Iowa State Salary Book Dissected Project Report

Wendy Zheng, Alex Scarlatos, Shengchun Liu, Liam Wang, and Nicholas Cataldo

I. INTRODUCTION

Salary data is an excellent indicator of high-level trends. An individual's salary is affected by a wide variety of factors, ranging from their gender to their level of education to their employment history. While there is high variability from person to person in their salary, when observing large sets of data, we can draw general patterns. These patterns can be very useful, since they can help identify social inequality, or factors that have a negative impact on a population. By understanding what trends exist in salary data, we can begin to address bigger questions, such as how to tackle poverty, and how to alleviate economic downturn.

We selected the Iowa state salary book for several reasons. It was a good dataset since it had entries for individuals rather than precomputed averages, so we could examine the data very closely. In addition to their income, a lot of information was given on each individual, including their gender, and what job they had and where they lived for different years. Iowa is a fairly standard U.S. state, with areas ranging from very rural to urban. It was also useful to only examine one state at a time to avoid the need to account for independent events happening across the country.

In this paper, we examine a variety of attributes for individuals, and see how those attributes can be used to predict different trends. Specifically, we will focus on the effects of a person's gender, a person's area of residence, the rate at which a person changes jobs, a person's place of employment, and national economic influences. We use a variety of statistical techniques to investigate these topics, including linear regression, mean comparison and distribution comparison to derive several statistically significant results.

We find that men on average make more than women and have higher career mobility, but that women on average work full-time more than men. We find that a high level of education in a person's place of residence has a positive impact on their salary, and that several economic and health factors have a significant impact as well. We find that higher job mobility has a positive impact on salary, and that salary distributions vary greatly between different departments. Finally, we find that inflation rate and federal funds are good predictors for overall average income, and that general economic health varied greatly following the Great Recession.

II. DATASETS

A. Iowa State Salary Book

The Iowa state salary book contains approximately 660,000 rows of data, containing salary data from 2007 to 2017. Each row in the Iowa state salary book dataset contains a record of each person that worked a particular job in a particular year. The columns in the salary book are Fiscal Year, Department, Agency/Institution, Name, Gender, Place of Residence, Position, Base Salary, Base Salary Date, Total Salary Paid, and Travel & Subsistence.

Department, Agency/Institution, Name, Gender, Place of Residence, and Position are self explanatory; used together, these columns help identify individual people and their jobs.

Fiscal Year is the year in which the job took place. Base Salary contains the payment method (ex. hourly, bi-weekly) and the amount of those payments. Travel & Subsistence is any money provided for transportation for an employee during the year. Base Salary Date is the same as fiscal year, except the date starts on July 1st instead of January 1st.

During parsing, Base Salary was split into two columns, one for the payment method and one for the payment amount. These columns were used to estimate the hourly wage and the hours worked estimate for each employee. Based off of the payment method and the payment amount, a column for hourly rate was made. Using hourly rate and the total salary for a given year, it is possible to estimate the number of hours worked in the year.

The inflation rate data was obtained from the Bureau of Labor and Statistics for years 2007 to 2017 [12]. To calculate the cumulative inflation from 2007 to 2017, the product of the inflation was carried from year to year. Inflation moved roughly 20% from 2007 to 2017.

The effective federal funds rate data was taken from the NY Federal Reserve [13]. The daily data was grouped by year and averaged to get a yearly federal funds rate.

B. Iowa Regional Data

In order to examine the effects of regional attributes on salary, we needed to collect data about different regions in Iowa. We collected datasets from the Open Data Network that were made available by Socrata [7]. We collected a dataset on health [8], population [10], regional data [9], and education [11]. These were originally for the whole United States, and were then filtered down to just include Iowa regions.

Each of these had a similar format, with a column for region name, year, variable name and value for that variable. Since the Iowa salary data existed from 2007-2017 and we wanted to merge the salary and regional datasets, we filled in missing years in the regional data. For missing years for each variable, we used the most recent existing value, or the closest value in the future if no previous values existed. We then merged each regional dataset using a key pair of area name and year, with a column for each variable.

Our final variable set was {adult obesity value, children in poverty value, income inequality value, median household income, some college (value and percent), unemployment value, violent crime value, uninsured value, population density, population count, percent graduate or professional degree, percent less than 9th grade, percent associates degree, percent bachelors degree or higher, percent high school graduate}. Finally, we merged this dataset on the salary data using an inner join. Topics that did not depend on regional data were not analyzed using this merged dataset.

III. Gender Gap

Is the gender gap real? Undoubtedly, this question has crossed our minds at some point, especially if you are a woman. The United States (US) has come a long way from addressing inequality between men and women since the 19th Amendment to the US Constitution in 1920, which granted women the right to vote. But how much of the gap have we really closed? The Iowa State Salary Book Dataset is a great dataset for exploring this topic because it contains approximately 660 thousand entries of salary, region, and employment data for state employees, of which more than half identify as females. There were other genders in the dataset but they were removed for the purposes of this topic as we only consider the male and female genders.

Since the Iowa State Salary Book dataset is, in fact, a salary book, the first question that comes to mind in regards to the gender gap is the pay gap. Do men earn a higher salary on average than women? Another related question we want to answer is: do more women work part time than men? Finally, we will consider the glass ceiling, sticky floor, and glass escalator theories by looking at career mobility rates between the two genders. More precisely, we examine the following question: do women have a lower or equal career mobility rate than men?

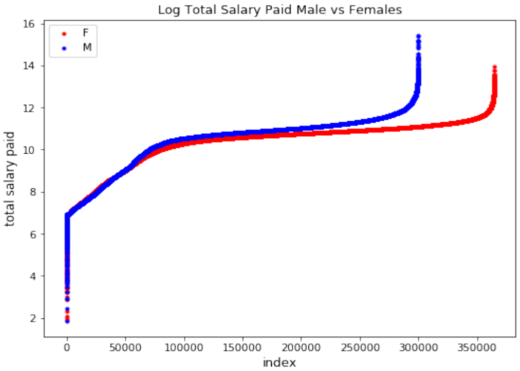
The main method used in testing the hypotheses was the Wald's test and **we make the assumption that the male and female populations are independent.** In all three hypotheses, the parameter we are estimating is either the Bernoulli maximum likelihood (MLE) estimate of the mean or the non parametric sample mean and sample variance estimates. **Since both the MLE estimator and the sample mean and sample variance estimators are asymptotically normal, the Wald's test is applicable.**

A. Gender Pay Gap

1) Hypothesis:

- H_0: Women earn an equal or lower salary than men on average
- H_1: Men earn a higher salary on average than women

2) Methods: Before summing up the total salary paid for both groups, rows in which the Total Salary Paid column was missing, 0, or a negative value were dropped. There were 27 and 15 rows that were dropped in the female and male datasets respectively. When we plotted the log of the total salary paid of all the male and females, it was visually clear that the males had a higher log total salary paid values overall, as can be seen in the following graph.



Since we don't know which distribution the log total salary comes from, we use the sample mean and sample variance plug in estimators when calculating the Wald's test statistic.

3) Results: For the first hypothesis, we obtain a Wald's test statistic value of -85.22. Since -85.22 is less than $Z_\alpha = 1.645$, for $\alpha = 0.05$, we reject the null hypothesis and conclude that men earn a higher average total mean salary than women.

B. Who Works Longer: Men vs Women

1) Hypothesis:

- H_0: Either more men work part time than women or they work the same amount of time
- H_1: More women work part time than men

2) Methods: Whether an individual works part time or full time is determined by their Hour's Worked Estimate value and/or their Total Salary Paid value if the former is missing. We set the threshold for full time to be more than 40 hours, and part time status is assigned otherwise. Since an individual can either be full time status or not (part time), we thus use the Bernoulli MLE estimate for the mean, which is also the sample mean, and plug in estimate for the variance; namely, $p^{hat}(1-p^{hat})$, where p^{hat} is the probability of an individual being full time.

3) Results: For the second hypothesis, we obtained a Wald's test statistic value of 29.8. Given that the null hypothesis was the fact that more females work full time then men, and since 29.8 is not less than $Z_\alpha = 1.645$, for $\alpha = 0.05$, we accept the null hypothesis and reject the alternative.

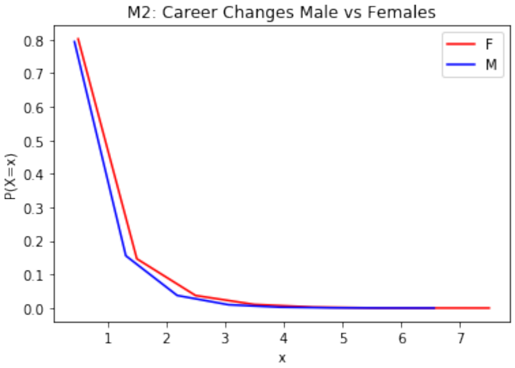
C. Career Mobility

1) Hypothesis:

- H_0: Women have a higher career mobility rate than men
- H_1: Women have a lower or equal career mobility rate compared to men

2) Methods: We defined career mobility as the movement of employees across grades/positions, both upward and downward, or a complete change in occupation [3]. Two metrics were formulated to measure career mobility rate for an given individual with the above definition in mind. For the first metric, mobility rate is defined as an indicator variable; ie, it will be equal to 1 if a given individual changed jobs at least once. The second metric is defined as the total number of job changes for a given individual over their career. Since the first metric is Bernoulli with p=probability of having changed jobs, we thus use the Bernoulli MLE estimate for the mean, which is also the sample mean, and plug in estimate for the variance; namely, $p^{hat}(1 - p^{hat})$, where p^{hat} is the probability of an individual having changed jobs. Since we don't know what distribution the second metric comes from, we use the sample mean and sample variance plug in estimators when calculating the Wald's test statistic. In addition to calculating the Wald's test statistic for both metrics, we ran permutation tests as well, using the absolute difference in means as the test statistic. We calculate the p-value by then measuring how many tests have a test statistic that is larger than the observed value after shuffling the genders to simulate a random sample.

3) Results: For the third hypothesis, we obtained Wald's test statistic values of -4.52 and -2.7 for the first and second metrics respectively. Since both values are less than $Z_\alpha = 1.645$, for $\alpha = 0.05$, we reject the null hypothesis and conclude that men have a higher mobility rate than women. Additionally, the p-values obtained after running 1000 trials were 0.009 and 0.008 for metric 1 and metric 2 respectively, providing further evidence that the null hypothesis should be rejected. The following is the probability density function plot of the number of job changes (metric 2) for males and females.



You can see around x=3 that women have a higher mobility rate for 0 to 3 job changes; however, beyond that, males have a higher mobility rate, and thus a higher mobility rate overall.

D. Wrapping Up the Gender Gap

The first result was not very surprising. As progressive as the country has been since the 1920s, inequality between between the two genders still exists in different forms, and the gender pay gap is one of them. According to a study done by the American Association of University Women (AAUW) in 2016, Iowa was ranked 41 out of 51 for having a pay gap statistic of 77%, meaning women were paid 77% of what men were paid [4]. Thus while Iowa is not dead last, the pay gap is still very transparent.

The second result was quite interesting; we believed that because women usually still had to take care of the household, with regards to the children, cooking, cleaning, etc, more of them would work part time than the men. However, it is not the case, since in fact more women work full time than men. The third result was also expected. The term *glass ceiling* was first used by Marilyn Loden during a 1978 speech [5]. Related terms such as *glass escalator*, and *sticky floor* came afterwards. Whether these concepts are a myth or not have been debated and studied by bodies such as the US Federal Glass Ceiling Commission. Nonetheless, our result clearly shows that males have a higher career mobility rate than females.

The results above were determined after testing various hypotheses formulated by examining the Iowa State Salary Book. Thus while the gender gap is evident in Iowa, at least at the public level, one can argue that Iowa is not representative of the country, or that the gender gap does not exist on the private level, or even that due to some loopholes and discrepancies in the data, it does not exist in Iowa on the public level either. Since additional data would be required to counter the first two arguments, we will instead address the third.

As stated by the Iowa Legislature, the Iowa State Salary Book data was received from the Department of Administrative Services without additional verification or editing [2]. In consequence, there is the possibility that some of the data entries are wrong, or have been altered by noise. In fact, upon inspection of the dataset, over 1 million values are missing, with 270 of them being the gender value. We also note however, that the gender column is missing the least number of entries than all other columns that are missing data. Furthermore, all other non male and female gender entries were removed from the dataset so that we could focus on the select two. **Thus, from the data perspective, we have tried to minimize the number of errors so that the data is as reflective of the true data as possible. In regards to the noise, we believe that the noise is not significant enough to change the results of our hypothesis testing.** The difference between the Wald test statistic and the Z_α value is simply too large for the first two hypotheses. Only the Wald's test statistic in the third result is remotely close to the Z_α, but even the permutation test p-values show that it takes more than luck and random chance to arrive at the results we have. We thus conclude that the gender gap is real in the Iowa public sector. If the results are reflective of anything, it is that Iowa is most likely not the only state in the nation with an evident gender gap.

IV. AREA OF RESIDENCE

The purpose of this topic is to discover how different attributes of a person's area of residence can affect their salary. This is a difficult thing to do, since there are many possible contributors to a person's salary that may be hidden to us, and within any given region there is a high variance in income. But by looking at overall trends and distributions of wages, we can start to find patterns that hint at factors underlying differences in income.

A. *Linear Regression on Area Attributes and Average Salary*

1) Hypothesis: For each area attribute, we want to find if there is a strong correlation with the average income across different areas. We use the linear regression technique to investigate this. We have one hypothesis for each area attribute, where the null hypothesis is that the attribute is not a strong contributor to average income, and the alternative hypothesis is that it is a strong contributor.

Assumptions:

- The data is linear
- The variances of the errors are equal
- The errors are normally distributed with a mean of 0

While we cannot fully assert that the data is truly linear for any of the attributes, we can check the distribution of the estimated errors and determine if they are normal.

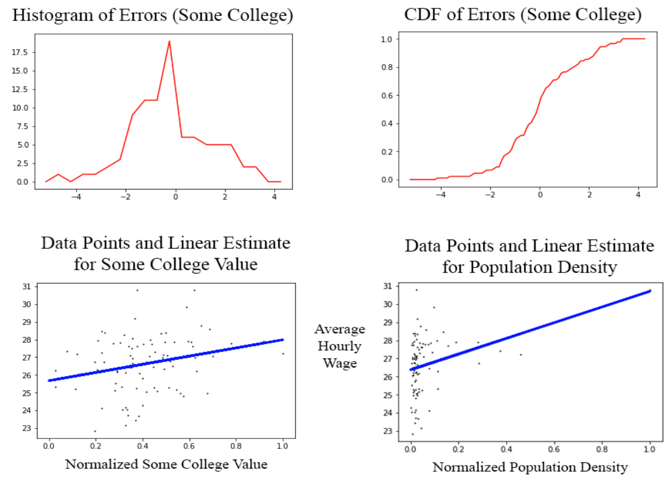
2) Methods: To set up the linear regression, we define our regressors as the attribute values for an area (normalized between 0 and 1), and the dependent variable as the average hourly wage for an area. We perform Simple Linear Regression (SLR) on each attribute independently, as well as Multiple Linear Regression (MLR) on all attributes at once. We will focus more on the SLR results, but examine MLR coefficients as a sanity check.

The reason that our observations are average area incomes rather than salaries for each individual is that using individuals would cause a region with a high population to outweigh the others. We also only examine one year at a time (the results below were performed on 2016 data). The reason we do this is to nullify effects that occur over time and focus on effects over area instead. It is well known that incomes in general increase over time, and if there was another variable that increased consistently over time, it would be falsely correlated with income.

3) Results: We found T-values for each attribute using the SLR coefficient and the standard error relative to the SLR predicted values. The degrees of freedom (df) for the T-test are relative to the number of sampled data points, and we had 89 different residences that had values for each attribute. At df = 80, the T-value for α = .20 is 1.282, at which the following attributes were found to be significant:

- Uninsured Value (SLR coef: -2.79, MLR coef: -11.15, T-value: 1.88)
- Population Density (SLR coef: 4.33, MLR coef: 0.01, T-value: 2.92)
- Population Count (SLR coef: 4.18, MLR coef: 0.00, T-value: 2.81)
- Some College Value (SLR coef: 2.31, MLR coef: 3.23, T-value: 1.52)
- Percent Graduate or Professional Degree (SLR coef: 2.36, MLR coef: -0.10, T-value: 1.53)
- Percent Less Than 9th Grade (SLR coef: -2.71, MLR coef: -0.02, T-value: 1.80)
- Percent Bachelor's Degree or Higher (SLR coef: 2.59, MLR coef: 0.08, T-value: 1.72)

Assumption Check: We examined the distributions of the errors to investigate the validity of using linear regression on the attributes.



Both the histogram and CDF closely resemble a normal distribution, though there are some small differences. We can see that the peak in the histogram is slightly to the left of the mean. In addition, on the CDF, we see a long tail on the left of the graph. This can likely be explained by the fact that we are only examining one variable at a time, and other unseen variables also have an impact on the average salary. Interestingly, the error graphs look very similar for all attributes.

However, the homoskedasticity of the errors seems to vary across attributes. We can see that the Some College Value errors appear to be more homoskedastic than the Population Density Errors. Interestingly, Some College Value also has a much higher MLR coefficient than Population Density. This may be an indicator that for some variables, such as Population Density, while there is a linear correlation, other attributes may also have an impact.

4) Conclusion: We found that linear regression pointed to a high correlation between several attributes of regions and their average incomes. We found with significance that education has a strong effect on income, with degree earners in an area boosting income and pre-highschool graduates lowering income. Population and population density had a very strong positive effect, and lack of insurance had a very strong negative effect.

B. *Comparing Individuals from Contrasting Areas*

1) Hypothesis: We want to discover how the salaries of individuals from highly contrasting areas differ, with respect to different area attributes. For each area attribute, we will compare the salaries of individuals who are from regions within the first quartile of that attribute's values, to those from regions within the fourth quartile of that attribute's values.

We will use a Wald's test to compare the salary means of these population pairs, and the KS test to compare their distributions. There are two null hypotheses for each attribute: 1) that the salary means of the two populations are the same and 2) that the salary distributions are the same.

Assumptions:

- Wald's test assumes that the statistic is asymptotically normal for the parameter in question, and a mean test satisfies this requirement.
- KS test has no assumptions.

2) Methods: We first restrict all data to a single year (2016 used for this analysis) to nullify the effects of changes over time. For each attribute, we separate regions that are in the first and fourth quartiles for the values of that attribute. Then we create two populations: one for individuals living in the regions with the lower percentile and one for individuals living in the regions with the higher percentile. We then perform a Wald's test to compare the mean hourly wage of both populations, and a KS test to compare the distributions of the hourly wage of both populations.

3) Results: The following are significant at α = .01:

- Average of Upper Quartile higher than Average of Lower Quartile:
  - Income Inequality Value, Violent Crime Value, Population Density, Population Count, Some College Value, Percent Graduate or Professional Degree, Percent Bachelor's Degree or Higher
- Average of Lower Quartile higher than Average of Upper Quartile:
  - Adult Obesity Value, Uninsured Value, Percent Less than 9th Grade, Percent Associate's Degree, Percent High School Graduate

In addition, when comparing distributions with the KS-test, we found all attributes to yield significantly different distributions, with the exception of Children in Poverty and *Percent Graduate or Professional Degree*.

4) Conclusion: When comparing the results of linear regression and the quartile method shown above, we found that the directions of the trends always agreed for significant variables. That is, for every variable that showed significance for either test, if the Upper Quartile had a higher mean than the Lower Quartile then the SLR coefficient was positive, and visa versa. In addition to the SLR results, this test reveals that *adult obesity* has a negative impact on salary, and rather concerningly, that *income inequality* and *violent crime* have a positive impact, and that percent only having completed an Associate's degree has a negative impact. While these attributes may not be direct influences on salary, we can infer that they are strong indicators.

C. Comparing Newcomers to Long-Time Residents

1) Hypothesis: The goal here is to discover how the incomes of people who have just moved into a city differ from the incomes of people who have lived in that city for a long time. We examine the differences of incomes of these populations when the newcomer population just moves into the city, and then examine the incomes of these two populations several years in the future. The null hypothesis is that the averages of these populations always remains the same, and we will use Wald's test to investigate this.

Assumptions:

- Wald's test assumes that the statistic is asymptotically normal for the parameter in question, and a mean test satisfies this requirement.

2) Methods: We define the requirement for being a long-time resident in area a as someone who has lived there for n years in a row, at a given year y. A newcomer is someone who moved into a at year y. Then, at year y + t, we take the difference of the average incomes of these populations, and add that difference as an observation. We collect all possible observations of this kind across all areas and all start years to construct our population for (n, t).

For all combinations of $n \in \{2,..,4\}$ and $y \in \{0,..,5\}$, we do a Wald's test on the population mean, hypothesizing that the average salary difference is 0.

One complication was that the data for this test was sparse, since we had to track the same residents throughout time, and if they move out of a city we lose them as a data point. Because of this (and to save on computation time) we only looked at residences that were in the top 10 for number of datatpoints.

3) Results: We noticed that the same pattern occurred for all values of n, so we will show the results for n = 2, since it contained the most data. The difference is long-time - newcomers, so a negative value indicates that the newcomer average was higher.

- y = 0, Avg Wage Difference: 3.01, W: 10.67
- y = 1, Avg Wage Difference: 1.97, W: 6.22
- y = 2, Avg Wage Difference: 1.29, W: 3.71
- y = 3, Avg Wage Difference: 0.45, W: 1.40
- y = 4, Avg Wage Difference: -0.49, W: 1.18
- y = 5, Avg Wage Difference: -1.20, W: 3.00

We found that the wages of long-time residents were initially higher than those of newcomers to an area. However, after approximately 3.5 years, the average income of the initial newcomers actually exceeds that of the long-time residents.

4) Conclusion: There are many possible reasons that the average salaries of newcomers will eventually overcome those of long-time residents. A possible one is that people moving into a new area are younger and are pursuing a new career, and will end up succeeding. Another is that those who have lived in the area for a while may not be changing jobs and thus their incomes have stagnated.

V. Position Changes

We are interested in the question - do position changes influence the salary change rate, R_s? Furthermore, since we are trying to figure out whether the gender gap is real, another question comes to us - does your gender affect your R_s?

In this section we use Wald's test to test our hypotheses and we use the permutation test to get the p-value. **Walds test assumes that the statistic is asymptotically normal for the parameter in the question.** Since the parameters μ and σ^2 are estimated by MME, they satisfy the requirement of Wald's test. We use Hourly Wage Estimate to calculate the R_s. Since the original data set is not clean, we try to extract meaningful data with the procedures listed below. The data is dirty in a few ways, one was is that some of the rows do not have position information and some of the rows indicate that some people have multiple positions during the same fiscal year.

1) Remove the data from the dataset if it does not have Position information or Hourly Wage Estimate information. The size of the original data shrinks from 665325 rows to 255157 rows after cleaning the data. Issues with Position or Hourly Wage Estimate values indicate issues with Base Salary Payment Method, Base Salary Payment Amount, Total Salary Paid, or Position values for a particular row. This explains why filtering removed approximately 61.65% of data.
2) Remove outliers by using a modified Tukey method. We defined an outlier as any observation with its Hourly Wage Estimate value outside the range: $\big[Q_1 - k_0(Q_3 - Q_1), Q_3 + k_1(Q_3 - Q_1)\big]$, where k_0 is 1.5 and k_1 is 3.0; Q_1 and Q_3 are the first quarter and the third quarter of the Hourly Wage Estimate. The size of the dataset shrinks from 255157 to 252328 after applying this filter rule. The number of the unique names was reduced from 175208 to 45081 after both data cleaning and outlier removal.
3) Group the data with Name and Fiscal Year. If a name appears more than once in a fiscal year, then we remove all rows containing that name for all years. The main reason why we want to remove these rows is that we cannot distinguish whether they represent a single person or multiple people. The number of the unique names decreased from 45081 to 43208.
4) We split the dataset into groups, where the people in each group have worked for the same number of years. The groups span from people who have only worked one year to those who have worked for eleven years. When grouping by years worked, we get the results shown below.

Years worked	# of Unique Names
1	9316
2	4776
3	5208
4	2732
5	2115
6	1972
7	2319
8	1729
9	1756
10	2307
11	8978

We used the group where people have worked for eleven years because we have a full record of their work history from 2007 to 2017.

The R_s, salary change rate, is computed as follows:

$$R_s = \frac{HWE_i - HWE_{i-1}}{HWE_{i-1}}$$

, where HWE is Hourly Wage Estimate, and i is the fiscal year from 2007 to 2017.

Moreover, the R_s is computed every year for each person and each result is placed into one of two groups: change rates that correspond with a job change or those without a job change.

A. *General case: Do position changes influence the Rs*

1) Hypothesis:

- H_0: Position changes will not influence the R_s
- H_1: Position changes will influence the R_s

2) Results:

- Estimated parameters

position status from 2007 to 2018	$\hat{\mu}$	$\hat{\sigma}^2$
with position change	0.11221	0.35404
without position change	0.06049	0.07432

- Test results

w-statistic	p-value
15.6556	0.0

- Since the w-statistic is larger than $Z_\alpha = 1.96$, for $\alpha = 0.05$ and p-value is much less than 0.05, which shows strong evidence against the null hypothesis, we reject the null hypothesis and conclude that position change will influence the R_s.

B. *Gender comparison without position change*

1) Hypothesis:

- H_0: Different genders will not influence the R_s
- H_1: Different genders will influence the R_s

2) Results:

- Estimated parameters

gender	$\hat{\mu}$	$\hat{\sigma}^2$
female	0.06263	0.08067
male	0.05493	0.06038

- Test results

w-statistic	p-value
3.3236	0.0

- Since the w-statistic is larger than $Z_\alpha = 1.96$, for $\alpha = 0.05$ and p-value is much less than 0.05, which shows strong evidence against the null hypothesis, we reject the null hypothesis and conclude that genders will influence the salary change rate when someone does not have any position change in the previous years.

C. *Gender comparison with position changes*

1) Hypothesis:

- H_0: Different genders will not influence the R_s
- H_1: Different genders will influence the R_s
- Estimated parameters

gender	$\hat{\mu}$	$\hat{\sigma}^2$
female	0.10552	0.33216
male	0.10936	0.48436

- Test results

w-statistic	p-value
-0.5818	0.575

- Since the w-statistic is larger than $Z_\alpha = -1.96$, for $\alpha = 0.05$ and p-value is larger than 0.05, which means weak evidence against the null hypothesis, we can not reject the null hypothesis and conclude that genders will not influence the salary change rate when someone has position change.

D. Conclusion

It is interesting that employees of different genders have different R_s when they have not changed positions at all over the past ten years. On the other hand, if an employee has had at least one position change then the gender has no statistical significance in showing that gender plays a role in R_s.

VI. Departments

A. Is the total salary paid in each department normal?

1) Hypothesis: The distribution of total salary paid in a department is normal. We think it is meaningful to find the shape of the salary distribution in a department. Since in really life, very few people are paid very high, very few people are paid very low, whereas the majority of people are paid at moderate level. Furthermore, the histograms of the salaries indicate that the distribution of the salaries in some departments might be normally distributed. So our proposed hypothesis is that the distributions of the total salary paid in all departments is normal.

2) Methods: MME (method of moments) is used to infer the parameters μ and σ from the data. Then we use KS(Kolmogorov-Smirnov) test to check if the normal distribution model is a good fit for the data. Samples are standardized (using the μ and σ inferred from MME) and compared with a standard normal distribution. In the analysis, we only consider the departments that have more than 30 employees, arguing that a sample size less than 30 is not big enough to get a valid results. Also outliers are removed using the Tukey outlier detection algorithm. The null hypothesis is that our standardized samples are in the same distribution as the standard normal distribution. The KS test is used to check if they are the same distribution.

3) Results: Only 5 departments out of 52 have the total salary paid in normal distribution. Image A(in the end of this section) is a PDF plot of total salary paid in Department of Aging, versus the fitted normal distribution. All of the 5 departments that have total salaries paid in a normal distribution are:

Department	KS results (D,CV)	p-value
Aging	0.0660, 0.11	0.11
Elder Affairs	0.074, 0.176	0.50
Energy Independence	0.11, 0.21	0.22
Campaign Disclosure	0.12, 0.21	0.16
Student College Aid	0.05, 0.081	0.10

The D statistic in the table is the absolute max distance (supremum) between the CDFs of the two samples. The null hypothesis is rejected when D>critical value. The CV is the critical value calculated using $c(\alpha)\sqrt{n}$, where $c(\alpha) = 1.95$ and n is the sample size of the data set. The results in the tables show that we failed to reject the null hypothesis, which means the salaries of all of the 5 departments in the above table are normally distributed.

B. The distribution of total salary paid in similar departments

1) Hypothesis: The salary distributions in similar departments are the same. In our dataset, the departments that are similar are: University of Iowa, Iowa State University and Iowa Northern state university because there are all universities. We did a CDF (cumulative distribution function) plot of the three universities. The lines in the graph almost overlap each other and this also indicates that the distribution of the total salary paid might be the same. The null hypothesis H_0 is salary distribution in department X is the same as the salary distribution in department Y, where department X and department Y are any combination of the three universities above.

2) Methods: The KS test is applied to test if any two of the three universities have the same distribution in total salary paid.

3) Results: In the table, UIS, UI, and UNI represent University Of Iowa State, University of Iowa, and University of Northern Iowa respectively. The D statistic in the table is the absolute max distance (supremum) between the CDFs of the two samples as defined before, and CV here is the critical value calculated using $c(\alpha)\sqrt{\frac{m+n}{mn}}$, where $c(\alpha) = 1.95$ and m, n are the sample sizes. If D>critical value then we reject the null hypothesis. As we can see from the table, all the D values are greater than the critical value(CV) so all the null hypotheses are rejected. The three universities, which we think are similar departments, do not have similar distributions in the salary paid.

compare pairs	KS result(D,CV)	reject null
UIS vs UI	0.042, 0.0086	yes
UIS vs UNI	0.041, 0.0150	yes
UI vs UNI	0.045, 0.0139	yes

C. Total Salary Paid in Beta Distribution

1) Hypothesis: The distribution of the the total salary paid in the 3 universities: University of Iowa, Iowa State University, and University of Northern Iowa is in Beta distribution. Depending on the university we are testing the three hypotheses, which are in the following format: H_0: distribution of the total salary paid in University of Iowa / Iowa State University/University of Northern Iowa is in Beta distribution.

2) Methods: MME is used to infer the parameters α, β in Beta distribution. $\alpha = \bar{X}\frac{\bar{X}(1-\bar{X})}{S^2} - 1$, and $\beta = (1-\bar{X})\frac{\bar{X}(1-\bar{X})}{S^2} - 1$ where $\bar{X}$ is the sample mean and S^2

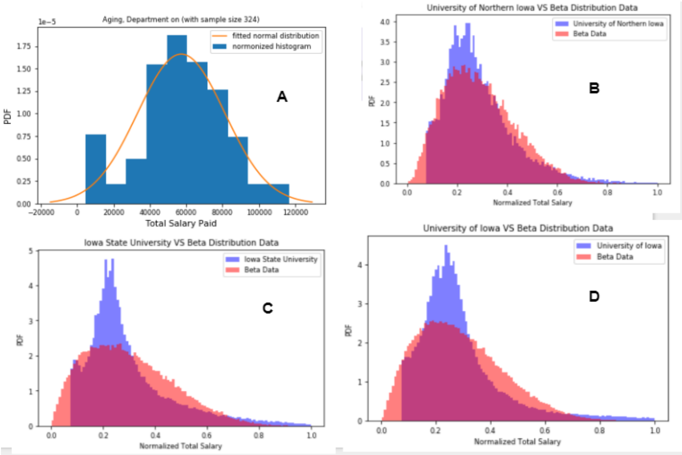
is the sample variance. The KS test is used to check if the distribution is a Beta distribution.

Basically, after we get the α and β values, we use these parameters to generate random data from a Beta distribution, with the same α and β parameters. Then the KS test was applied to both the generated sample data and the original data to test whether they are from the same distribution. If this is the case, then the original data is shown to be from a Beta distribution with the inferred α and β parameters.

3) Results:

	KS result(D,CV)	reject null
UIS	0.744, 0.0072	yes
UI	0.101, 0.0045	yes
UNI	0.057, 0.013	yes

Image B, Image C and Image D are the PDF plot of the total salary paid in 3 different universities versus the PDF plot of the fitted Beta distribution. The PDF plot of the original total salary data looks very similar to the PDF plot of the Beta distribution with the fitted parameters from the original data. Yet, it can be seen from the table all D>CV, which means all null hypotheses are rejected. Non of the universities have total salary paid in Beta distribution. UIS,UI,UNI,D,CV are defined as before.



D. Conclusion

For the majority of departments the total salary paid are not normally distributed. Even though their CDF plots almost overlap each other, the distribution of total salary paid in similar departments (in our case the three universities) are not the same. The salary distribution in the three universities, University of Iowa, Iowa State University and Northern Iowa State University are not in Beta distribution, even though their PDF plot looks very similar to that of a Beta distribution.

VII. Effects of National Trends

Do national trends relate to peoples salaries? Can national trends and events be related to salary? In short, yes. We use events such as the Great Recession and trends like inflation rate and federal funds rate to inspect salary data.

Inflation is important to consider because it affects people's buying power. If people's salaries do not increase with inflation then their income will be less effective. If salaries do not move with inflation then that can indicate issues within an organization.

Comparing years 2007 and 2008 are important because it shows the immediate effects of the 2008 stock market crash on salaries. Comparing years 2008-2012 and 2013-2016 are important because they show the effects of the recovery period and the expansion period after the recession. The recession was considered over in mid 2009, but the stock market did not surpass its pre-crash value until early 2013. The years 2013 to 2016 show the market post the high 2008 levels. Year 2017 was not included to nullify any possible impact following the presidential election.

Federal funds rate is used to help control inflation in the economy. If rates increase or decrease the economy can be affected and salaries should theoretically be affected as well.

A. Effect of Inflation Rate on Salary

1) Hypothesis:

- H_0: Inflation Rate and Salary Increase are not correlated
- H_1: Inflation Rate and Salary Increase are correlated

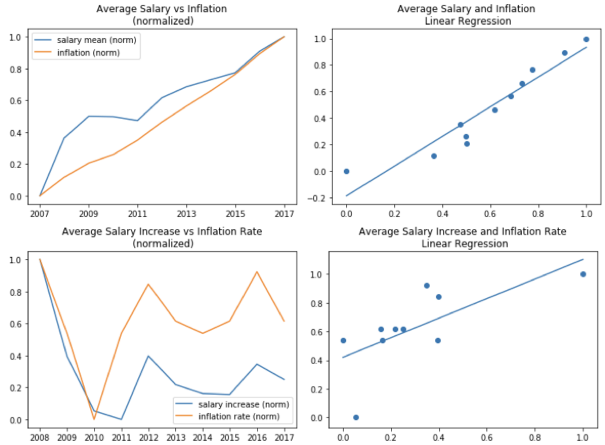
2) Methods: Simple linear regression was used to test whether there was a correlation between inflation rate and salary.

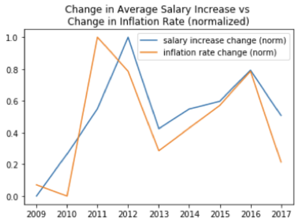
Cumulative inflation, inflation rate, and change in inflation rate were compared to salary, salary increase, and change in salary increase, respectively. This gave a more complete picture of how salary was moving in relation to inflation. For these linear regressions the data was normalized.

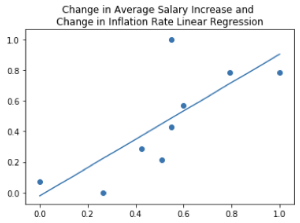
Assumptions:

- The data is linear
- The variances of the errors are equal
- The errors are normally distributed with a mean of 0

The graphics below show the plots of both salary and inflation and the corresponding linear regression.







3) Results: Linear Regression shows that:

- Average salary and cumulative inflation are positively correlated with a regression coefficient of 1.1222 and a T-value of 11.5219.
- Average salary increase and inflation rate are positively correlated with a regression coefficient of 0.6820 and a T-value of 3.5864.
- Change in average salary increase and change of inflation rate are positively correlated with a regression coefficient of 0.9253 and a T-value of 4.3484.

This shows that inflation is a good indicator of how salaries will behave. As a consequence, we reject the null hypothesis and conclude that inflation rate and salary are positively correlated.

B. Effect of Events on Salary

1) Hypothesis:

- H_0: Salaries between years of events are not significantly different
- H_1: Salaries between years of events are significantly different

2) Methods: The salary data in this test was adjusted for inflation. This was done by dividing the cumulative inflation by the salary for that year. The salary data was separated into individual years and into groups of years. Salary data from year 2007 was compared to year 2008 to look at the immediate effect on the stock market crash. The years 2008-2012 and 2013-2016 were also compared in order to explore the recovery and post recovery salary values since the Great Recession.

Wald's test was used to determine whether there was a significant difference in salary between the two time periods to quantify the potential effects that the event had.

3) Results: Walds Test shows that salary data between 2007 and 2008 differs significantly (wald statistic = 26.583), which reflects the potential effects caused by the stock market crash in 2008. Similarly, Wald's test shows that the salary data between years 2008-2012 and 2013-2016 are also significantly different (wald statistic = 8.907). This shows that the salary from the recovery years of the great recession are different from the expansion years coming out of the recession.

C. Effect of Federal Funds Rate on Salary

1) Hypothesis:

- H_0: Federal Funds Rate is not correlated with Salary and Salary Increase
- H_1: Federal Funds Rate is correlated with Salary and Salary Increase

2) Methods: Simple linear regression was used to test whether there was a correlation between effective federal funds rate and salary. The effective federal funds rate was tested against the yearly salary and the increase in yearly salary. A derivative of the funds rate was not taken in this test and the salary in this test was not inflation adjusted because the federal funds rate is partially used to control inflation.

Assumptions:

- The data is linear
- The variances of the errors are equal
- The errors are normally distributed with a mean of 0

3) Results: Linear Regression shows that average salary and federal funds rate are negatively correlated (LR coef: -0.773, T-value: 3.751) while average salary increase and federal funds rate are positively correlated (LR coef: 0.3563, T-value: 5.4389). The federal funds rate during the recession was set to increase the growth in the economy. This shows that the federal funds rate is correlated with the increase in salary.

D. Conclusion

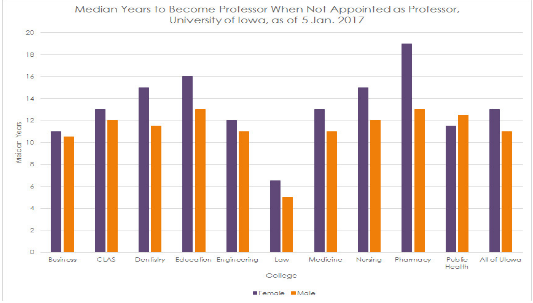
National trends are important and affect real people and their income. Inflation and funds rate can be used as indicators for how salary will change from year to year. Events also have short and long term effects on people's income.

VIII. Prior Work

1) Equity At Iowa Project: The Equity at Iowa project is a collaboration project between faculty at University of Iowa and independent researchers to explore the open data found in the Iowa State Employee Salary Book [6]. The project mainly focuses the distribution male and female professorships at the University of Iowa across different fields. In particular, they examine the number of full male and female professors across colleges, median number of years to become a full professor, median base salary of faculty in a specific college, and hiring initiatives of select college departments.

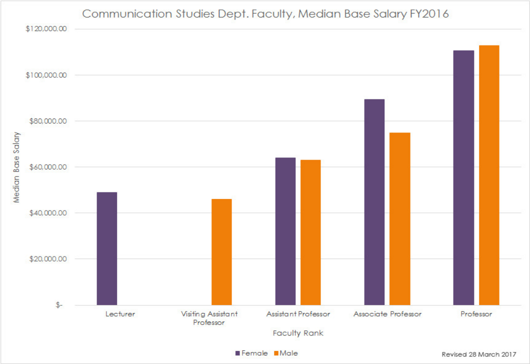
The main sections that we have in common are the median number of years to become a full professor and the median base salary of faculty in a specific college, which we will compare with our gender gap analysis. Since our department analysis mainly focused on what distributions the salaries come from, we cannot compare that analysis with the results from the Equity at Iowa project.

The following graph shows the median number of years it takes for the female faculty and male faculty to achieve the rank of full professor as taken from their project site.



The main observation we can draw from the graph is that it takes longer for women to reach full professor rank than men across all colleges except for Public Health; however, even for Public Health the difference is approximately one year whereas the max difference is 6 years for women faculty in the Pharmacy department compared to men in the same department. Their analysis concludes that for faculty in all of the colleges combined, the median time to full professor rank is 2 years longer for female faculty than it is for male faculty. We can loosely compare this result with the result from our gender gap analysis on mobility rate. Recall from that analysis that men have a higher mobility rate than women; we can thus correlate that result with the fact that men reach full professor rank faster than women across most colleges. The main difference between our analyses is that the gender gap analysis on mobility rate did not focus on specific jobs but across all public jobs, whereas their analysis focused on university professorships across colleges.

The following graph shows the median base salary for faculty of the Communication Studies department for fiscal year 2016 as shown on their project website.



The main observation we can draw from the graph is that women assistant professors and associate professors make a higher median base salary than men in the Communication Studies Department; although the difference for assistant professor is not very large. Male professors make a higher median base salary than females in the same department but the difference is not very large either. This is somewhat contradictory to the result obtained from our gender gap analysis on pay gap, which said that men earn an higher average salary than females in general. However, again, note that their analysis focused only on one department while ours averaged over all public jobs. Furthermore, note that their Communication Studies department is small, containing only 18 people, and that eight of them are cross listed across multiple departments, thus giving these individuals a much higher salary despite their position. Thus we conclude that the data their graph is displaying is skewed by salary factors beyond the individual's department.

2) U.S Income Distribution: Similar work has been done to analyze the income distribution by race and ethnicity in the US[14], Whereas our work tries to find a statistical model for the distribution of the salary as a whole. In their paper they found out that there is racial and ethical difference in the distribution of the income/salary. For example, more black householders or Hispanic householders have income lower than $50000.

3) Job Mobility and Earnings: Prior work has been done to analyze the effects of job mobility on earnings[15]. They compared the effects of mobility between genders in the early and late stages of their careers, whereas our work tries to find the relation between position changes and salary change rate. They found evidence showing that changing multiple employers within a year tends to have negative effects on hourly earnings; on the other hand, staying with the same employers within a year tends to have positive effects on hourly rate of pay. We cannot do the same analysis since our dataset doesn't have the information of job change within a year.

IX. Future Work

As concluded in the gender gap section of the report, more women work full time than men. While it is true that in today's society, men are no longer the sole breadwinners of the family, an interesting question to consider is whether, given that more women work full time than men, are more men taking up the household side of things? Furthermore, we would like to delve deeper into the gender pay gap by looking at why women are paid less. Is it because of their job sector or is it because despite more women working full time, they still have household responsibilities, and thus they have a harder time advancing in their career to higher paying positions? These are all additional questions in the gender gap topic that we would like to look into in the future.

Additional information on each individual would have allowed for a more detailed analysis for certain variables. For instance, instead of just examining how an area's average level of education affects the income of its residents, it would be interesting to see how the education of an individual affects their personal income. Health factors for each individual, and information on family, such as number of children, would also be very interesting. We could have also benefited from more fine-grained area information, since the salary book only reported county of residence, but economic factors can vary greatly within some counties.

There are many interesting hypotheses related to position changes can be tested. For example, we can count how many times an employee changes position during a fixed length time duration. Then try to answer the question: does the frequency of position changes influence the salary change rate? And is the position change department-oriented?

More work needs to be done to see if we can find statistical models for the distribution of salary paid in different departments or as a whole. Finding valid statistical models for the salary distribution will provide insights to problems such as: Are employees paid fairly? And, is the current distribution of payment (income) reasonable for a stable society?

Doing analysis with local trends is the next step in future analysis. Getting data for local school budgets and for local government revenues would enable more questions to be answered on a community scale instead of a state scale. Getting data on private employment could give a more complete picture of the economic status of communities and what industries mainly provide income to the people on Iowa.

ACKNOWLEDGMENT

We would like to thank Professor Anshul Gandhi for offering CSE 544: Probability and Statistics for Data Science in Spring of 2018 and for teaching us various methods and techniques required to complete this project.

REFERENCES

[1] GitHub Project Page
[2] State of Iowa Salary Book Dataset
[3] Career Mobility Definition
[4] AAUW Pay Gap
[5] Glass Ceiling
[6] Equity At Iowa
[7] Socrata
[8] ODN RWJF Health Behaviors
[9] ODN Geographic Area
[10] ODN Population
[11] ODN Education
[12] Inflation Data
[13] Federal Funds Rate Data
[14] The U.S. Income Distribution: Trends and Issues
[15] The Effects of Job Mobility and Intermittent Work History on Earnings: A Comparison of Men and Women