odds of winning in Documentary (Feature), Documentary (Short Subject), Film Editing, Sound, Sound Editing, or Sound Mixing.

Another relationship this study explored was that of runtime and winning in each category. We discovered that in Animation, Documentary (Feature), Documentary (Short Subject), and Short (Live Action) categories, each additional minute of runtime increases the odds of winning an Oscar based on our data from 2000-2019. This gives producers an edge when planning a movie, as they can make their movie longer and be more likely to win an Oscar.

Finally, while looking into the relationship of genre and winning, we ran into issues similar to the issues with investigating race in the Oscars. The issue being that the distribution of the genres was heavily weighted in the Drama category, and all the other categories had miniscule amounts in comparison. We attempted to adjust that by creating a combination category for categories that had small amounts of nominations, just as we did for the people of color in race. Though the test resulted in genre not being an effective predicting variable for winning an Oscar, we did find some interesting odds comparisons for different categories. Mainly, how a lead actor and a supporting actor have drastically different odds of winning when comparing the genres. We found that a lead actor has the best odds of winning while in a Drama movie, and a supporting actor has the best odds of winning while in a Comedy movie.

Conclusion

With all of these factors considered, the statistical testing results lead to each category having different criteria for predicting a winner. There is no cookie-cutter answer for the perfect formula for a movie to win an Oscar, because there are different factors in each category that boost the odds of winning.

Opportunity for future study

There are many topics surrounding the Oscars and different award shows that have yet to be studied. To build off of this research, a data set that includes all movies and actors eligible for an Oscar in the years studied would assist in questions that compare the winners to the general population of movies. For example, questions regarding race and genre in this study would have been able to answered more concretely with data on all Oscar eligible entities.

Another opportunity for future study would be for a comparison of time blocks in the Oscars. This study specifically focused on 2000-2019, but a study that would be able to compare each 20 years of the Oscars and find differences in the trends and predictors would be an amazing opportunity to show how the award show has progressed.

Finally, a similar study could be done with data on different award shows, such as the Grammys or the Golden Globes. All of these award shows in the entertainment industry seem to have traceable trends, and it would be interesting to see how they compare with one another.

Acknowledgements

I would like to thank some professors at Florida Southern College for assisting me in completing my honors thesis. Specifically, I would like to thank Dr. Eicholtz for assisting me early in the process in finding a portion of my data, as well as Dr. Jelsovsky for his expertise and

assistance in the categorical statistical modeling processes. I would especially like to thank Dr. Susan Serrano for all of her dedication to this project, and for being an excellent mentor throughout my college career.

References:

Britannica, T. Editors of Encyclopaedia (2020, August 7). Academy Award. Encyclopedia Britannica. https://www.britannica.com/art/Academy-Award

Hunt, D. D., Ramon, D. A.-C., & Tran, M. (2019). *Hollywood Diversity Report*. UCLA College of Social Sciences. https://socialsciences.ucla.edu/wp-content/uploads/2019/02/UCLA-Hollywood-Diversity-Report-2019-2-21-2019.pdf

Kuiper, S., & Sklar, J. (2013). *Practicing Statistics: Guided Investigations for the Second Course.* Pearson.

Leonardo DiCaprio - Awards. (n.d.). IMDb. https://www.imdb.com/name/nm0000138/awards?ref_=nm_ql_2

Mertler, C. A., & Vannatta, R. A. (2010). *Advanced and Multivariate Statistical Methods* (4th ed.). Pyrczak Publishing.

Pardoe, I., & Simonton, D. K. (2008). Applying discrete choice models to predict Academy Award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society),171*(2), 375-394. doi:10.1111/j.1467-985x.2007.00518.x

Pollock, R. (2018). *Academy Awards Oscars: Nominees and Winners 1927 to Present*. Datahub. https://datahub.io/rufuspollock/oscars-nominees-and-winners#r

SAG-AFTRA. (n.d.). *About the Screen Actors Guild*. SAG-AFTRA. Retrieved 2021, from https://www.sagaftra.org/about

Simonoff, J. S., & Sparrow, I. R. (2000). Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. *CHANCE*, *13*(3), 15-24. 10.1080/09332480.2000.10542216

Simonton, D. K. (2002). Collaborative Aesthetics in the Feature Film: Cinematic Components Predicting the Differential Impact of 2,323 Oscar-Nominated Movies. *Empirical Studies of the Arts,20*(2), 115-125. doi:10.2190/rhq2-9uc3-6t32-hr66

Simonton, D. K. (2004). Film Awards as Indicators of Cinematic Creativity and Achievement: A Quantitative Comparison of the Oscars and Six Alternatives. *Creativity Research Journal,16*(2), 163-172. doi:10.1207/s15326934crj1602&3_2

Simonton, D. K. (2004). The "Best Actress" Paradox: Outstanding Feature Films Versus Exceptional Womens Performances. *Sex Roles,50*(11/12), 781-794. doi:10.1023/b:sers.0000029097.98802.2c

Literature Review References, not cited in paper:

Collins, A., & Hand, C. (2006). Vote Clustering in Tournaments: What Can Oscar Tell Us? *Creativity Research Journal,18*(4), 427-434. doi:10.1207/s15326934crj1804_2

Eliashberg, J., & Shugan, S. M. (1997). Film Critics: Influencers or Predictors? *Journal of Marketing,61*(2), 68-78. doi:10.1177/002224299706100205

Gilberg, M. (2000). Male Entertainment Award Winners Are Older Than Female Winners. *Psychological Reports,86*(1), 175. doi:10.2466/pr0.86.1.175-178

Hedley, M. (2002). The Geometry of Gendered Conflict in Popular Film: 1986-2000. *Sex Roles,47*, 201-217.

"Operations Research Analysts : Occupational Outlook Handbook." *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 4 Sept. 2019, www.bls.gov/ooh/math/operations-research-analysts.htm.

Pardoe, I. (2005). Just How Predictable Are the Oscars? *Chance,18*(4), 32-39. doi:10.1080/09332480.2005.10722753