**Correlation Between Movie Success and Various Factors**

This study explores a fascinating dataset of different movies and analyzes the factors that make them successful. These components include the impact of the director, the year of release, and particular financial indicators like box office revenue. The goal of this study is to find underlying patterns and connections that could provide insight into the complex nature of cinematic success by using statistical studies and probability theory.

The film industry is a wide and fascinating one, with many different aspects influencing a movie's potential for success. With an emphasis on features including director, release year, length, genre, rating, metascore, and gross profits, this study examines a dataset of highly rated films. The goal is to examine these elements and understand how they affect a film's critical and financial success.

Directors are crucial in determining the direction and style of a movie. The dataset shows that there is a tendency for well-known directors' films to earn a lot of money at the box office and positive reviews. For example, films helmed by film legends such as Steven Spielberg and Christopher Nolan not only do extraordinarily well in terms of box office receipts, but they also garner praise and higher metascores. This connection implies that a director's popularity and style have a big impact on a movie's box office and critical reception.

A film's year of release can have a significant impact on its commercial success. Due to greater audience availability, movies released during specific times of the year, such as the summer or the holidays, frequently see higher box office returns. Historical background and technological developments also come into play. The reception and financial success of films released in the early 2000s, for example, may be driven by the significant developments in digital cinematography during that time.

The fundamental criteria to evaluate a film's commercial success is its box office gross. Films with strong production quality, broad appeal to audiences, and significant marketing campaigns are frequently the highest grossing titles in the dataset, such "The Dark Knight" and "Schindler's List." The genre has an impact on box office results as well, both being action and adventure films, for example, which typically draw larger audiences and earn more money.

While box office revenue reveals a film's ability to make money. Reviews and metascores reveal a film's level of artistic and technical quality. The dataset shows how these two aspects interact in an intriguing way. While some movies succeed in both the box office and the critics' eyes, others could be strong in just one. The discrepancy emphasizes how the film business sometimes struggles to maintain artistic excellence with profitability.

The success of a film is determined by several important aspects. As was previously mentioned, directing ability plays a big role. Another factor is genre, such as action and drama, for example, which appeal to a wider audience. Both the production budget and the celebrity cast play a role. Larger expenses and more well-known performers typically translate into better box office results.

In conclusion, a film's success is an unpredictable occurrence that depends on a number of factors, including the director's style, the genre, the timing of its release, and the interaction between its box office and critical reception. This study shows that, despite some clear patterns and tendencies, the film business is still a dynamic, complex field where commercial and creative success are not always guaranteed. Comprehending these subtleties highlights the complexity of factors that defines a successful movie production and offers insightful information to reviewers, spectators, and filmmakers.

**Questions**

**Chapter 2:**

1. In the given dataset, let's consider `Drama` as the genre of interest for the first week of movies. If `D` denotes drama movies released in the year 1994, `A` denotes action movies released in the year 2008, and `B` denotes biography movies released in the year 1993, construct a set `U` representing this scenario. What is the set when only `D` happens? What is the set when both `D` and `B` happen?

2. Assuming we are looking at the release years of movies in the second week, let `` be the event of a movie released in 1994, `` be the event of a movie released in 1972, `` be the event of a movie released in 2008, and `` be the event of a movie released in 1957. If ``, what is the probability of a movie from 1957 being a drama? Now, what will happen if we exchange the probabilities of ``, ``, and ``?

3. Continuing with the second week of movies, suppose there were ten movies in the dataset from 1994. Out of those, three movies have a metascore of more than 80. Suppose that three movies are surveyed to see which have a metascore of exactly 82.

- List the sample space.

- Identify the simple events in each of the following events:

- `A` = At least two movies have a metascore of 82.

- `B` = Exactly two movies have a metascore of 82.

- `C` = Exactly one movie has a metascore less than 82.

4. An experiment consists of looking at the movies from two different years, where each year has ten movies in the dataset.

- Use combinatorial theorems to determine the number of sample points in the sample space `S`.

- Find the probability that the sum of the ratings of movies from 1994 and 2008 is more than 18.

5. Suppose that we ask `n` randomly selected movies whether they are directed by Christopher Nolan.

- Give an expression for the probability that no movies share the same director. (Total movies directed by Christopher Nolan are 1)

- How many movies do we need to select so that the probability is at least 0.5 that at least one movie is directed by Christopher Nolan?

6. For a certain selection of movies, the percentage of movies with a rating greater than 9 was as shown in the accompanying table. That is, of all the movies listed, 47% had a rating greater than 9, 20% had a rating between 8 and 9, and 33% had a rating below 8. A movie is to be selected randomly from this population. Let `A` be the event that the movie has a rating greater than 9 and let `M` be the event that the movie has a rating below 8.

- Are the events `M` and `A` independent?

7. Looking at a random selection of movies from the dataset, assume there are events `A` and `B`, where `A` is a movie having a metascore above 90 and `B` is a movie grossing more than $100M. If the probability that a movie has a metascore above 90 is `A = 0.47`, the probability that a movie is grossing more than $100M is `B = 0.29`, and we know that both events cannot occur at the same time, find the following:

- When a movie is selected, find the probability that either it has a metascore above 90 or it is grossing more than $100M.

- Find the probability that neither of the events will occur.

8. A population of movies contains 40% with a rating above 9 and 60% with a rating below 9. It is reported that 30% of the high-rated movies and 70% of the lower-rated movies are dramas. A movie chosen at random from this population is found to be a drama. Find the conditional probability that the movie has a rating above 9.

**Chapter 3:**

1. You and a friend are guessing the metascore for two randomly selected movies. If both movies have metascores above 90, you win $1; if both have metascores below 90, you win $2; if one movie has a metascore above 90 and the other below, you lose $1. Assuming that no movie has a metascore exactly 90, give the probability distribution for your winnings, Y, on a single play of this game.

2. A film critic wants to analyze the movies from two different eras where era 1 includes movies released before 2000 (A), and era 2 includes movies released after 2000 (B). The critic asks four interns to help with analyzing the movies. Each intern is randomly given three movies, two from A and one from B. Each intern is asked to predict the era each movie was released. Suppose that the interns have equal odds of guessing the correct era for each movie. Let Y be the number of interns predicting the era correctly for all three movies.

- Find the probability function for Y.

- What is the probability that at least three of the four interns predict all eras correctly?

- Find the expected value of Y.

- Find the variance of Y.

3. Suppose that 30% of the movies in the dataset are directed by Steven Spielberg. Movies are selected at random, and each selection is independent of each other. Find the probability that the first movie is part of one of five Spielberg-directed movies in the dataset.

4. Refer to Problem 3. What is the expected number of movies you need to select until you find the first Spielberg-directed movie.

5. An analyst is given a list of 20 movies, of which ten have a rating above 9, three have a rating between 8 and 9, and two have a rating below 8. Six movies are to be selected from the list, one at a time without replacement. What is the probability that all six movies selected have a rating above 9?

6. Refer to Problem 5. The analyst wants to know the mean and variance of the number of high-rated movies (rating above 9) in a sample of six movies.

7. From a random selection of movies, the probability that a movie has a metascore above 90 follows a Poisson distribution with an average of 0.35 wins per movie. During a random selection of movies, what are the probabilities that:

- No more than four movies have a metascore above 90?

- At least three movies have a metascore above 90?

- Exactly two movies have a metascore above 90?

8. Would you rather bet on a movie directed by Christopher Nolan or a movie with a runtime longer than 150 minutes? If you bet on a Nolan movie, there is a 29% chance it's in the dataset; if you bet on the runtime, there is a 24% chance. There are a total of 50 movies in the dataset. Find the answer if you bet on the runtime.

- What is the expected value of the number Y of movies that you bet on that you will win?

- Find the standard deviation of Y.

- Calculate the intervals μ±2σ and μ±3σ.

**Chapter 4:**

**Problem 1:**

A movie's rating contributes to its overall success points. Let's say a rating of 8.5 and above gives a movie three success points, a rating between 7.0 and 8.4 gives it one point, and a rating below 7.0 gives zero points. The total amount of points gained from a movie's rating is a random variable *Y* with a probability density function given by:

-

- the cumulative distribution function, and graph it.

- Find the probability that a movie will gain between 0 to 1 point from its rating.

- Find the probability that a movie will gain more than 1 point from its rating.

**Problem 2:**

After analyzing the metascores of movies from the dataset, assume the metascores are uniformly distributed between 50 to 100. Find the probability that the metascores for a movie selected at random:

- Are below 75.

- Are above 85.

**Problem 3:**

Refer to Problem 2. Find the expected value of metascores of movies if they are uniformly distributed between 50 and 100.

**Problem 4:**

If the success points Y for a movie's rating have a density function as described in Problem 1:

- Find the mean and variance of *Y*

**Chapter 5:**

**Problem 1:**

For a certain group of films, let ​ denote the proportion of films that received a critic rating above 85%, and ​ denote the proportion of films that received an audience score above 85%. We want to understand how often critics and audiences agree on the high quality of a film.

- Construct a joint probability table for ​ and ​ given a sample of 10 films.

- Calculate the probability that at least half of the films are highly rated by both critics and audiences.

**Problem 2:**

Assume a director's success is judged by box office earnings. We categorize a film as a huge success (A) if it earns more than $250 million, a moderate success (B) if it earns between $100 million and $250 million, and not successful (C) if it earns less than $100 million. Let ​ be the number of huge successes in the director's career of 15 films.

- Find the probability distribution for ​

- If three films are randomly selected from the director's portfolio, find the probability that two are huge successes.

**Problem 3:**

Consider a dataset of 20 movies where each movie has a rating. Let ​ denote the proportion of movies rated above 8.0, and ​ denote the proportion of movies rated between 6.0 and 8.0. Assume that no movie is rated below 6.0.

- Determine a probability function that models the behavior of ​ and ​ given their continuous nature.

- Find the probability that at least 30% of the movies are rated above 8.0 but less than 60% are rated between 6.0 and 8.0.

**Answer Key**

**Chapter 2:**

**Problem 1:**

- Given set `U = {D, A, B}` where `D` represents drama movies from 1994, `A` represents action movies from 2008, and `B` represents biography movies from 1993.

- When only `D` happens, the set `U` would only include drama movies from 1994. Since there's only one drama movie from 1994 in the dataset (The Shawshank Redemption), the set would be `{D}`.

- When both `D` and `B` happen, the set `U` would include drama movies from 1994 and biography movies from 1993. In the dataset, Schindler's List is a biography movie from 1993. Thus, the set would be `{D, B}`.

**Problem 2:**

- We have the probabilities ``, and ``.

- The probability of a movie from 1957 being a drama, ``, could be assessed by looking at the given dataset. Since 12 Angry Men is the only movie from 1957 and it is a drama, the probability is 1 (or 100%).

- If we exchange the probabilities of ``, ``, and ``, we would have ``, ``, and ``. The probability of a movie from 1957 being a drama would remain 1 because the outcome is certain regardless of the probability associated with the year.

**Problem 3:**

- The sample space for movies from 1994 with a metascore of more than 80 would include The Shawshank Redemption (metascore 82).

- The simple events would be:

- `A` = {SS}, where `SS` represents The Shawshank Redemption.

- `B` = There are no such events since there's only one movie from 1994 with a metascore of 82.

- `C` = There are no such events since there's only one movie from 1994 and it has a metascore of 82.

**Problem 4:**

- Using combinatorial theorems, the number of sample points in the sample space `S` (for movies from two different years with ten movies each) would be `C(20, 10)` because we are choosing 10 movies from a combined set of 20.

- The probability that the sum of the ratings of movies from 1994 and 2008 is more than 18 is certain (1 or 100%) because the ratings of The Shawshank Redemption and The Dark Knight are 9.3 and 9.0, respectively, summing to more than 18.

**Problem 5:**

- The probability that no movies share the same director is the probability that none of the selected movies is directed by Christopher Nolan. Since there's only one movie by Christopher Nolan, as long as we select any other movie, the probability is 1.

- To have at least a 0.5 probability that one movie is directed by Christopher Nolan, we could use the complement rule. However, since there's only one movie by Nolan, selecting any one movie would give us a probability of 0.5.

**Problem 6:**

- To determine if the events `M` and `A` are independent, we need to check if ``. Since these are mutually exclusive events (a movie cannot both have a rating above 9 and below 8), ``. Therefore, they are not independent.

**Problem 7:**

- The probability that either the home team wins or the away team wins is ``. Since `A` and `B` cannot both happen (mutually exclusive), `P(A ∪ B) = P(A) + P(B) = 0.47 + 0.29 = 0.76`.

- The probability that neither event will occur is the complement of the union of `A` and `B`, which is ``.

**Problem 8:**

- Let `H` be the event of a home win (40%), and `A` be the event of an away win (60%). Given that 30% of `H` and 70% of `A` are easy wins, the probability that a randomly selected easy win is an away win is ``. We don't have enough information to find ``, so we can't calculate this probability without additional data.

**Chapter 3:**

**Problem 1:**

To determine the probability distribution for your winnings, Y, we need to consider all possible outcomes of the two movies' metascores.

Given:  
**Solution:**

Let's denote the event of a movie having a metascore above 90 as A and below 90 as B. We need to find P(AA), P(BB), and P(AB∪BA).

If we assume the movies are equally likely to have metascores above or below 90, and if 30% of movies have metascores above 90 (based on an assumption since exact data is not provided), then:

The events are independent, hence:

**Problem 2:**

**Solution:**

The probability that an intern guesses all three eras correctly is the probability of guessing one era correctly raised to the power of three. If we assume a 50% chance of guessing correctly (since there are two eras), then:

The probability function for Y, where Y is the number of interns guessing all movies correctly, follows a binomial distribution with parameters n=4 (number of trials) and p=0.125 (success probability).

**Problem 3:**

30% of movies are directed by Spielberg.

**Solution:**

**Problem 4:**

**Solution:**

This is a geometric distribution problem where we are looking for the expected number of trials until the first success (selecting a Spielberg movie). The expected number of trials, E(X), in a geometric distribution is:

**Problem 5:**

Given:  
10 high-rated, 3 mid-rated, 2 low-rated movies.

Solution:

We are choosing 6 movies without replacement from a pool of 20 movies.

**Problem 6:**

**Solution:**

For a hypergeometric distribution, the mean is ​ and the variance is:

where *n* is the sample size, *K* is the number of success states in the population, and *N* is the population size.

**Problem 7:**Given:

Average of 0.35 away team wins per game.

Solution: For a Poisson distribution with an average rate of λ=0.35 wins per game, the probability of exactly *k* events occurring is given by:

You can then calculate the probabilities for accordingly.

**Problem 8:**Given:

29% chance for Nolan movie, 24% for runtime > 150 mins.

**Solution:**

You should bet on the event with the higher probability if you have no other information. In this case, betting on a Nolan movie is better with a 29% chance of winning compared to a 24% chance for a long runtime movie.

For the expected value, standard deviation, and intervals, you would use the binomial distribution formulas for *n* = 296 games, and *p* = 0.29 for Nolan or *p* = 0.24 for runtime.

Expected value

Variance

Standard deviation

Intervals

**Chapter 4:**

**Problem 1:**

Cumulative distribution function (CDF):

Probability of gaining between 0 to 1 point:

**Problem 2:**

Given:

Uniform distribution between 50 to 100.

Probability of metascore below 75:

Since it's a uniform distribution,

Probability of metascore above 85:

**Problem 3:**

Given:

Uniform distribution between 50 to 100.

Expected value

**Problem 4:**

Mean �(�)E(Y):

To find the mean, you integrate

Variance To find the variance, you first need the mean (calculated above) and then integrate

Let's calculate the mean and variance using the given PDF.

For the mean

For the variance

**Chapter 5:**

**Problem 1:**  
Given:

A director's career of 15 films.

​ is the number of films earning over $250 million.

To solve:

We find the probability distribution for ​

**Solution:**

The probability that a given film earns over $250 million will depend on industry averages which we do not have. However, for the sake of this problem, let's assume this probability is *p*. The number of successes in 15 films, ​ ​, would follow a binomial distribution.

The probability distribution for ​ ​ is

where *k* is the number of films earning over $250 million.

For the second part: If *p* is the probability of a film being a huge success, then the probability that exactly two out of three randomly selected films are huge successes is given by the binomial probability formula:

**Problem 2:**

Given:

A sample of 10 films.

​ ​ is the proportion of films with a critic rating above 85%.

​ ​ is the proportion of films with an audience score above 85%.

To solve:

We need to create a joint probability table for ​ ​and ​ ​.

Then we calculate the probability that at least half of the films are highly rated by both critics and audiences.

Solution: Since we do not have the specific proportions, we can consider a hypothetical continuous distribution where the density function of the joint probability reflects the likelihood of films falling into these categories.

For the integration part, let's assume the density function for ​ ​ and ​ is uniform over the interval [0,1] for each variable:

The probability that at least half of the films are highly rated by both is given by the integral over the appropriate domain:

**Problem 3:**

Given:

A dataset of 20 movies.

​ is the proportion of movies rated above 8.0.

​ ​ is the proportion of movies rated between 6.0 and 8.0.

To solve:

We determine a probability function for and ​

We find the probability that at least 30% are rated above 8.0 but less than 60% are rated between 6.0 and 8.0.

**Solution:**

For continuous random variables, the joint probability function is often given by a joint probability density function (PDF). Let's assume is a bivariate uniform distribution.

The probability that at least 30% are rated above 8.0 but less than 60% are rated between 6.0 and 8.0 is:

If the density function is uniform and the variables are independent, then can be taken as a constant *k* where:

Given that (no overlap in rating categories), we would integrate over a triangle in the unit square with vertices at (0,0), (1,0), and (0,1). The area of this triangle is 1/2, so

.

The final integral to solve for the given probability is: