University of California Employee Pay

Daniel Jaso, Kian Sheik, Haiyu Liu
University of San Francisco
MATH 371 - Statistics with Applications

May 10th, 2018

Contents

1 Introduction

The data used in the following research was obtained from the University of California. The UC releases annual compensation data for each employee. This data can be obtained publicly via `https://ucannualwage.ucop.edu/wage/`. The data, at the time this was written, spans the years 2010-2016 for each campus. We will focus on UCLA in the year of 2015, as this data was most plentiful.

UCLA states that, with 42,000 full- and part-time jobs, UCLA is among the top five regional employers. In total, there were 3,766 professors employed in the year of 2015. These titles are quite varied within the dataset, so we used external directories to organize them into separate departments.

We aim to find an expected salary for Psychology and Math Professors using these directories. Once we find these values, our goal is to use them to create a confidence interval for the difference in pay among these departments. This will give us insight as to the disparagement of pay among departments.

2 Extraction of Data

We used the Wayback Machine `http://archive.org/web/` provided by the Internet Archive to access a 2015 versions of the the UCLA department websites of the Psychology `https://www.psych.ucla.edu/faculty` and Mathematics `https://www.math.ucla.edu/people`, respectively. By cross-referencing the directories of each department with the data provided by UC Annual Wage, we were able to compile statistics on each department to gain insights on which department has a higher expected salary.

Using the network tab of the Google Chrome Developer tools, one can see all of the network traffic moving to and from any given webpage. When you query a search while this tab is opened, you are given the URL `https://ucannualwage.ucop.edu/wage/search.action`. This URL delivers the data in JSON, which is much more digestible than the HTML table initially provided to us. By default, the number of rows this returns is set to 20. We can edit this in the query string to be 9999999. After some processing to

transform the data into CSV, we were able to import it into R for exploration. Here is the full url for the purpose of downloading reproducible data congruent with this paper.:
`https://ucannualwage.ucop.edu/wage/search.action?rows=9999999&page=1&sidx=EAW_LST_NAM&sord=asc&year=2015&location=Los+Angeles&firstname=&lastname=&title=&startSal=0&endSal=9999999`

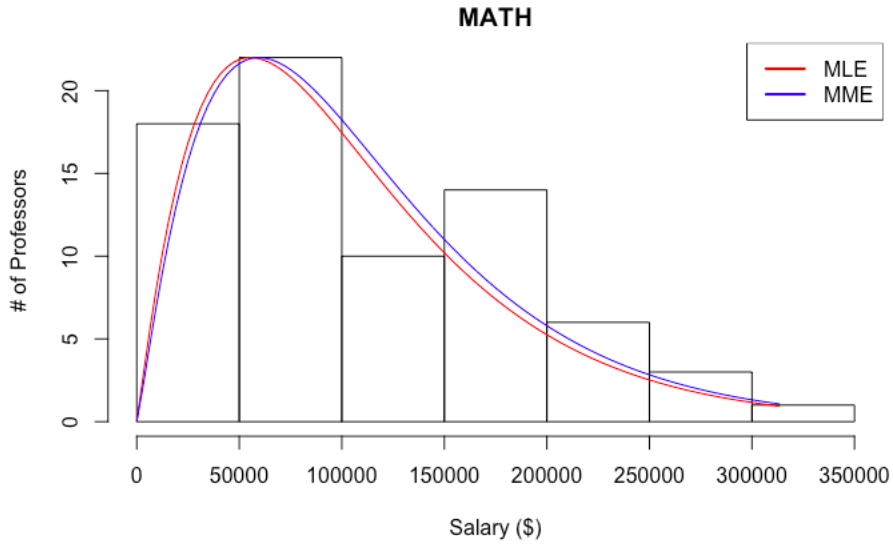
3 Method of Moments Estimators (MME)

We first graphed the histogram of each department to decide which distribution to fit the data to. We hypothesize that each department follows: $X_i \sim GAM(\theta, \kappa)$.

MME of θ, κ:

$$\hat{\theta} = \frac{(n-1)S^2}{n\bar{X}}$$

$$\hat{\kappa} = \frac{n\bar{X}^2}{(n-1)S^2}$$

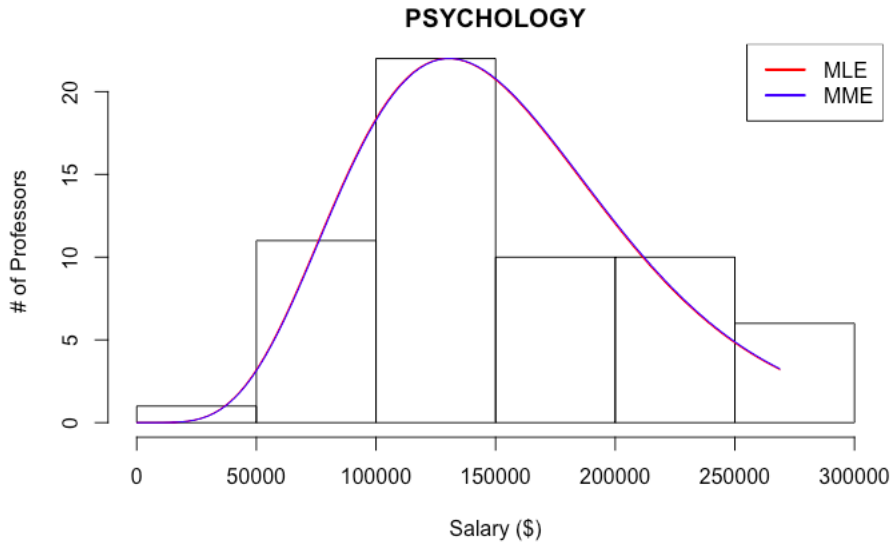


The distribution of salary amongst math professors. The raw MME cal-

culations for this distribution can be seen below. $\bar{X} = \$110,755.30$ and $s = \$75,872.87$

$$\hat{\theta} = 51274.291 = \frac{(74-1)5756692540}{74 * 110755.3}$$

$$\hat{\kappa} = 2.160 = \frac{74 * 110755.3^2}{(74-1)5756692540}$$



The distribution of salary amongst psych professors. The raw MME calculations for this distribution can be seen below. $\bar{X} = \$153,643.20$ and $s = \$59,926.18$

$$\hat{\theta} = 22983.732 = \frac{(60-1)3591147461}{60 * 153643.2}$$

$$\hat{\kappa} = 6.684 = \frac{60 * 153643.2^2}{(60-1)3591147461}$$

4 Maximum Likelihood Estimator (MLE)

Using the MME values from above as starting points, we compute the numerical MLE using the mle function from the stats4 package. We found that

the difference between our MME and MLE are negligible; the graphs of the MLE are also shown above in the previous figures.

5 Hypothesis Testing with Cross Validation

In order to test whether our MLE Estimators are accurate, we used cross-validation. This means that we divided our populations into two random samples: 75% training-data, 25% testing-data. This allows us to see whether our estimators are victims of overfitting or if they are in fact valid estimators. We used hypothesis testing against the test samples to validate our estimators.

H_0: $\mu = \mu_0 = \hat{\kappa}\hat{\theta}$
H_a: $\mu \neq \mu_0 \neq \hat{\kappa}\hat{\theta}$

Critical region:

$$R = \left\{ x : \frac{\bar{X} - \mu_0}{\sqrt{\frac{\kappa\theta^2}{n}}} \geq Z_{0.95} \right\}$$

Based on our hypothesis testing, we found that the data for the Psychology department follows a distribution of GAM($\theta = 22983.732$, $\kappa = 6.662$) with 95% confidence. The Mathematics department, however, did not pass our hypothesis testing given the best-fit MLE estimators we found. With a 5% significance level, we reject that the distribution of the salaries the among Math department follow GAM($\theta = 51274.291$, $\kappa = 2.08$).

6 Confidence Interval

Given two random samples X, Y:

$$T = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}} \sim t(v) \tag{11.5.13}$$

With degree of freedom:

$$v = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{{n_1}^{-1}} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{{n_2}^{-1}}} \tag{11.5.14}$$

We calculated the difference between the Math and Psychology department's expected salary. With a 5% significance level, we found that a Psychology professor can expect to be paid from $10,203.72 to $53,892.24 more than a Math professor each year.

7 Conclusion

Based on our findings, we can say with a 95% confidence interval that the expected salary of a Math professor at UCLA in 2015 was between $93,468.33 and $128,042.26. The Psychology professors, on the other hand, can expect to receive between $138,480.10 and $168,806.40. We were unable to fit the Math department's pay distribution to a Gamma distribution, however we were able to fit the Psychology department to a Gamma distribution with a 95% confidence interval.

Aside from these findings, we were able to say with a 5% significance level that a Psychology professor can expect to be paid from $10,203.72 to $53,892.24 more than a Math professor each year. These results were surprising to us as we were expecting the math professors to have a higher average salary.

The biggest challenge in this project was acquiring, cleaning and grouping the data so that we may make meaningful inferences. The initial data was poorly labeled and the provider of the data offered no insights as to what the shorthand names of the Job Title column referred to. The data required external forensics to be grouped at a department level. Given these constraints, less of our time was spent on the actual analysis as we would have liked. Given more time or more complete data, we would have liked to analyze the expected salary amongst all departments.

This project showed us the the real-world burdens of data scientists and statisticians. Although there are only a few confidence intervals in this document, a lot of work went into producing the data that generated them. The skills we learned and displayed in order to compute the confidence intervals shown in this document are an example of the fact that the data provided will not always be perfect; if you wish to make novel discoveries, you must be capable of shaping the data to properly answer the questions being asked.