

STAT 430:

Bayesian Statistics Final Project

Group Members:

Taylor Thiel, Aditya Subramanian,
Shairoz Sohail, Thomas Holzhauser

Net IDs (Respectively): tmthiel2, subrmnn4, ssohai3, holzhar2

Description:

Exploring gender based salary differences in higher education employees of public schools from the state of Illinois using standard frequentist approaches as well as Bayesian hierarchical analysis and Monte Carlo Markov chains.

Introduction:

When given public official's salaries from the state of Illinois, including employees of public schools, we can pursue analyzing some potential contributing factors since these are also made public information. With our end goal being to determine whether or not there is a gender gap in the higher education system in Illinois, we also consider other potential candidates for disparity in wages, namely: employer and job title.

There has been research that has been conducted in regards to the effect of gender on salaries in industry. The term 'gender gap' in regards to salaries has been widely cited, leading back to an analysis of 2013 U.S Census data that concludes "the typical woman working full time, year-round in the United States earned 78 percent of men's earnings. That number rose approximately 1 percent since 2012 — a change that is not significantly different."¹ We wish to determine whether this same phenomenon is present in regards to our dataset with consideration to two unique statistical methods: frequentist and bayesian approaches.

The goal at hand is to analyze our dataset and further determine the parameter of interest, θ , the difference between mean salaries of male and mean salaries of female employees at post-secondary institutions in Illinois. Data² is available online for the salaries of all employees of post-secondary institutions in Illinois for the year 2014 as this is a subset of public Illinois employees; this is the data we utilized.

After prepping our data we performed a frequentist analysis. We then appended gender onto our dataset and performed both a frequentist and bayesian hierarchical modeling methods on the residuals of said frequentist analysis. Initial analysis of data revealed it to not be roughly normal. However, the residuals did appear to be roughly normal, with the exception with some skewness. For the prior Normal distribution, we utilize hyper priors of an Inverse Gamma for sigma squared (and Gamma for tau squared), and Normal (for mu). These hyper priors are assigned extremely differing values in order to test for sensitivity in the bayesian model. Informative priors were not selected due to the fact that we are not modeling salary but the residuals of salary after modeling both school and job title.

¹ U.S Census Bureau, AAUW (2014)

² Better Government Association (2014)

$$P(\theta_M|y) \propto P(\theta_M|\mu, \tau^2) * P(y_i|\theta_i) * P(\mu) * P(\tau^2)$$

$$P(\theta_F|y) \propto P(\theta_F|\mu, \tau^2) * P(y_i|\theta_i) * P(\mu) * P(\tau^2)$$

θ_M = Mean Salary of Men in Academic Positions

θ_F = Mean Salary of Women in Academic Positions

y_i = Observations of i^{th} employee's salary in gender





Data Preparation:

The data used in the ensuing analysis is made available through the state of Illinois as a public database. The data contains salary information on all public officials including public school employees. First, we limited our searches to “All Higher Ed Employees” before using Python scripts to scrape the data from the HTML website. The script used regular expressions to carefully parse the document looking for all variables provided on said website. This data was outputted into a CSV file, where we then used excel for some data preparation before transitioned into using SAS for further data cleaning. The initial variables provided by the website include the following:

Variable Name	Variable Description
First Name	First name and middle initial of employee.
Last Name	Last name of employee.
Salary	Employee's salary.
Title	Position and title of employee.
Department	Always stated as “All Higher Ed Employees”.
Employer	University and campus of employment.
Year	Always stated as “2014”.

Excel was used to arrange the data in a format which could be easily read in by SAS. There were 4 main problems that needed to be addressed - all of which can be found in the next image as an example. To be discussed later on, gender was appended on using the employee's first name; we therefore needed a ‘first name’ variable. The blue arrow represents an initial preceding the first name, which was voided in our analysis. Similarly, represented by the green arrow, the initial here is intended for the employee's middle name. Once again, the initial was voided. The black arrow represents the concatenation of salaries, as the comma delimited salaries were read into the CSV as two independent accounts. Lastly, the red arrow represents the

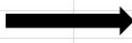
concatenation of titles, which was caused by a problem similar to that of the salaries and comma delimited inputs.

E Geoffrey			Geoffrey
Love			Love
\$147	840		147840
Asst Prof			Asst Prof
All Higher Ed Employees			All Higher Ed Employees
U Of I At Urbana/Champaign			U Of I At Urbana/Champaign
2014			2014
Craig A			Craig
Conrad			Conrad
\$147	672		147672
Chairperson	Dept Of Management		Chairperson Dept Of Management
All Higher Ed Employees			All Higher Ed Employees
Western Illinois University			Western Illinois University
2014			2014

With all of our data now in a single column, we were able to read the CSV into SAS quite easily. We then converted each group of eight observations (seven variables and one blank space) into a single observation containing seven variables. With our data now in a manageable and intuitive format, we proceeded to append gender onto the dataset using the employee's first name. This was done using the open source dataset 'babynames'³ which can be found in the R-package of the same name. This dataset was saved as a CSV and read into SAS. The dataset contains the number of females and males born of a given name in a given year. All names, which fell in the "unknown" gender category after attempting to append gender using the 'babynames' dataset, were voided. This accounted for 2,159 individuals being removed, which created a potential bias against unique and gender neutral names. We believe this impact does not negatively affect our analysis, as very few genders would be misclassified at this point. Furthermore, we do not believe that this bias targets genders of a specific pay grade, therefore, we believe this is a reasonable assumption; any bias would essentially be a random removal of an observation in terms of our *salary* analysis. We grouped by name and summed the counts of total females born of a given name as well as males. For example, consider the following (for simplicity, only years 1970 and 1971 were used):

³ Wickham (2014)

Year	Sex	Name	Prop						
1970	F	Trevor	13						
1970	M	Trevor	987						
1971	F	Trevor	13						
1971	M	Trevor	1096						



Name	Total Count	Proportion Male
Trevor	2109	0.988

In this example, we summed the total number of people with the name Trevor and then determined what proportion of them were Male; this was done for all names over all years. We then only considered names with at least a total count of 5 and a proportion male of greater than 90% or less than 10%. We then had a table to merge gender onto the dataset using first names.

Lastly, we needed to consolidate the number of unique levels of the variables job title (2372 levels) and university (17 levels). This was done by looking at different forms of the same value. For example, “Prof/...”, “Prof.”, “Prof”, “Prof of...”, and “Professor” are all values that should be treated the same. We therefore replaced substrings within the variable title with a single form of each title. We then only considered the first five characters of the title, as there are often extremely descriptive titles that follow the preface value of “Prof.”. After completing this process, the first 23 most frequently occurring job titles comprised about 95% of the observations. We therefore considered these values as the official title (called `grouped_title`) and all other observations were categorized as ‘Other’. Next, we consolidated the university variable into one value per institution. For example, U of I has three unique campuses and a fourth value for U of I employees not located on a given campus. We decided to consolidate based on institution to remove issues with some specific institutions having very low counts despite having multiple campuses. At this point, we had 18,293 data points which we could run analysis on. In order to further clean the data set before starting our analysis, we removed all employees with salaries under \$25,000 (in an attempt to remove part time employees), as well as those salaries three times of the Inter-Quartile Range above the third quartile. This upper bound was equivalent to salaries above \$247,678. The resulting amount left us with 14,740 data points. Our final cleaning took place with respect to dropping individuals who did not have a gender accounted for from the ‘babynames’ package. Limiting our results to those individuals specifically identified with male and female genders left us with 12,580 individuals. We therefore performed our analysis on approximately 70% of the original dataset, after considering extreme salaries and unclassified genders.

Frequentist Approach:

For our frequentist observations, the goal was to determine whether or not there was a significant difference between genders in salary. However, other than gender, we have realistically two other variables in consideration: school and job title. For example, say we wanted to observe the distributions of the instructor salaries versus the distribution of department chair salaries. The histogram for the instructor salaries (See Appendix A) shows a distribution which barely overlaps with the values found in the histogram from the department chair salaries (See Appendix B). Instructor salaries ranged from \$25,020.00 to \$103,649.00, and department chair salaries were spread out from \$37,666.00 to \$213,504.00. We proceeded to conduct a series of analyses which would judge the normality of the data, the overlap of the distributions, the associated variances for each position, amongst other frequentist metrics (See Appendix C for example T-Test comparing these two job-titles salary means).

Appendix F depicts a histogram with the original distribution of the raw salary data. Initially observing a large amount of data on the left, we acknowledged that the data was right skewed due to a tail of extreme values distributed towards the right. We proceeded to fit a log of the salary to better model the distribution of the data, as illustrated in Appendix G. The histogram distribution after considering the log of the salary removed the previously noted skew from the data. Although our data is still not normally distributed (according to tests of normality), it is symmetrical and a large enough sample size that we felt comfortable moving on with our frequentist analysis assuming approximate normality.

As was stated previously with our initial example of professor and department chair titles, our two other variables (school and job title) do appear to be significant factors when considering salaries. For this reason, we wanted to look into salary based on gender, both before and after considering these other variables. The results agreed with our initial suspicions. Looking at just gender first (Appendix D) the T-Test results suggest that there is a significant association of gender with the log of salary. Looking at the corresponding histogram, it fully appears that males have a larger presence in the skewed ends, which are seen as the tails of the salary histograms. Furthermore, the pooled difference

confidence interval displayed that the 95% confidence interval does not contain the value of 0, suggesting that there is a significant difference when looking at gender. Having noted this, we then wanted to look at the significance of gender when looking at the residuals remaining after modeling on our other two accessible variables: employer and job title. We observed the R-Square statistic of 0.5022 to show that both terms were indeed significant.

After modeling log salary using school and job title using a simple generalized linear model, our resulting residuals (observed-exp(predicted)) appear to be approaching a bit more normalized state as seen in Appendix H. However, many considered transformations would complicate future modeling issues and we could not do yet another log transformation due to the negative values. Since we are working with a fairly large dataset, we decided to move forward without any further transformations. When considering a T-Test on the residuals and looking at gender (See Appendix E), we note that there still exists a significant difference between gender, although it did decrease. Prior to considering the other two variables, the mean difference in salary was approximately \$14,000 and afterwards it was approximately \$6,500. While this analysis is by no means exhaustive (perhaps something like years in position could just as well explain the differences in salary that gender is currently doing) it does suggest a potential for there still being a salary gender gap in Illinois' public employees in higher education. Thus concludes our frequentist analysis. In the next section, we shall consider what results ensue when instead using a bayesian approach on the residuals after considering school and job title on salary.

Bayesian Approach:

Unfortunately, since we needed to consider the residuals of salary after considering two variables specific to our dataset, we had a difficult time in finding informative priors. We therefore decided to cover our grounds by considering many different priors in order to test for the sensitivity of our model. Specifically, we considered 15 different priors for each model. We use the open source program OpenBUGS⁴ to model the evolution of our model through different priors and to collect information about our posterior distributions.

⁴Lunn, Spiegelhalter (2009) Creators of OpenBugs

Since the question at hand, as mentioned in the introduction, was to determine whether or not there was a significant difference in mean salary of females and males, we modeled the two genders separately. After doing so, assuming the diagnostics were acceptable, we could then consider both of their 95% credible intervals to see if they overlapped in our analysis of 15 different priors, we performed a total of 100,000 iterations. We shall first proceed by discussing the sensitivity and diagnostics of these models followed by considerations towards the results themselves.

Bayesian Approach (Sensitivity and Diagnostics):

In order to address the issue of model sensitivity, we used a series of priors that were constructed from the initial mean observed on the male salary of the dataset. We made changes to the parameters, as addressed in the previous section. All 15 priors that were used, resulted in completely different values, and were not similar the whole time.

Initially, we ran models for the tau parameter, on both the male and female salary datasets. Based on the BGR plots, we saw that all four plots converged rather quickly, with the example of a male burn in at 10,000 iterations. Given the various priors, we saw that the data was centered near the estimate of the frequentist data's mean.

The differences between the parameters both went to the expected values from the mean results of male and female salaries, and will be talked about in the following section.

Bayesian Approach (Results):

Before commenting on the results for the mean differences in the residuals present between genders, we shall first discuss the results for tau squared. After burning the first 5,000 iterations, the resulting 95% confidence intervals for tau squared for females and males (Appendix 0 and Appendix Q respectively) are the following: Female CI: (1.759e-9,1.839e-9), Male CI: (1.11e-9,1.207e-9). These credible intervals are far from overlapping, suggesting that there is very likely a difference in the variance of residuals based on gender. However, this is a somewhat unsurprising result. This is due to the combination of two things: first, there is a significant difference in the variances between job titles and secondly, that there is an association

with job title with gender. It therefore makes sense that this difference in variation would carry over to when considering the residuals segregated by gender.

Next, we shall consider μ for males and females. The resulting 95% confidence intervals for μ for females and males (Appendix K and Appendix M respectively) are the following: Female CI: (254.0,1591.0), Male CI: (6721.0,8294.0). Furthermore, their estimated means, from the same Appendices, were the following: 925.5 and 7507. This is consistent with the frequentist approach as, once again, the expected difference is approximately 6,500. We therefore claim that this is a significant difference due to the fact that the credible intervals come nowhere near overlapping with each other.

It is worth noting that neither of the credible intervals contain the value of zero for the estimates of μ for both females and males. This could suggest that modeling using school and job titles is not sufficient. However, this is an unsurprising result as things such as years of experience as specific department would be obvious candidates for further predicting salaries.

Closing Remarks:

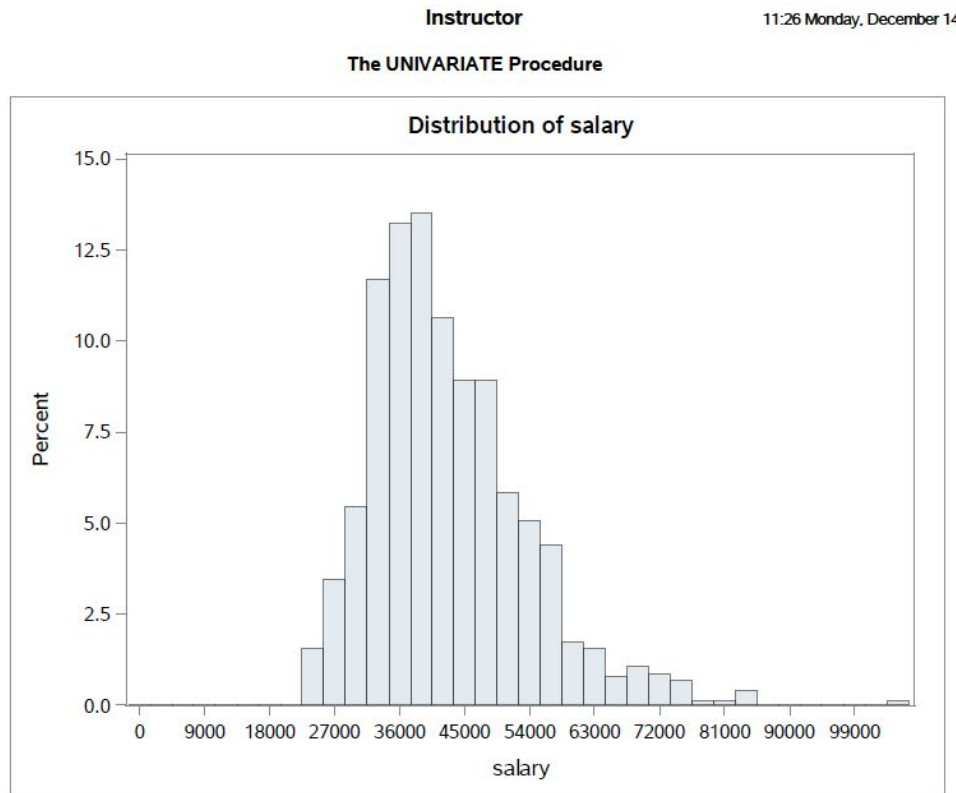
In closing, our analysis shows that there is reason to believe that there is some gender wage gap present within Illinois' higher education employees. Furthermore, this result is extremely similar independent of whether a frequentist or bayesian approach is utilized. However, it is worth noting that the number of informative covariates accessible in our analysis was limited: more specifically, only 2 categorical variables plus gender, but only after appending this variable utilizing the employee's first name.

Works Cited

- Better Government Association, Educational Employee Payroll Database (2014)
- Cowles, Mary K. (2015) Applied Bayesian Statistics- With R and OpenBUGS. Springer.
- Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions, *Journal of the Royal Statistical Society, Series B*, 71(2), 3049-3067
<http://www.springer.com/us/book/9781461456957>
- U.S. Census Bureau, AAUW, By the Numbers: A Look at the Gender Pay Gap (2014)
- Wickham, Hadley. babynames: US baby names 1880-2013 (2014)

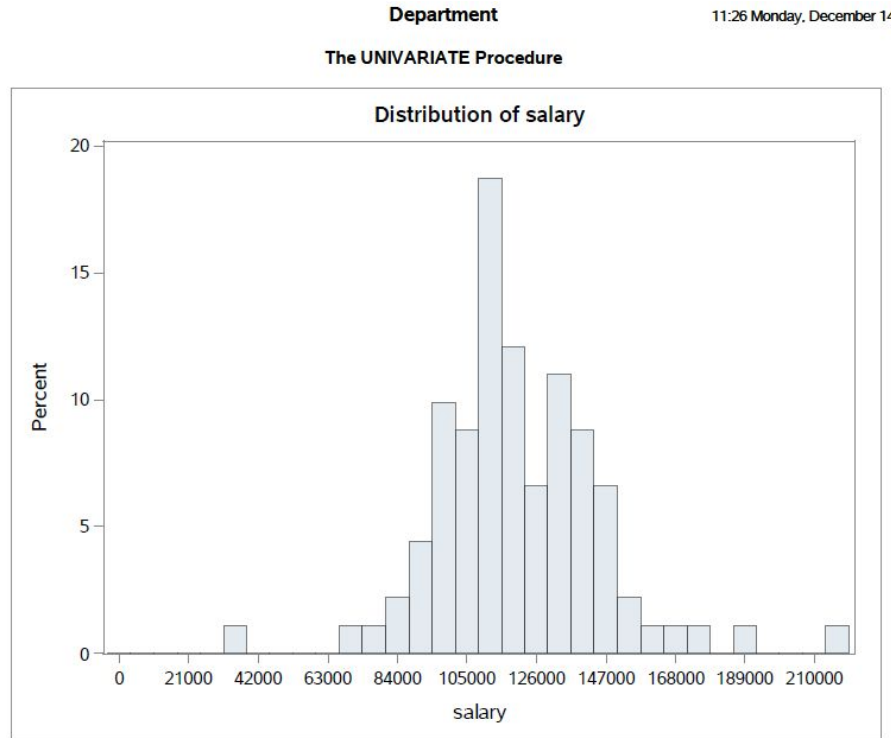
Appendices

Appendix A:



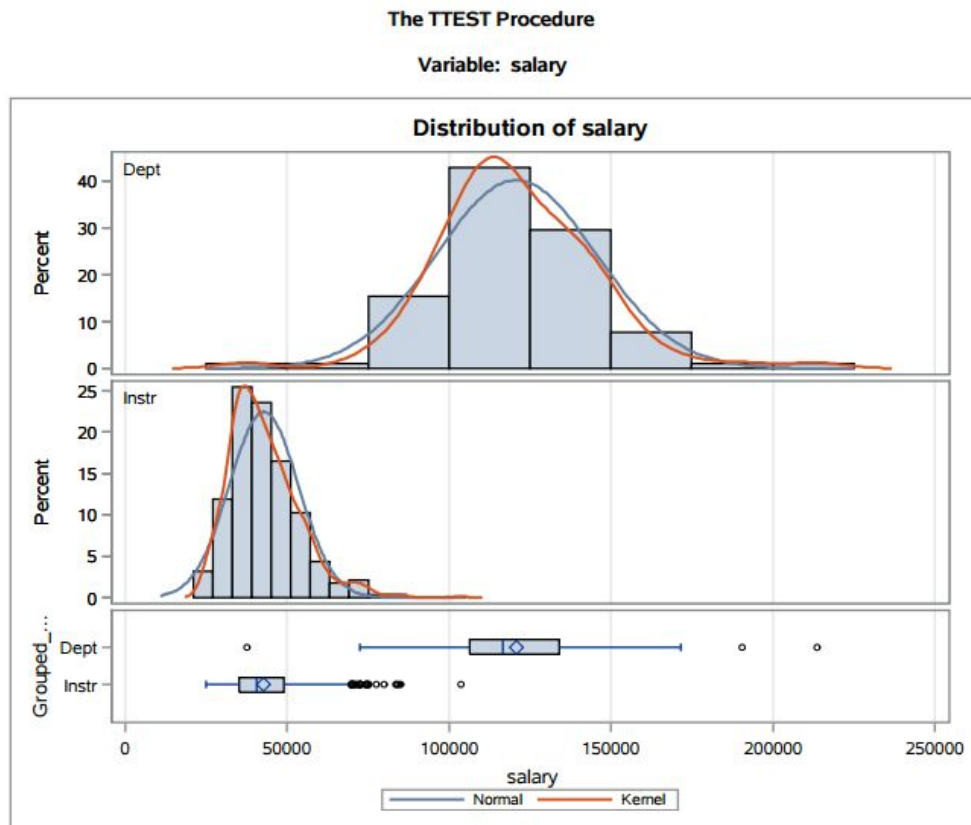
Histogram depicting the distribution of salaries for Illinois state school employees with the title of “Instructor”

Appendix B:



Histogram depicting the distribution of salaries for Illinois state school employees with the title of “Departmental Chair”

Appendix C:



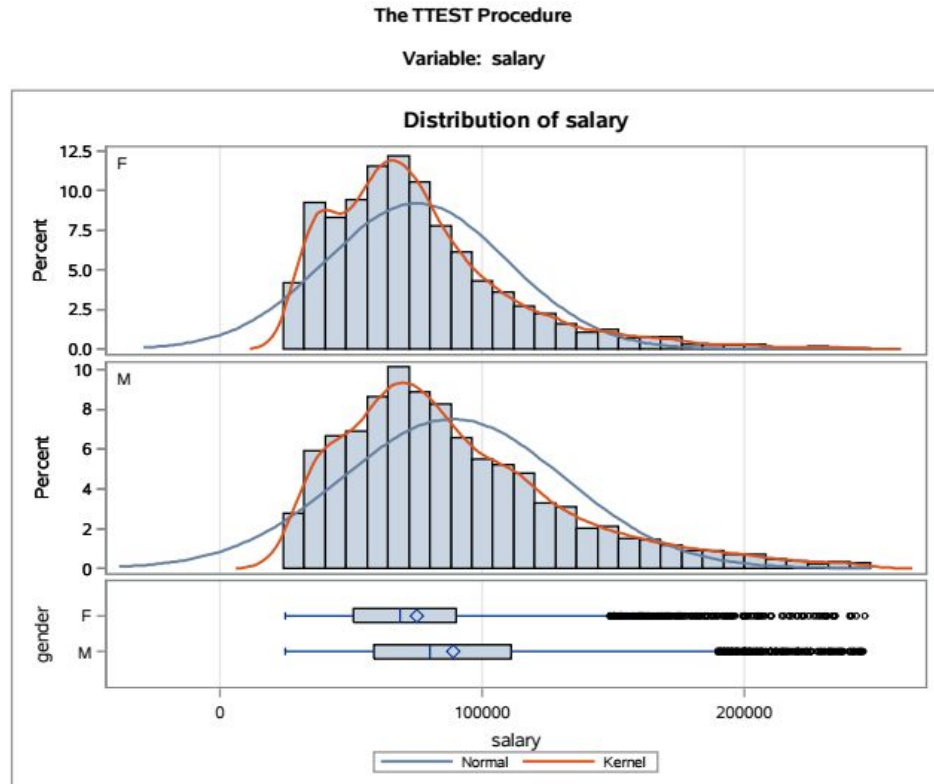
Grouped_Title	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Dept		120758	115591	125925	24808.7	21653.8	29048.2
Instr		42832.8	42185.8	43479.8	10648.9	10210.7	11126.6
Diff (1-2)	Pooled	77925.1	75269.6	80580.6	12382.1	11892.4	12914.1
Diff (1-2)	Satterthwaite	77925.1	72719.3	83130.9			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1132	57.58	<.0001
Satterthwaite	Unequal	92.915	29.73	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	90	1042	5.43	<.0001

General distributions depicting the of salaries for Illinois state school employees with the title of “Instructor” and “Departmental Chair”

Appendix D:



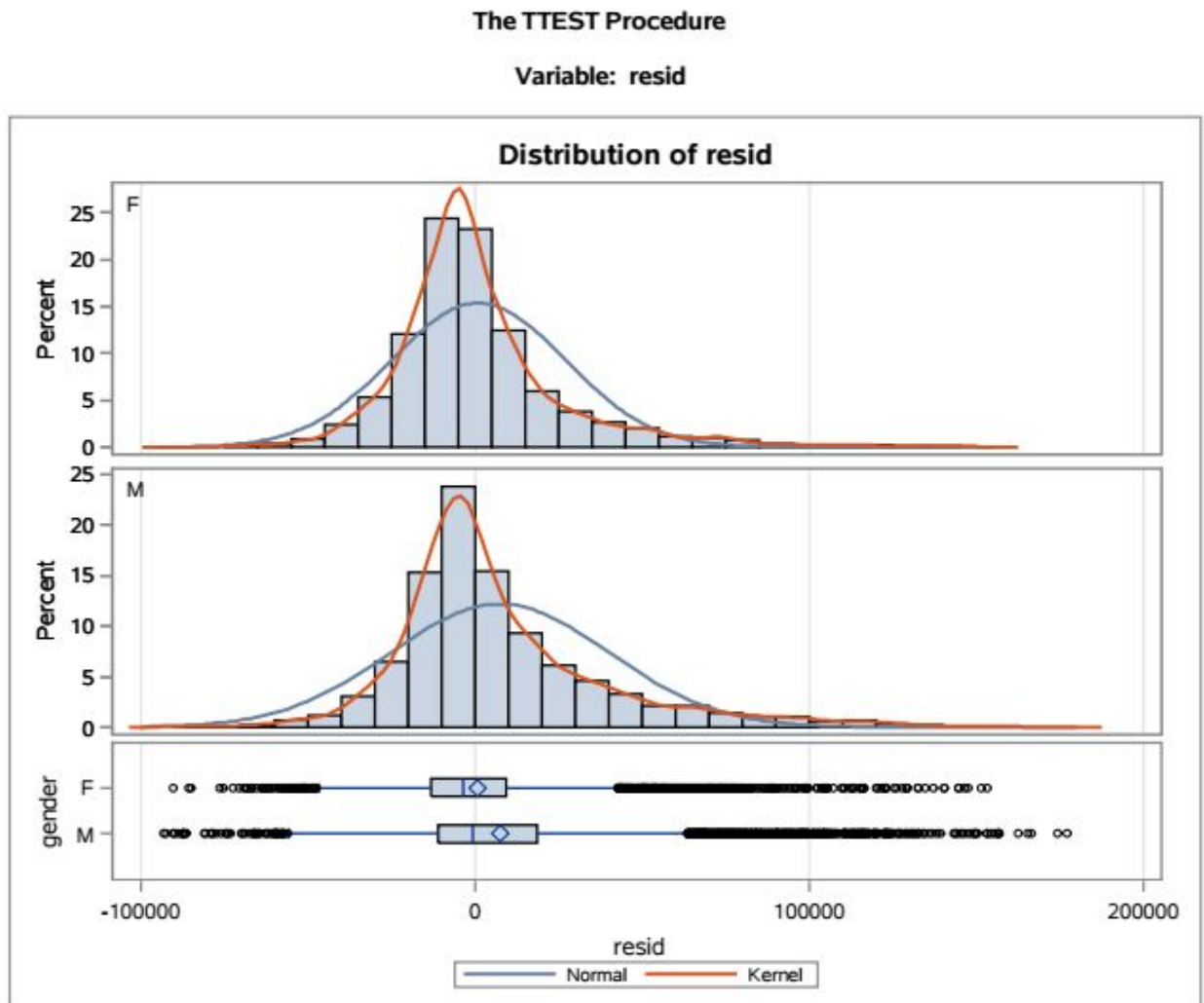
gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		75000.4	74115.2	75885.6	34709.7	34095.0	35347.1
M		88796.0	87777.6	89814.4	42432.1	41724.1	43164.6
Diff (1-2)	Pooled	-13795.6	-15161.1	-12430.1	38995.7	38519.8	39483.7
Diff (1-2)	Satterthwaite	-13795.6	-15144.8	-12446.4			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	12578	-19.80	<.0001
Satterthwaite	Unequal	12500	-20.04	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6670	5908	1.49	<.0001

Histograms depicts the distribution of salaries after accounting for gender. Pooled t-test results also suggest there being a significant difference in salaries based on gender.

Appendix E:



Description on next page.

Appendix E (continued) :

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		926.9	263.9	1590.0	25962.8	25502.4	26440.3
M		7507.5	6721.8	8293.2	32695.6	32149.4	33260.7
Diff (1-2)	Pooled	-6580.6	-7622.8	-5538.3	29724.5	29361.2	30096.9
Diff (1-2)	Satterthwaite	-6580.6	-7608.5	-5552.6			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	12545	-12.38	<.0001
Satterthwaite	Unequal	12402	-12.55	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6654	5891	1.59	<.0001

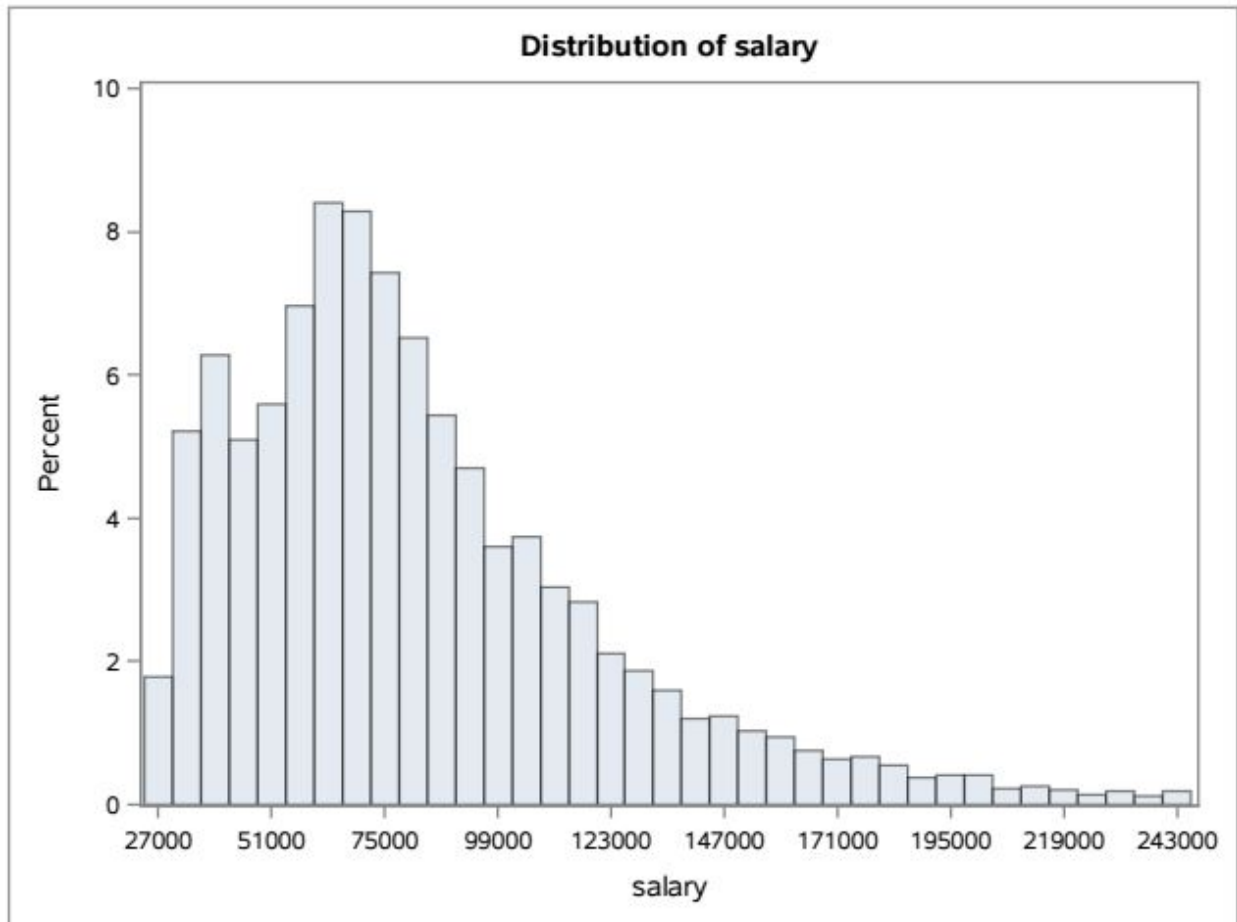
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	135331140552	135331140552	153.17	<.0001
Error	12545	1.1084053E13	883543512.5		
Corrected Total	12546	1.1219385E13			

R-Square	Coeff Var	Root MSE	resid Mean
0.012062	672.9119	29724.46	4417.288

Source	DF	Anova SS	Mean Square	F Value	Pr > F
gender	1	135331140552	135331140552	153.17	<.0001

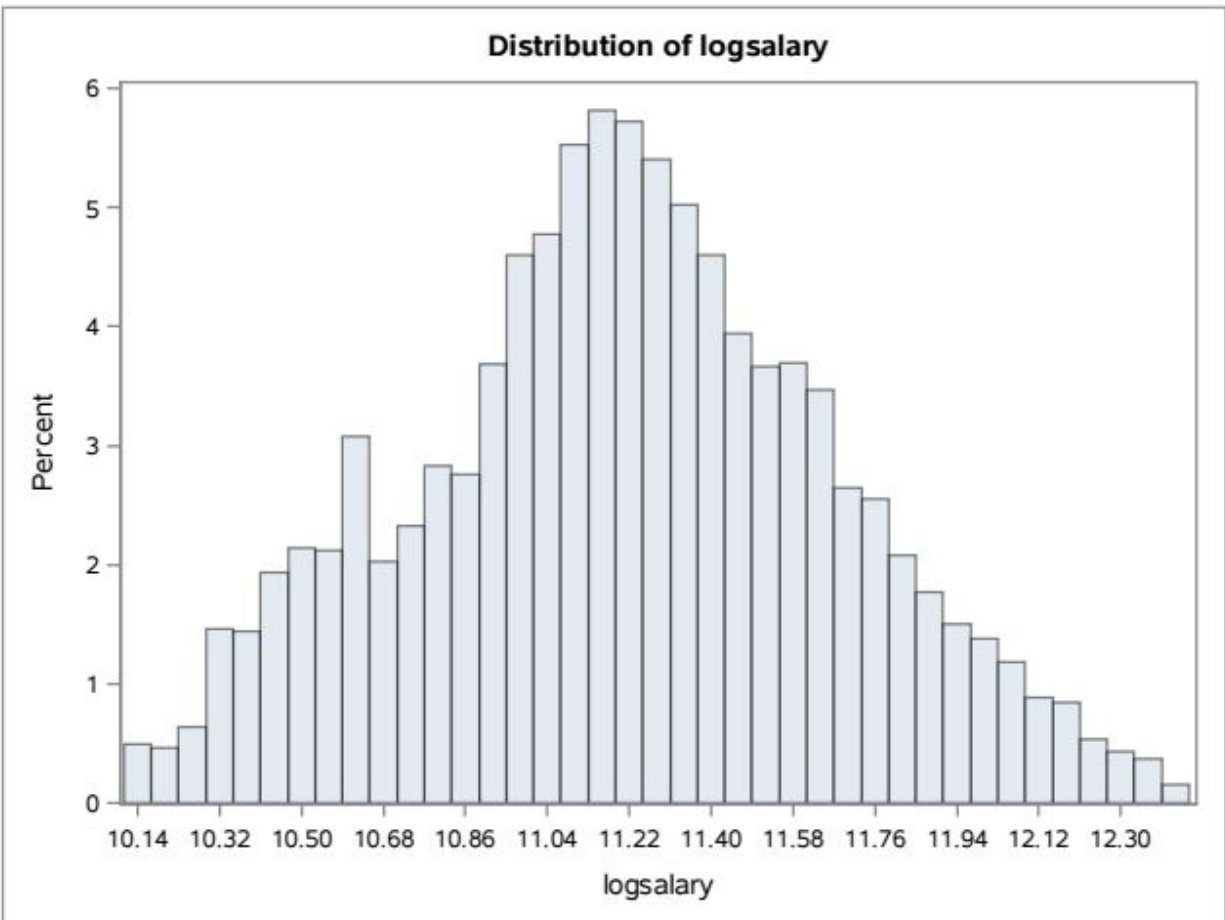
These are the results from performing both a t-test as well as an ANOVA on the residuals of salary after modeling both school and job title. Note: Modeled the log of salaries and then transformed predictions back to dollar values. Therefore, these analysis were done on non-transformed data.

Appendix F:



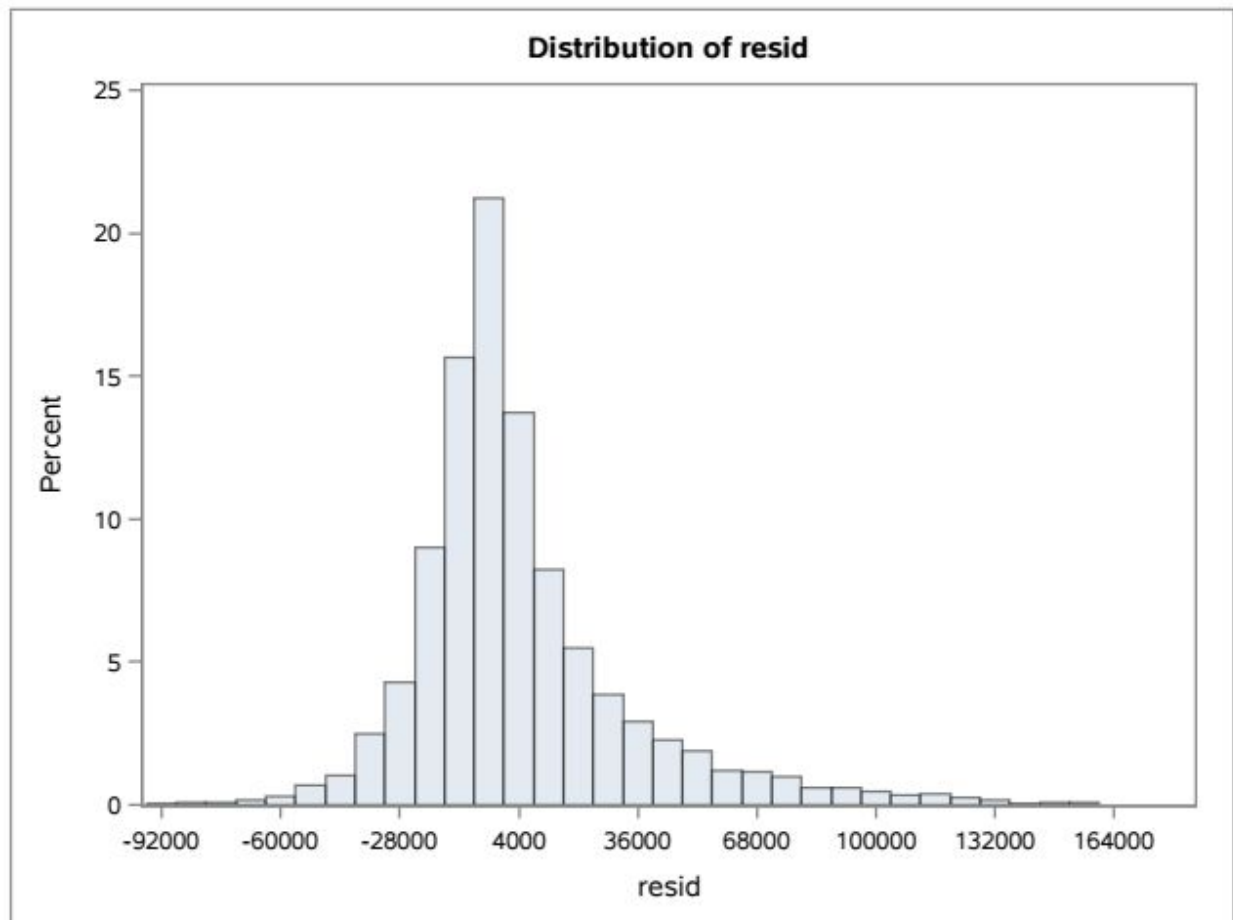
Histogram depicting the distribution of salaries for Illinois state school employees, highlighting on the skew of the data to the right.

Appendix G:



Histogram depicting the log distribution of salaries for Illinois state school employees.

Appendix H:



This is the distribution of the residuals after modeling school and job title. This was done by taking actual and subtracting $\exp(\text{predicted})$.

Appendix I:

Effects:	Intercept school Grouped_Title
-----------------	--------------------------------

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	30	1350.69627	45.02321	420.94
Error	12516	1338.69770	0.10696	
Corrected Total	12546	2689.39398		

Root MSE	0.32705
Dependent Mean	11.21178
R-Square	0.5022
Adj R-Sq	0.5010
AIC	-15466
AICC	-15466
SBC	-27785

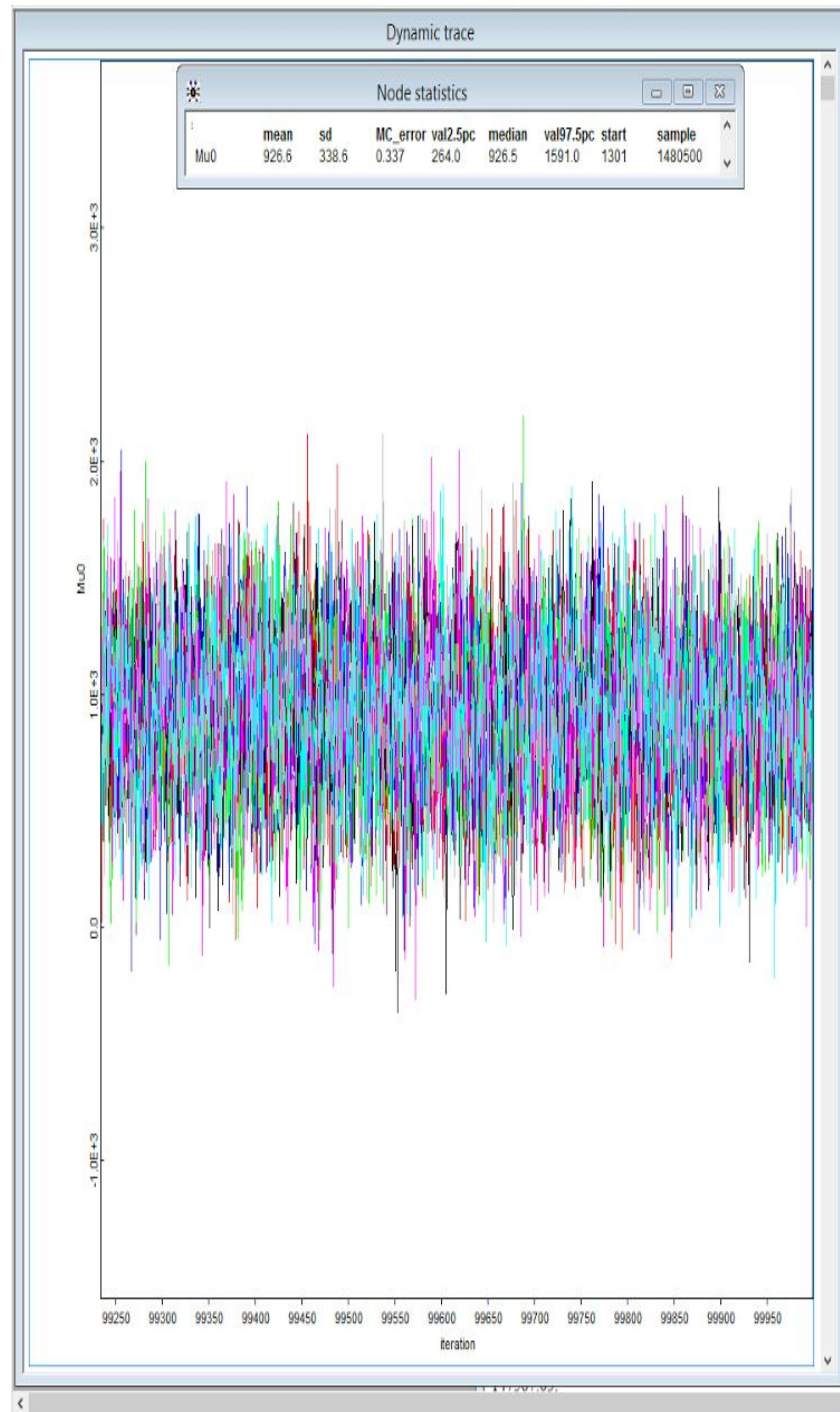
Results correspond to the GLM using school and job title. As can be seen from the F-Value, the model is clearly significant. These variables also appear to do a fair bit of explaining the variation in the data as can be seen by the R-Square of approximately 0.5.

Appendix J:

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Grouped_Title Dean	1	1.204016	0.036747	32.76
Grouped_Title Dept	1	0.992473	0.039540	25.10
Grouped_Title Dir	1	0.586788	0.017292	33.93
Grouped_Title Exec	1	0.841533	0.034917	24.10
Grouped_Title Instr	1	-0.073458	0.018711	-3.93
Grouped_Title Lectu	1	-0.060540	0.020785	-2.91
Grouped_Title No Ra	1	-0.268597	0.190036	-1.41
Grouped_Title Other	1	0.485998	0.020923	23.23
Grouped_Title Part-	1	-0.081546	0.327707	-0.25
Grouped_Title Postd	1	-0.212439	0.020757	-10.23
Grouped_Title Prof	1	0.787811	0.015957	49.37
Grouped_Title Res	1	0.191530	0.026201	7.31
Grouped_Title Sr Di	1	0.893675	0.058641	15.24
Grouped_Title Sr Le	1	-0.122953	0.031583	-3.89
Grouped_Title Tch	1	-0.049848	0.048794	-1.02
Grouped_Title VP	1	1.260631	0.045486	27.71
Grouped_Title Vst	0	0	.	.

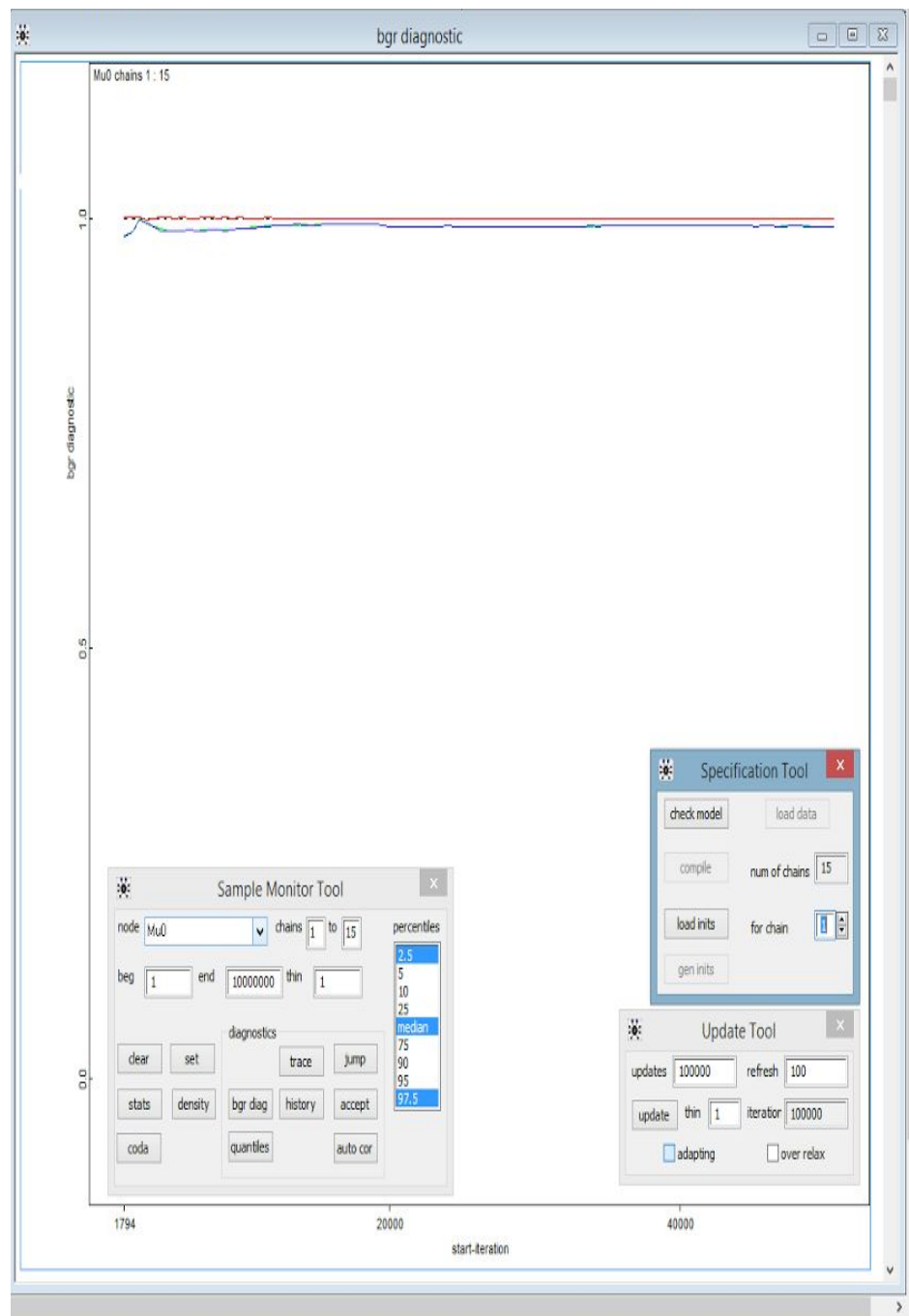
Chart depicting the different titles of employees in the state of Illinois, and their corresponding statistics for comparison between positions.

Appendix K:



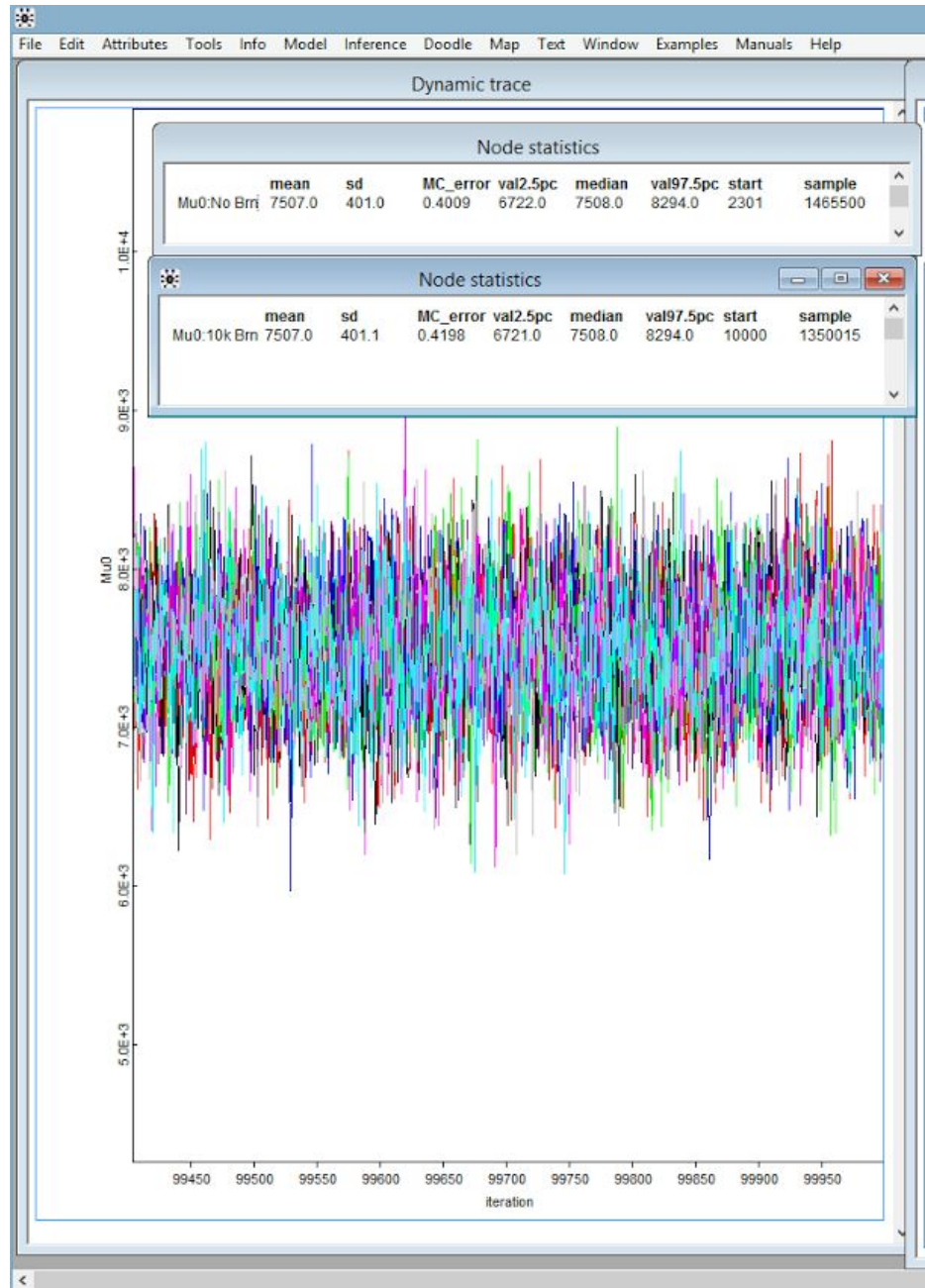
This is the Dynamic Trace plot at the end of 100,000 iterations for estimates of μ_0 for females. As can be seen by the chart itself, it appears that all 15 different priors converged towards the same values.

Appendix L:



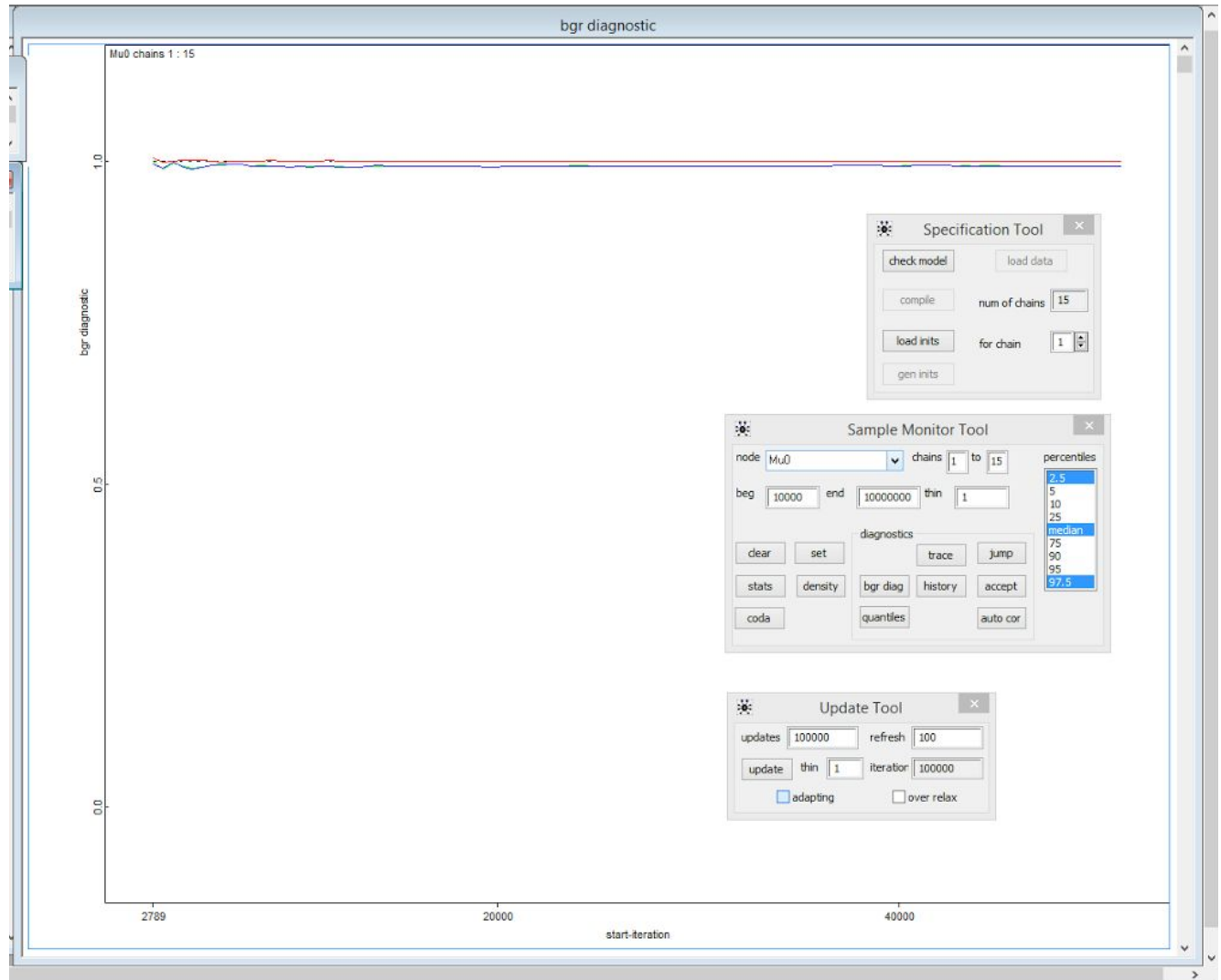
This is the BGR plot at the end of 100,000 iterations for estimates of μ_0 for females. As can be seen from this chart, there is an extremely quick convergence. We therefore did not find it necessary to have a burning set before compiling the model statistics.

Appendix M:



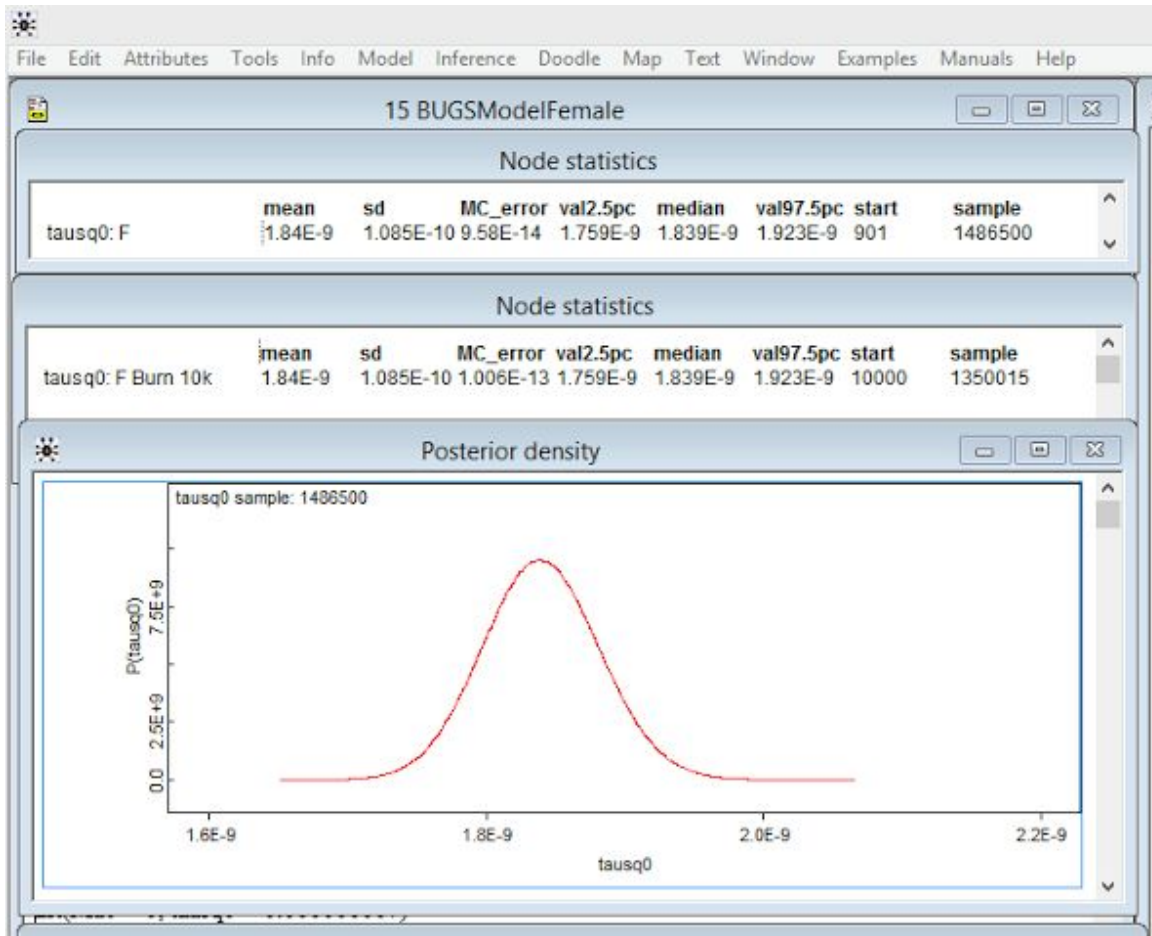
This is the Dynamic Trace plot at the end of 100,000 iterations for estimates of μ_0 for males. Also, in this screenshot the stats of the simulation can be seen. Furthermore, from the chart itself, it appears that all 15 different priors converged towards the same values.

Appendix N:



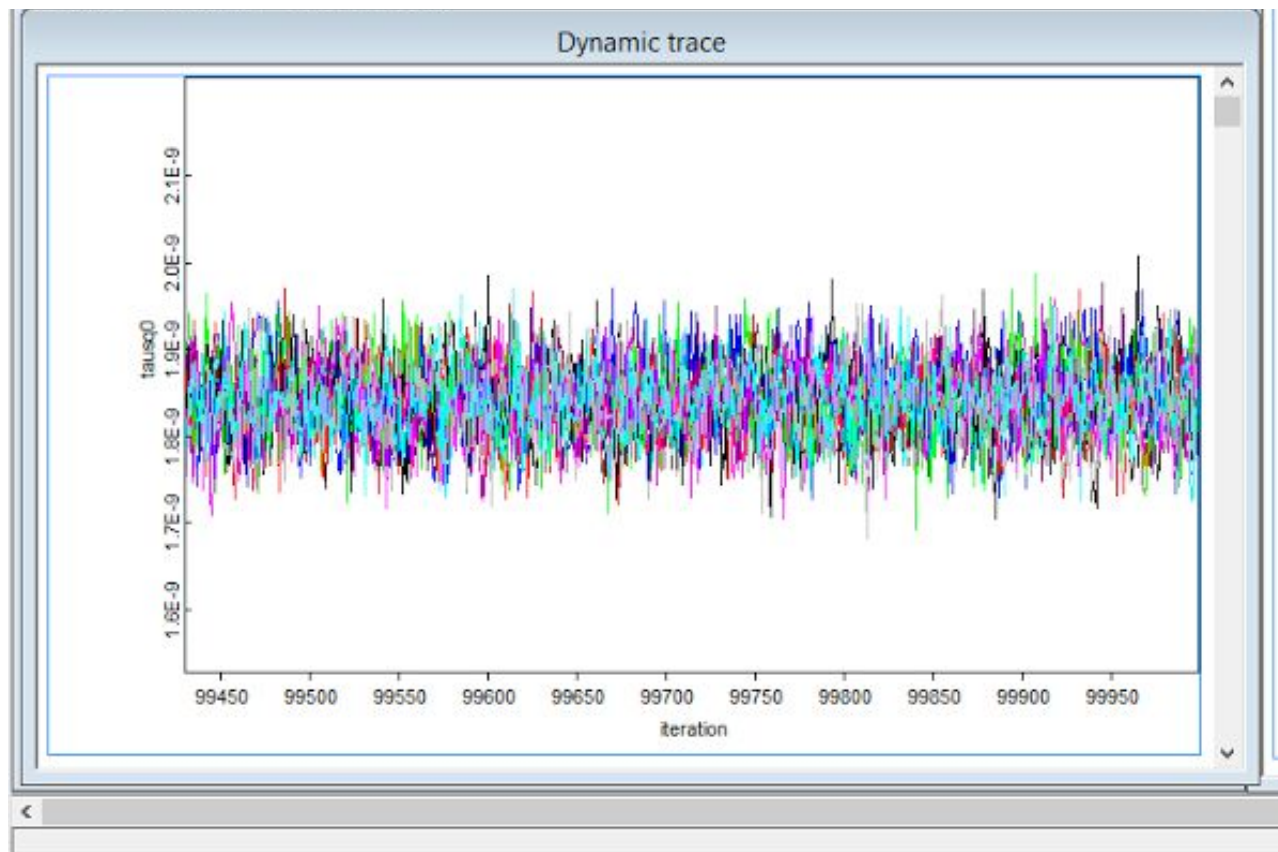
This is the BGR plot at the end of 100,000 iterations for estimates of μ_0 for males. Although this plot also suggests a very quick convergence, we decided to burn 10,000 values due to the slight ebb and flow at some of the beginning values.

Appendix O:



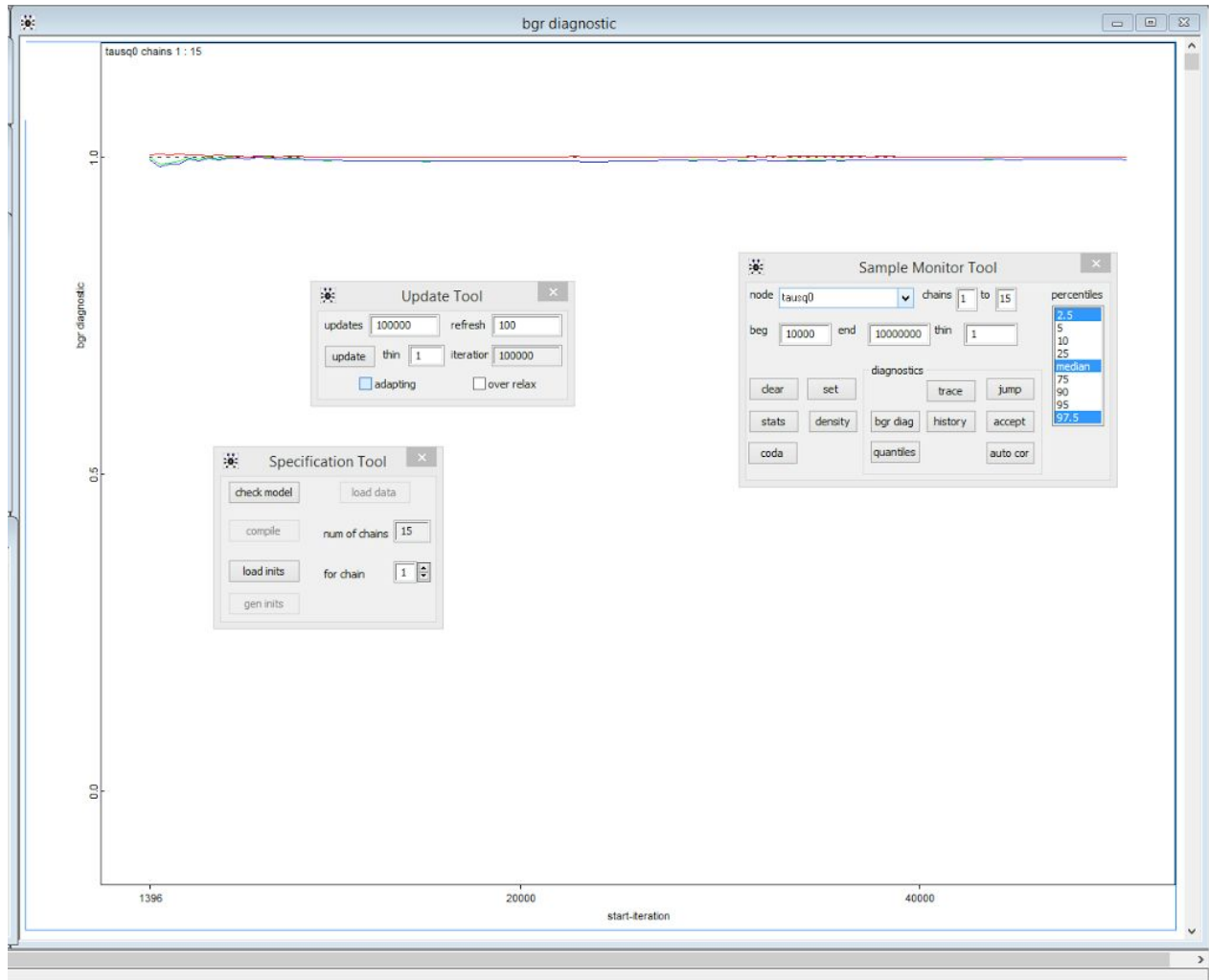
This is the posterior density at the end of 100,000 iterations for estimates of Tausq0 for females as well the summary statistics both before and after burning the first 5,000 iterations.

Appendix P:



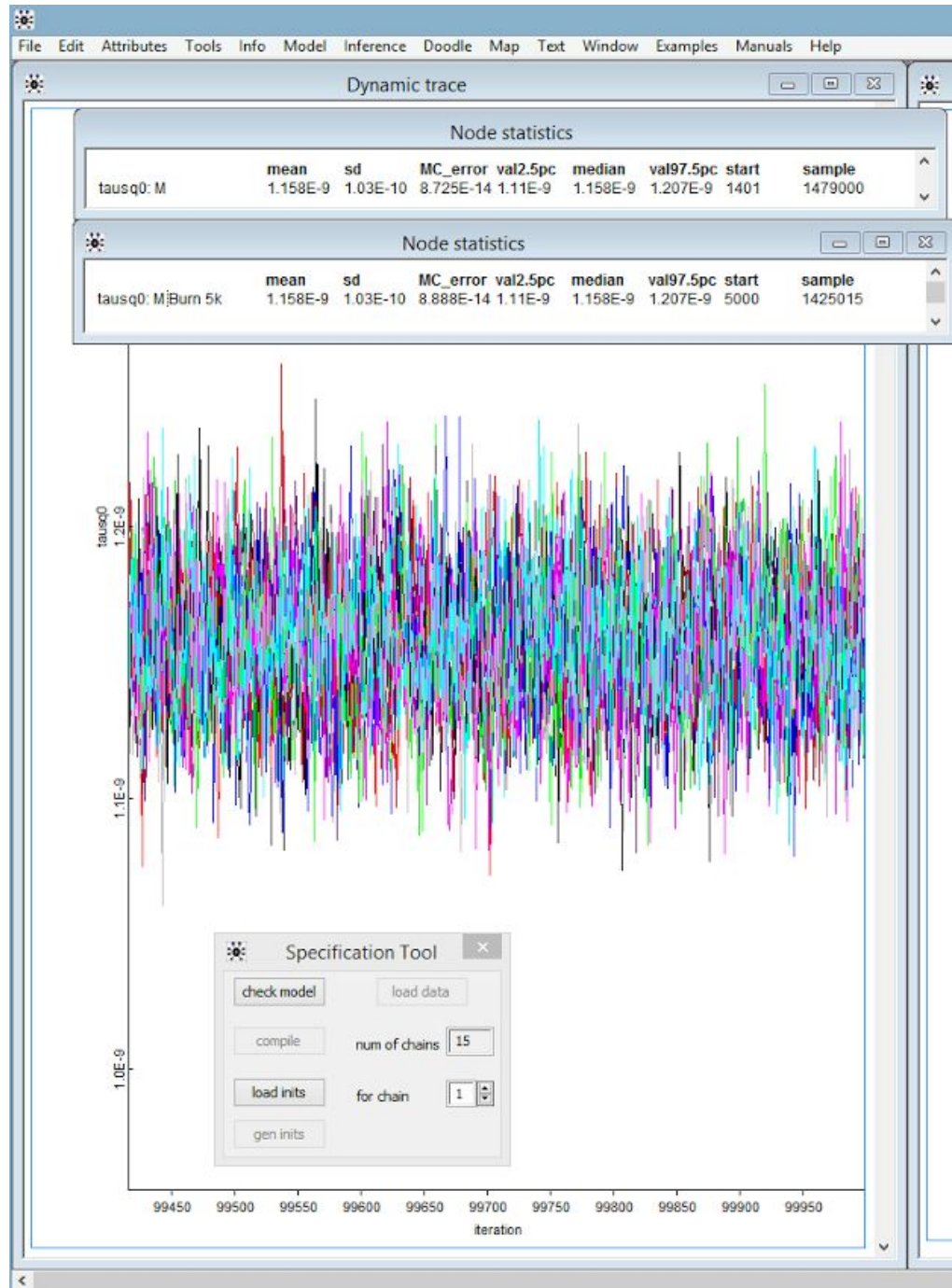
This is the Dynamic Trace plot at the end of 100,000 iterations for estimates of Tausq0 for females. From the chart itself, it appears that all 15 different priors converged towards the same values.

Appendix Q:



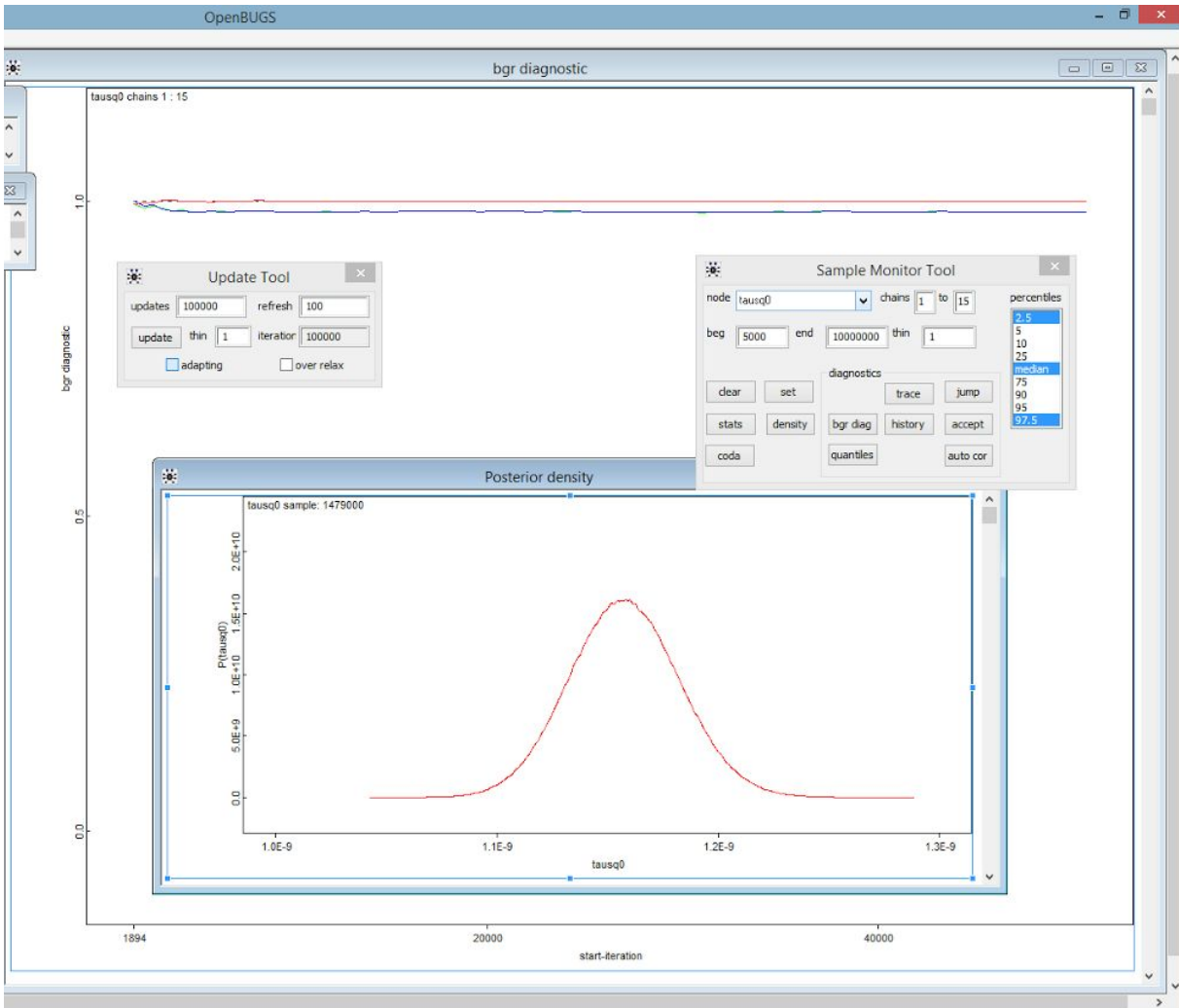
This is the BGR plot at the end of 100,000 iterations for estimates of Tausq0 for females. This plot also suggests a very quick convergence, we decided to burn 5,000 values due to the delay in convergence.

Appendix Q:



This is the Dynamic Trace plot at the end of 100,000 iterations for estimates of Tausq0 for males. Also, in this screenshot the stats of the simulation can be seen. Furthermore, from the chart itself, it appears that all 15 different priors converged towards the same values.

Appendix Q:



This picture contains both the BGR plot as well as the posterior density at the end of 100,000 iterations for estimates of Tausq0 for males. The BGR plot also suggests a very quick convergence but we decided to burn 5,000 values to get rid of the first few larger values.

Appendix R:

0 data_scrapping.py

```
from lxml import html
import requests, csv, sys

sys.stdout=open("stat_430_project_data.txt","w")
initial_page =
'http://www.bettergov.org/payroll-database?combine=&Employer=All%20Higher%20Ed%20Employees&DataYear=2014&Salary[min]=0&Salary[max]=1000000&page=0'
strings = []

for i in range(0,366):
    strings.append(initial_page+str(i))
for i in range(0,len(strings)):
    page = requests.get(strings[i])
    tree = html.fromstring(page.content)
    first_name = tree.xpath('//td[@class="views-field views-field-FirstName"]/text()')
    last_name = tree.xpath('//td[@class="views-field views-field-LastName"]/text()')
    salary = tree.xpath('//td[@class="views-field views-field-Salary active"]/text()')
    title = tree.xpath('//td[@class="views-field views-field-Title"]/text()')
    Employer = tree.xpath('//td[@class="views-field views-field-Employer"]/text()')
    department = tree.xpath('//td[@class="views-field views-field-Department"]/text()')
    DataYear = tree.xpath('//td[@class="views-field views-field-DataYear"]/text()')

    if (len(first_name) == len(salary) == len(last_name) == len(title) == len(Employer) == len(department) == len(DataYear)):
        for i in range(len(first_name)):
            print(first_name[i] + last_name[i] + salary[i] + title[i] + Employer[i] + department[i] + DataYear [i])
```

4 Data Prep Code File.sas

```
/* Once below 2 pathnames are updated, you can run this entire file to generate all relevant data preparation and final dataset used for eventual analysis */

/*Update pathname1 to be the path location of the dataset file that was provided under the name "1 dataclean"*/
%let pathname1=C:\Users\tmthiel2\Downloads\2 dataclean.csv;

/*Update pathname2 to be the path location of the dataset file that was provided under the name "babynamescsv"*/
%let pathname2=C:\Users\tmthiel2\Downloads\3 babynamescsv.csv;

/* Brief description of each dataset created from this file*/
/*Final Data: Final_dataset_dropped_low_sal*/
/*Babynames*/
/* Read in babynames csv into sas*/
/*Babynames2/Babynames3*/
/* Intermediate steps to get from Babynames to Babynames4*/
/*Babynames4*/
/* Condensed Babynames to have only one observation per unique name. Contains counts of each gender.*/
/*Dataset*/
/* Dataset that was created after condensing each set of 8 rows to 1 from pathname1 dataset. This creates one observation per employee.*/
/*Data_w_gender*/
/* This is the same as (Dataset) but after having appended gender.*/
/*Final_dataset*/
/* Data is done being cleaned. All variables are in final form.*/
/*Final_dataset_dropped_low_sal*/
/* Same as (Final_dataset) but after having dropped observations with salaries less than $25,000*/
/*Merged*/
/* Result of merging (Dataset) and (Babynames4) based on first names.*/
/*Rawdata*/
/* Dataset that was read in from pathname1.*/
/*Rawdata2*/
/* Intermediate step from (Rawdata) to (Dataset)*/
/*Testing*/
/* Just looking at observations with salary<$10,000.*/
/*Testprop*/
/* Where thresholds are set to append gender. Final decision was, % of 1 gender must be>90% and must have at least 5 occurrences */
/* Here we read in the two provided csvs that were mentioned above in the 2 pathnames */
proc import datafile="&pathname1"
    out=rawdata
    dbms=csv;
    getnames=no;
run;
```



```

proc import datafile="&pathname2"
    out=babynames
    dbms=csv;
    getnames=yes;
run;

/* We then create an indicator to determine which variable we currently have stored */
/* x1-x8 all have a value of 0 every 8th observations */
data rawdata2;
    set rawdata;
    n=_n_;
    x1=mod(n+7,8);
    x2=mod(n+6,8);
    x3=mod(n+5,8);
    x4=mod(n+4,8);
    x5=mod(n+3,8);
    x6=mod(n+2,8);
    x7=mod(n+1,8);
    x8=mod(n,8);
run;

/* Using the indicators from the previous dataset, we retain all the values until */
/* reaching the eighth observations. At which point, we then output a single observation */

data dataset;
    set rawdata2;
    retain first_name last_name salary title employer university year ;
    if x1=0 then first_name=var1;
    if x2=0 then last_name=var1;
    if x3=0 then salary=var1+0;
    if x4=0 then title=var1;
    if x5=0 then employer=var1;
    if x6=0 then university=var1;
    if x7=0 then year=var1;
    if x8=0 then output;
    keep first_name last_name salary title employer university year ;
run;

/* The first row is just an empty set. Here we drop observation 1 */
data dataset;
    set dataset (firstobs=2);
run;

/*      Examples of someone with non realistic salary      */
/*      We believe this is due to part time employees      */
/* Later on, this will be addressed by dropping observations that do not meet a certain threshold */

```

```

data testing;
  set dataset;
  if salary<10000;
run;

/* Based on this histogram, 25,000 dollars appears to be a reasonable cutting off point.
Any observations with salaries below this we will assume are part time employees and therefore will be ignored */
proc univariate data=dataset ;
  var salary;
  histogram;
  where salary<100000;
run;

/* We create a variable to sort by that is the concatenation of name and gender */
data babynames2;
  set babynames;
  asdf=cat(name,sex);
run;

proc sort data= babynames2;
  by asdf;
run;

data babynames2;
  set babynames2;
  lag=lag(asdf);
  id=_n_; /* Id was created so that we can return to the original order after flipping the dataset */
run;

proc sort data= babynames2;
  by descending id;
run;

/*Looking at the previous value of an upside down dataset is looking at one observation in the future*/
data babynames2;
  set babynames2;
  fut=lag(asdf);
run;

/* return dataset to original order */
proc sort data= babynames2;
  by id;
run;

/* Aggregate the counts of a given gender*name combination. Only output once total sum has been calculated */
data babynames3;

```

```

set babynames2;
retain sum 0;
sum=sum+n;
if fut ne asdf then do;
    output;
    sum=0;
end;
keep sex name sum asdf id;
run;

/* Same process as before, but now we need the value of sex in the future observation */
proc sort data= babynames3;
    by descending id;
run;

data babynames3;
    set babynames3;
    futname=lag(name);
    futsex=lag(sex);
run;

proc sort data= babynames3;
    by id;
run;

/* Same process as before, but now we are aggregating each pair of observations
for different genders and calculating the number of people with a given name as
well as the proportion of which were male */

data babynames4;
    set babynames3;
    retain proportion 0 malecount 0 femalecount 0;
    if sex="M" then malecount=sum;
    else femalecount=sum;
    if futname ne name then do;
        proportion=malecount/(femalecount+malecount);
        totalcount=femalecount+malecount;
        output;
        malecount=0;
        femalecount=0;
        proportion=0;
    end;
    keep totalcount femalecount malecount proportion name;
run;

```

```

/* we create a new names variable to ensure that merging is done properly */
data babynames4;
  set babynames4;
  trimmed_names_upcase=upcase(trim(name));
run;

/* we create a new names variable to ensure that merging is done properly */
data dataset;
  set dataset;
  trimmed_names_upcase=upcase(trim(first_name));
run;

/*We then sort the reference table (babynames4) as well as our dataset(dataset)
by the new name variable. We can then merge the tables*/
proc sort data=dataset;
  by trimmed_names_upcase;
run;

proc sort data=babynames4;
  by trimmed_names_upcase;
run;

/* Merge gender based on name. Only keep observations that came from dataset.
E.g., drop observations that consist of only variables from the reference table
since there may have been no match */
data merged;
  merge dataset (in=x) babynames4 (in=y);
  by trimmed_names_upcase;
  if x=1;
run;

/* Each observation now has corresponding proportions male associated with their
given name. We then only assign a gender if the total count>4 and the proportion
corresponds to at least 90% of the given gender. Otherwise, gender receives a value
of unknown */
data testprop;
  set merged;
  length gender $10;
  if proportion>.9 and totalcount>4 then gender="M";
  else if proportion<.1 and totalcount>4 then gender="F";
  else gender="Unknown";
run;

```

```

/* Just keep relevant variables*/
data data_w_gender;
  set testprop;
  keep first_name last_name salary title employer university year gender;
run;

/* Counts of each gender */
proc freq data = data_w_gender;
  table gender;
run;

/* Reduce dimensionality for variables: title and university */
/* we do not override the original values but create new variables */
/* This way, one can still view the given employee's full title */
data Final_dataset;
  set data_w_gender;
  oldtitle=title;
  title=tranwrd(title, "Faculty ", "");
  title=tranwrd(title, "Continuing Ed ", "");
  title=tranwrd(title, "University ", "");
  title=tranwrd(title, "Academic ", "");
  title=tranwrd(title, "Full-Time", "");
  title=tranwrd(title, "Full Time", "");
  title=tranwrd(title, "Full", "");
  title=tranwrd(title, "Professor", "Prof");
  title=tranwrd(title, "Associate", "Assoc");
  title=tranwrd(title, "Assistant", "Asst");
  title=tranwrd(title, "Clinical", "Clin");
  title=tranwrd(title, "Adjunct", "Adj");
  title=tranwrd(title, "Visit", "Vst");
  title=tranwrd(title, "Visiting", "Vst");
  title=tranwrd(title, "Senior", "Sr");
  title=tranwrd(title, "Tech", "Tch");
  title=tranwrd(title, "Technical", "Tch");
  title=tranwrd(title, "Instructor", "Instr");
  title=tranwrd(title, "Director", "Dir");
  title=tranwrd(title, "Advsr", "Advis");
  title=tranwrd(title, "Devolopment", "Devlp");
  title=tranwrd(title, "Devoloper", "Devlp");
  title=tranwrd(title, "Operations", "Oper");
  title=tranwrd(title, "Operator", "Oper");
  title=tranwrd(title, "Resedential", "Res");
  title=tranwrd(title, "President", "Pres");
  title=tranwrd(title, "Vice Pres", "VP ");
  title=tranwrd(title, "Vice", "VP ");

```

```

title=tranwrd(title, "V ", "VP ");
title=tranwrd(title, "Executive", "Exec");
title=tranwrd(title, "Vstin", "Vst");
title=tranwrd(title, "Lect", "Lecture");
title=tranwrd(title, "Prof'", "Prof ");
title=tranwrd(title, "Vst", "Vst ");
title=tranwrd(title, "Adj", "Adj ");
title=tranwrd(title, "Res", "Res ");
title=tranwrd(title, "Dir", "Dir ");
title=tranwrd(title, "Tch", "Tch ");
title=left(title);
testingtitle=substrn(title, 1, 5);
testingtitle=tranwrd(testingtitle, "Asst.", "Asst");
testingtitle=tranwrd(testingtitle, "Count", "Exec ");
testingtitle=tranwrd(testingtitle, "Inter", "Other");
testingtitle=tranwrd(testingtitle, "Depar", "Dept ");
title=oldtitle;
if ~(testingtitle in ("Assoc", "Asst", "Prof", "Instr", "Dir", "Lectu", "Clin", "Vst", "Adj", "Postd", "Res", "Chair", "Sr Le", "Part-", "Exec", "Dean", "No
Ra", "VP", "Other", "Tch", "Advis", "Sr Di", "Dept")) then testingtitle="Other";
Grouped_Title=testingtitle;
campus=University;
if campus="All Higher Ed Employees" then campus = "";
school=campus;
school=tranwrd(school, "Chicago", "Chicago ");
school=tranwrd(school, "Eastern", "Eastern ");
school=tranwrd(school, "Governors", "Governors ");
school=tranwrd(school, "Illinois State", "Illinois State ");
school=tranwrd(school, "Northeastern", "Northeastern ");
school=tranwrd(school, "Northern", "Northern ");
school=tranwrd(school, "Southern", "Southern ");
school=tranwrd(school, "U Of I", "U Of I ");
school=tranwrd(school, "Western", "Western ");
school=trim(substrn(school, 1, 20));
binned_salary=left(put(round(salary,1000),12.));
drop testingtitle oldtitle year Employer;
run;

/* Revisiting salary*/
proc univariate data=Final_dataset;
var salary;
histogram;
where salary<50000;
run;

```

```

/* Drop observations with values less than 25000 */
data Final_dataset_dropped_low_sal;
  set Final_dataset;
  if salary >= 25000;
run;

/*proc univariate data=Final_dataset_dropped_low_sal noprint;*/
/* var salary;*/
/* output pctlpre=P_ pctlpts= 25, 75 to 100 by 25;*/
/*run;*/

data Final_dataset_dropped_low_sal;
  set Final_dataset_dropped_low_sal;
  if ~(salary > 103000 + (3 * (103000 - 54774)));
run;

proc univariate data=Final_dataset_dropped_low_sal ;
  var salary;
  histogram;
run;

```

6 GLM.sas

```
/**/  
/*proc freq data=test;*/  
/* table grouped_title*gender / out=procfreq norow nopercnt ;*/  
/*run;*/  
  
ods pdf file="C:\Users\tmthiel2\Downloads\GLM Frequentist Approach.pdf";  
/* only looking at the observations with values of male or female for gender */  
  
data test;  
  set Final_dataset_dropped_low_sal;  
  logsalary=log(salary);  
  if gender="M" or gender="F";  
run;  
  
/* We are going to model the log of salary due to there being an obvious right skew in the data */  
proc univariate data=test normal;  
  var salary;  
  histogram;  
run;  
  
/* Still non normal, but no obvious choice for transformation (looks like a triangle)*/  
proc univariate data=test normal;  
  var logsalary;  
  histogram;  
run;  
  
Gender is last to be selected but does make it into the model after grouped_title (1st) and school (2nd)*/  
proc glmselect data=test ;  
  class school grouped_title gender;  
  model logsalary=school grouped_title gender/selection=forward details=all;  
  output out=pout predicted=predicted ;  
run;  
  
/* No variables are kicked out of model when using backward selection*/  
proc glmselect data=test;  
  class school grouped_title gender;  
  model logsalary=school grouped_title gender/selection=backward;  
  output out=pout predicted=predicted;  
run;
```



```

data get_resid;
  set pout;
  resid=salary-(exp(predicted));
run;

proc univariate data=get_resid normal;
  var resid;
  histogram;
run;

/* Get residuals after modeling grouped_title and school. These residuals will then be used for bayesian analysis */
proc glmselect data=test;
  class school grouped_title;
  model logsalary=school grouped_title/selection=backward;
  output out=pout_2 predicted=predicted;
run;

data bayes_model_this;
  set pout_2;
  resid=salary-exp(predicted);
run;

proc univariate data=bayes_model_this normal;
  var resid;
  histogram;
run;

/* Frequentist approach */
proc anova data=bayes_model_this;
  class Gender;
  model resid=gender;
run;

ODS PDF CLOSE;

```

8 Histograms of Salaries Given Title or School.sas

*salary histograms based on title;

```
title "Adjunct";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Adj";
  histogram salary / midpoints = 0 to 150000 by 5000;
run;
```

```
title "Advisor";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Advis";
  histogram salary / midpoints = 0 to 90000 by 3000;
run;
```

```
title "Associate";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Assoc";
  histogram salary / midpoints = 0 to 240000 by 8000;
run;
```

```
title "Assistant";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Asst";
  histogram salary / midpoints = 0 to 210000 by 7000;
run;
```

```
title "Chair";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Chair";
  histogram salary / midpoints = 0 to 240000 by 8000;
run;
```

```
title "Clinical";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Clin";
  histogram salary / midpoints = 0 to 210000 by 7000;
run;
```

```
title "Dean";
proc univariate data=Final_dataset_dropped_low_sal;
  where Grouped_Title = "Dean";
  histogram salary / midpoints = 0 to 270000 by 9000;
run;
```

```

title "Department";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Dept";
histogram salary / midpoints = 0 to 210000 by 7000;
run;

```

```

title "Director";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Dir";
histogram salary / midpoints = 0 to 210000 by 7000;
run;

```

```

title "Executive";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Exec";
histogram salary / midpoints = 0 to 270000 by 9000;
run;

```

```

title "Instructor";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Instr";
histogram salary / midpoints = 0 to 90000 by 3000;
run;

```

```

title "Lecturer";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Lectu";
histogram salary / midpoints = 0 to 180000 by 6000;
run;

```

```

title "Other";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Other";
histogram salary / midpoints = 0 to 270000 by 9000;
run;

```

```

title "Postdoctorate";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Postd";
histogram salary / midpoints = 0 to 90000 by 3000;
run;

```

```

title "Professor";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Prof";
histogram salary / midpoints = 0 to 270000 by 9000;
run;

```

```

title "Residential";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Res";
histogram salary / midpoints = 0 to 240000 by 8000;
run;

```

```

title "Senior Director";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Sr Di";
histogram salary / midpoints = 0 to 240000 by 8000;
run;

```

```

title "Senior Lecture";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Sr Le";
histogram salary / midpoints = 0 to 150000 by 5000;
run;

```

```

title "Technical/Technician";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Tch";
histogram salary / midpoints = 0 to 90000 by 3000;
run;

```

```

title "Vice President";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "VP";
histogram salary / midpoints = 0 to 270000 by 9000;
run;

```

```

title "Visiting";
proc univariate data=Final_dataset_dropped_low_sal;
where Grouped_Title = "Vst";
histogram salary / midpoints = 0 to 210000 by 7000;
run;

```

*salary histograms based on school;

```
title "Chicago";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Chicago";
histogram salary / midpoints = 0 to 225000 by 5000;
run;
```

```
title "Eastern";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Eastern";
histogram salary / midpoints = 0 to 225000 by 5000;
run;
```

```
title "Governors";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Governors";
histogram salary / midpoints = 0 to 225000 by 5000;
run;
```

```
title "Illinois State";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Illinois State";
histogram salary / midpoints = 0 to 250000 by 5000;
run;
```

```
title "Northeastern";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Northeastern";
histogram salary / midpoints = 0 to 225000 by 5000;
run;
```

```
title "Northern";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Northern";
histogram salary / midpoints = 0 to 250000 by 5000;
run;
```

```
title "Southern";
proc univariate data=Final_dataset_dropped_low_sal;
where school = "Southern";
histogram salary / midpoints = 0 to 250000 by 5000;
run;
```

```
title "U Of I";  
proc univariate data=Final_dataset_dropped_low_sal;  
where school = "U Of I";  
histogram salary / midpoints = 0 to 250000 by 5000;  
run;
```

```
title "Western";  
proc univariate data=Final_dataset_dropped_low_sal;  
where school = "Western";  
histogram salary / midpoints = 0 to 200000 by 5000;  
run;
```

```
title;
```

11 TTests.sas

```
ods pdf file="C:\Users\tmthiel2\Downloads\12 TTest Results.pdf";
title "";
data test_2_titles;
  set final_dataset_dropped_low_sal;
  if grouped_title="Instr" or grouped_title="Dept ";
run;

proc ttest data=test_2_titles ci=equal;
  class grouped_title;
  var salary;
run;

data test_just_gender;
  set final_dataset_dropped_low_sal;
  if gender="F" or gender="M";
run;

proc ttest data=test_just_gender ci=equal;
  class gender;
  var salary;
run;

data test_gender;
  set bayes_model_this;
  if gender="F" or gender="M";
run;

proc ttest data=test_gender ci=equal;
  class gender;
  var resid;
run;

title "";
ODS PDF CLOSE;
```

15 BUGSModelFemale.odc

```
model {  
  for (i in 1:n){  
    MuHatF[i] ~ dnorm(Mu[i], .0000000076677473078539)  
    Mu[i] ~ dnorm(Mu0, tausq0)  
  }  
  Mu0 ~ dunif(0, 100000)  
  tausq0 ~ dgamma(0.001,0.001)  
}  
  
data  
list(  
  n=5892,  
  
  MuHatF = c(...All Relevant Data Points; See specifics in attachment...)
```

16 BUGSModelMale.odc

```
model {  
  for (i in 1:n){  
    MuHat[i] ~ dnorm(Mu[i], 0.0000000048758833887825)  
    Mu[i] ~ dnorm(Mu0, tausq0)  
  }  
  Mu0 ~ dunif(0, 100000)  
  tausq0 ~ dgamma(0.001,0.001)  
}  
  
data  
list(  
  n = 6655,  
  
  MuHat = c(...All Relevant Data Points; See specifics in attachment...)
```