

# SI 544 Stats Final, Winter 2008

Instructor: Lada Adamic

*This is your final exam. Turn in all R code, label your plots, and put in a sentence or two highlighting the results of each step of your analysis. Mention where appropriate things like p-values, what your null-hypothesis is, and whether you reject it or accept it. For each statistical method you choose, briefly justify your selection and comment on whether the assumptions needed in order to be able to use the method are met. This exam is open book/lecture notes/solution sets/web. You are to work on this without the assistance of others. Good luck!*

## 1 Movie genres in Asia

You've obtained some statistics on the number of movies that fall in one of four genres for four different countries. Some of these numbers correspond with your intuition about which types of movies are popular in each country. For example, you recall watching many good Hong Kong action flicks. You can load the data into R using the command

```
t = as.matrix(read.table("MovieGenresInAsia.txt",head=T,row.names=1))
```

where `t` is just a variable name that you can substitute with your favorite.

	India	Japan	China	Hong.Kong
Action	20	42	13	33
Comedy	32	28	5	13
Drama	66	65	24	34
Romance	40	16	9	11

You would like to know whether the proportion of movies made in each genre depends on the country.

- (10pts) Display these differences visually (plot the data!)
- (10pts) Test whether the proportion of movies in each genre does depend on the country where the movies are made.
- (5pts) Calculate the differences between the observed and expected number of movies in each table cell. Discuss your observations briefly.

## 2 Does the rating of a movie depend on its genre and the country it was made in?

You decide to sample at random 40 movies for each of 3 different genres (Drama, Action, Comedy), for 3 different countries and special administrative regions: USA, UK, Hong Kong. The data is limited to movies made since 1990 for which there were at least 500 **imdb** (Internet Movie Database) ratings. You can read in the file using the command:

```
movies = read.table("MoviesCountryGenre.txt",head=T,sep='\t')
```

- A. (10pts) You would like to visualize the data, to see if the mean rating depends on either genre or country, and whether there is interaction between the variables.
- B. (10pts) Now you would like to see if any of the differences by country, genre, or an interaction between the two variables are statistically significant. Present your analysis and discuss briefly.

## 3 Box office and ratings

Next you'd like to know whether spending more on a movie will be rewarded in terms of user ratings and box office income. You have gathered a list of movies made since 1990 where both the budget and US box office results are given in dollars, and where the number of **imdb** user ratings exceeds 1000. You can read the file using

```
mf = read.table("BoxBudgetRating.txt",head=T).
```

The file `BoxBudgetRating.txt` has 5 columns (`title`, `year`, `budget`, `boxoffice`, `avrating`), each row corresponding to one movie.

- A. (10 pts) What is the correlation between amount of money spent on a movie and box office revenues?
- B. (15pts) Perform a linear regression, and estimate the number of extra dollars made at the boxoffice, for each additional dollar spent in making the movie. Plot the data and your linear regression line. Was the use of a simple linear regression appropriate? Why or why not?
- C. (10 pts) Next you'd like to construct a multiple regression model that would use both the movie budget and average movie ratings to explain the box office outcome (for simplicity, do not include an interaction term). The ratings are given on a scale of 1 to 10. For each additional point increase in ratings (e.g.

from 6.3 to 7.3) how much do the box office revenues increase?

D. (10pts) Your friend claims that they don't make movies like they used to. Your friend is not all that old, so for him, old movies are movies made in the 1990s. You'd like to know whether the data support his claim, that is whether movies made in the 1990s receive higher ratings on average than movies made over the last 8 years.

E. (10pts) If the actual difference in means ratings were 0.1, and assuming that the sample standard deviation in movie ratings you measured is equal to the population standard deviation, how many movies would you need in each group in order to be have 80% power in rejecting the hypothesis that the means are equal? Assume your significance criterion is  $\alpha = 0.01$ .