Earnings by California County Employees

An analysis of public available data

Moises Evangelista

2016-06-10

Introduction

This report tries to mimic the report titled County Pay Practices, Report 2015-132. County Pay Practices report was created by the California State Auditor and it sheds light in the wage gap between females and male employees employed by california counties.

This report uses publicly available data to shed more light on the report created by the California State Auditor. The data the data source for this report is from transparentcalifornia.com

Empirical Analysis

The data from transparentcalifornia.com was downloaded for years 2013 and 2014 (the data for year 2015 is incomplete as of 2016-06-10). The data includes employee names but it lacks employee gender. To overcome the gender data not being included a processed was used to match the names to another list that is composed of names and gender. A list of names that includes gender was extracted from the R package `babynames`. 'The baby names list is composed from US baby names provided by the SSA. This package contains all names used for at least 5 children of either sex'. see this more more information.

The names from the county data were generally composed in one of two formats. The first format was `FIRST LAST`, the second format was `LAST, FIRST`.

The county data was processed and the first names extracted. Some names are used by male and females almost equally e.g. Alexy or Adel. The data from the R package was summarized and names were given a proportion of usage e.g. Alexy was classified as female and as was Adel. Some names are almost exclusively used by females e.g. Therese or Annabelle and some are almost exclusively used by males e.g. Rodrigo or Mauricio. Names from the county data were match with the names and the final gender was selected from the names hightest proporition. After the name matching processes some names were left unmatched e.g.Mcheko or SuiKwong.

For this report the median was used because it is less prone to be skewed by total pay observations that are too high or too low especially since the total earnings have long tails e.i. some records have cents and some records have over a million of total earnings.

Data issues

This analysis uses each row as if it was a unique worker and this method this creates some issues. Some workers may be employees in one department with a certain title e.g. Program Analyst and then promote another title within the department to e.g. Information Systems Analyst. In this situations the same worker may appear twice in the data in the same year. In addition, some workers may start working for the county for part of the year due to being new hire or other reasons and some workers may leave county employment due to retirement, death, or employment in the private sector. This situations will create duplicate persons.

Some salaries include back pay and court settlement payments, which make total pay increase and thus skew the data.

Suggestions to mitigate this situations and get a better idea of county pay is to include a unique identifier by worker (it does not have to be the employee number), and include the number of hours a worker was paid in the year and separate the payments due to court settlements. In addition, Kern County data was especially difficult to match.

Table 1: The table below shows the percent of names matched to males or females for years 2013 and 2014. Kern County has the highest percent of unmatched names because their name structure is in disarray

	bothMissing	FemaleLikely	MaleLikely
Alameda County	0.06	0.55	0.39
Amador County	0	0.54	0.45
Butte County	0.02	0.63	0.35

	bothMissing	FemaleLikely	MaleLikely
Calaveras County	0.01	0.56	0.43
Colusa County	0.03	0.58	0.39
Contra Costa County	0.06	0.61	0.34
Del Norte County	0.02	0.59	0.39
El Dorado County	0.01	0.55	0.43
Fresno County	0.04	0.54	0.42
Glenn County	0.02	0.59	0.39
Humboldt County	0.02	0.61	0.37
Imperial County	0.01	0.57	0.42
Inyo County	0.01	0.53	0.45
Kern County	0.5	0.06	0.44
Kings County	0.02	0.58	0.4
Lassen County	0.01	0.47	0.52
Los Angeles County	0.07	0.53	0.4
Madera County	0.02	0.58	0.4
Marin County	0.03	0.52	0.45
Mariposa County	0.01	0.57	0.41
Mendocino County	0.02	0.59	0.39
Merced County	0.03	0.54	0.43
Mono County	0.01	0.42	0.57
Monterey County	0.03	0.63	0.34
Napa County	0.03	0.58	0.4
Nevada County	0.02	0.55	0.43
Orange County	0.04	0.54	0.42
Placer County	0.01	0.57	0.42
Plumas County	0.01	0.54	0.45
Riverside County	0.04	0.59	0.37
Sacramento County	0.04	0.48	0.48
San Benito County	0.01	0.63	0.36
San Bernardino County	0.03	0.62	0.35
San Diego County	0.04	0.55	0.41
San Francisco	0.07	0.41	0.52
San Joaquin County	0.05	0.6	0.34
San Luis Obispo County	0.02	0.58	0.4
San Mateo County	0.05	0.57	0.38
Santa Barbara County	0.02	0.56	0.42
Santa Clara County	0.07	0.58	0.35
Santa Cruz County	0.02	0.58	0.4
Shasta County	0.01	0.61	0.38
Sierra County	0	0.5	0.5
Siskiyou County	0.01	0.51	0.48
Solano County	0.04	0.63	0.34
Sonoma County	0.02	0.55	0.43
Stanislaus County	0.04	0.63	0.33
Sutter County	0.04	0.58	0.39
Tehama County	0.01	0.63	0.36
Tulare County	0.03	0.57	0.4
Tuolumne County	0.01	0.58	0.41
Ventura County	0.02	0.56	0.41
Yolo County	0.03	0.6	0.37
Yuba County	0.03	0.56	0.41

For this analysis the data from Kern County represented a major concern since it does not follow any of the two name formats used by the other counties.

Table 2: Sample of five records that were matched and five records that were unmatched from Kern County data. For some records the names are scrabbled, first name appears first and sometimes it appears in the middle of the name, and commas are missing. In addition, Name initials and name suffix add to the complexity of Kern County data.

Employee.Name	FinalGender	Type
Andress Julie M	MaleLikely	Matched Name
Hernandez Jose	MaleLikely	Matched Name
Garcia Julissa	MaleLikely	Matched Name
Cervantes Alfred	MaleLikely	Matched Name
Bailey Shannon A	FemaleLikely	Matched Name
Cornelison Curtis	bothMissing	Unmatched Name
Mcsperitt Katelyn Cheyene	bothMissing	Unmatched Name
Lund Cheryl L	bothMissing	Unmatched Name
Gibbons Rebecca Lynne	bothMissing	Unmatched Name
Contreras Ernie	bothMissing	Unmatched Name

Median pay

To measure the gender gap, it is useful to know the number of employees who are above and below the median by county. For this the median pay was extracted, rather than the average, by county and by year. The records where then grouped into above or below the median and a Chi-squared test was complete by county. For most counties the number of records classified as female are above the median is lower that what is expected.

Table 3: Expected and observed females counts over the median pay by county in 2014. County where the p-value is highlighted show a statistical significant difference between expected and observed counts

ID	Agency	ActualCount	ExpectedCount	pvalue
1	Alameda County	2,288	2,714	**9.242e-73**
2	Amador County	94	115	**5.428e-05**
3	Butte County	927	1,037	**6.643e-16**
4	Colusa County	109	130	**4.429e-05**
5	Contra Costa County	3,277	3,570	**1.812e-31**
6	El Dorado County	493	615	**5.458e-26**
7	Fresno County	1,780	2,069	**4.32e-42**
8	Glenn County	127	147	**0.0002197**
9	Humboldt County	624	696	**3.358e-10**
10	Imperial County	566	656	**1.702e-14**
11	Inyo County	131	142	**0.04543**
12	Kern County	296	307	0.3448
13	Lassen County	5	5	0.8008
14	Los Angeles County	23,153	27,222	**0**
15	Madera County	268	353	**1.646e-23**
16	Marin County	687	760	**3.239e-08**
17	Mariposa County	146	168	**0.0002983**
18	Mendocino County	380	405	**0.005017**
19	Merced County	728	712	0.196
20	Monterey County	1,983	2,094	**8.335e-09**
21	Napa County	398	478	**5.673e-16**
22	Nevada County	243	279	**5.074e-06**
23	Orange County	4,099	5,044	**3.113e-179**
24	Placer County	634	834	**9.675e-51**
25	Plumas County	138	150	**0.04195**
26	Riverside County	5,737	6,960	**2.9e-249**
27	Sacramento County	2,454	3,033	**4.842e-99**
28	San Benito County	161	172	0.05964
29	San Bernardino County	5,690	6,843	**1.737e-240**
30	San Diego County	4,519	5,337	**3.192e-132**
31	San Francisco	6,632	7,941	**3.414e-172**
32	San Joaquin County	1,962	2,232	**3.178e-41**
33	San Luis Obispo County	762	918	**6.589e-30**
34	San Mateo County	2,535	2,565	0.176
35	Santa Barbara County	1,662	1,888	**5.511e-30**
36	Santa Clara County	4,957	5,399	**1.563e-43**
37	Santa Cruz County	687	769	**5.422e-11**
38	Shasta County	559	615	**6.327e-07**
39	Sierra County	31	34	0.2429
40	Siskiyou County	182	209	**0.0001521**
41	Solano County	905	1,016	**1.192e-16**
42	Sonoma County	1,280	1,437	**1.033e-18**
43	Stanislaus County	1,124	1,308	**7.416e-34**
44	Sutter County	286	324	**2.918e-06**

ID	Agency	ActualCount	ExpectedCount	pvalue
45	Tehama County	251	297	**3.184e-10**
46	Tulare County	846	1,136	**2.706e-79**
47	Tuolumne County	236	255	**0.008567**
48	Ventura County	2,179	2,707	**5.007e-108**
49	Yolo County	449	521	**3.666e-13**
50	Yuba County	262	298	**4.24e-06**

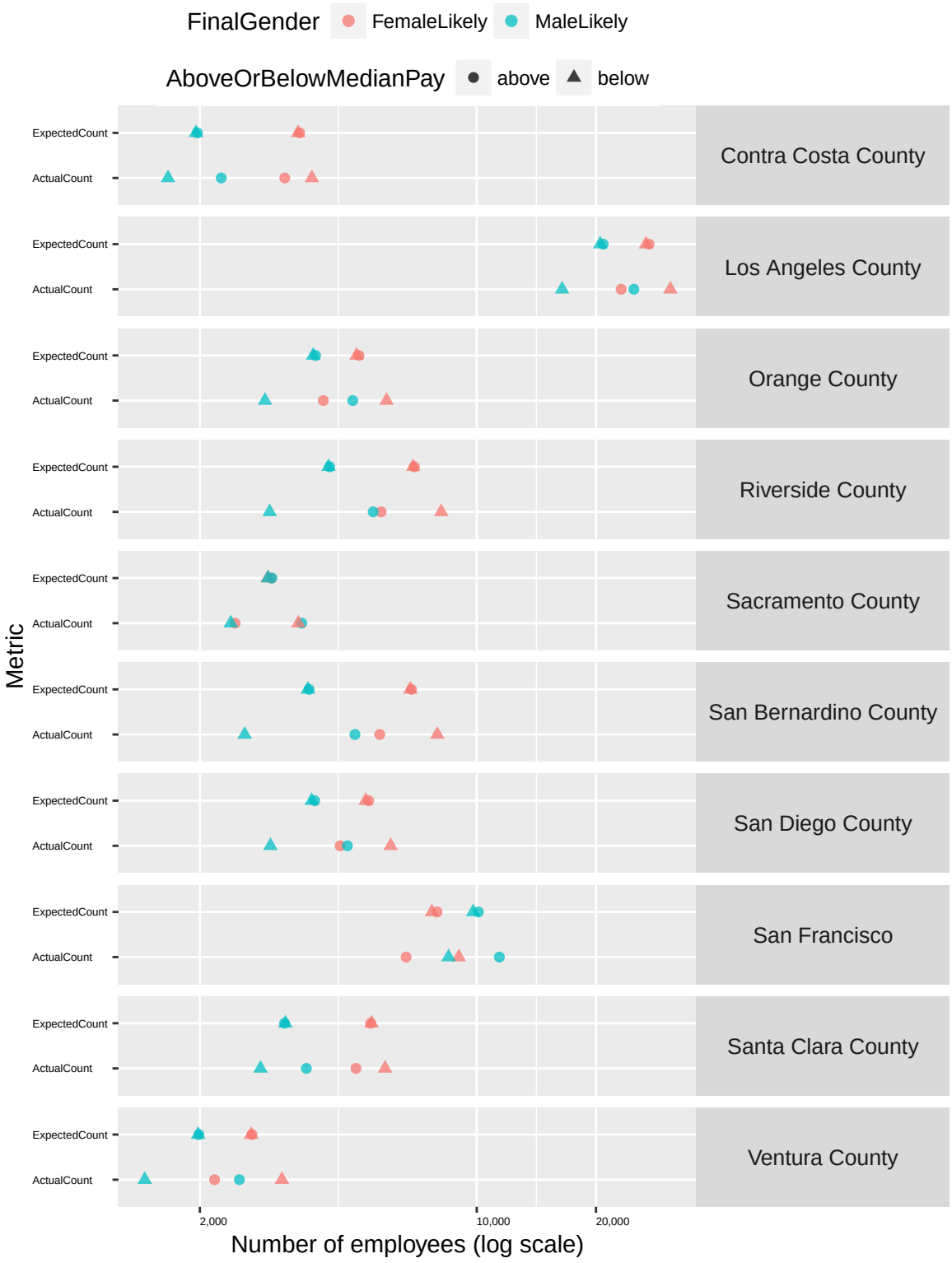

Figure 1: Top 10 counties in 2014 (representing about 70% of total records) broken down by expected and actual counts of those above and below the median total pay.

Table 4: Wage ratio by county for 2014 year. Some couties ratio is close or over to 1 (highlighted in column 'FamaleToMaleRatioWage')

Agency	FemaleLikely	MaleLikely	FamaleToMaleRatioWage
Alameda County	92,220	124,491	0.7408
Amador County	69,915	87,725	0.797
Butte County	39,940	57,987	0.6888
Colusa County	62,185	77,509	0.8023
Contra Costa County	74,968	95,053	0.7887
El Dorado County	62,762	87,468	0.7175
Fresno County	59,480	76,920	0.7733
Glenn County	55,292	65,400	0.8455
Humboldt County	55,101	65,936	0.8357
Imperial County	56,425	69,957	0.8066
Inyo County	68,969	75,283	**0.9161**
Kern County	79,694	82,326	**0.968**
Lassen County	105,630	101,484	**1.041**
Los Angeles County	77,417	102,387	0.7561
Madera County	52,797	67,940	0.7771
Marin County	83,358	95,791	0.8702
Mariposa County	46,127	58,803	0.7844
Mendocino County	56,235	63,415	0.8868
Merced County	67,745	63,568	**1.066**
Monterey County	66,873	77,949	0.8579
Napa County	81,888	104,389	0.7844
Nevada County	61,183	74,433	0.822
Orange County	77,719	105,566	0.7362
Placer County	75,153	105,113	0.715
Plumas County	38,408	52,897	0.7261
Riverside County	54,106	78,030	0.6934
Sacramento County	77,942	101,662	0.7667
San Benito County	54,188	63,133	0.8583
San Bernardino County	58,860	84,623	0.6956
San Diego County	71,084	91,901	0.7735
San Francisco	91,298	111,316	0.8202
San Joaquin County	67,552	89,866	0.7517
San Luis Obispo County	68,823	94,889	0.7253
San Mateo County	72,431	76,098	**0.9518**
Santa Barbara County	50,530	83,377	0.606
Santa Clara County	94,346	111,909	0.8431
Santa Cruz County	83,515	96,638	0.8642
Shasta County	46,255	54,179	0.8537
Sierra County	52,230	65,276	0.8001
Siskiyou County	53,369	67,090	0.7955
Solano County	78,141	95,437	0.8188
Sonoma County	92,516	111,819	0.8274
Stanislaus County	64,073	83,630	0.7661
Sutter County	68,363	85,805	0.7967
Tehama County	53,592	65,567	0.8174
Tulare County	49,777	70,934	0.7017
Tuolumne County	49,760	62,062	0.8018
Ventura County	70,054	103,925	0.6741
Yolo County	70,017	89,513	0.7822
Yuba County	60,706	74,210	0.818

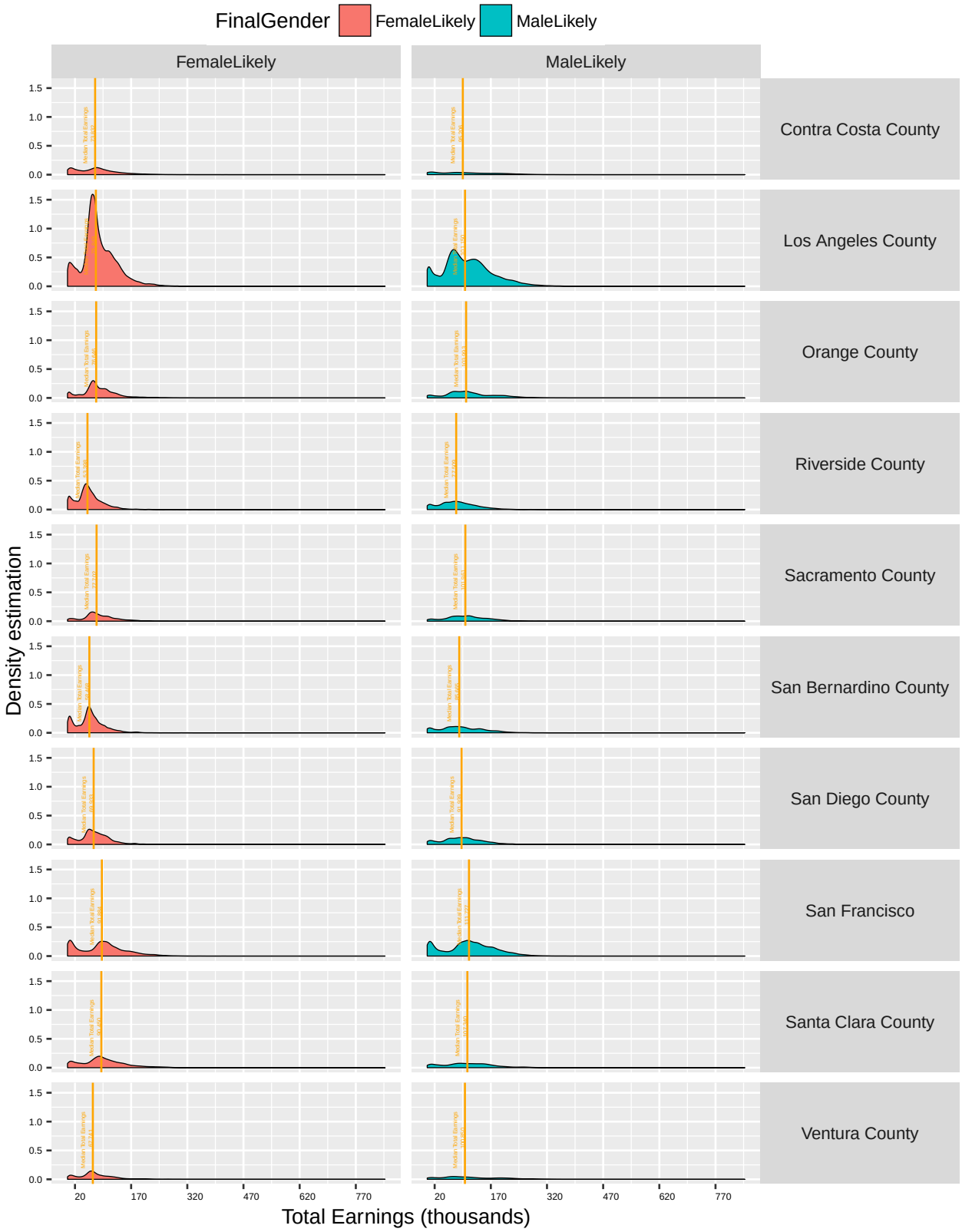

Figure 2: Top 10 counties in 2014 (representing about 70% of total records) broken down gender and county. It excludes records with over one million or less than 100 dollars total pay.

Figure 3: Tectonic changes are needed to close the gap between total earnings between males and females. For Riverside and San Diego Counties, the highest paid records are female in 2014.

Table 5: Descriptive statistics by year

	2013	2014	Test Statistic
	$N = 360372$	$N = 367650$	
FinalGender			$\chi^2_1 = 2,\ P = 0.1$[1]
FemaleLikely	57% (203899)	57% (208682)	
MaleLikely	43% (156473)	43% (158968)	
Total Earnings	49035 **77553** 115933 (85642± 57829)	48629 **79492** 119595 (87872± 60634)	$F_{1,728020} = 173,\ P < 0.001$[2]

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD.Numbers after percents are frequencies.Tests used:[1]Pearson test; [2]Wilcoxon test

Table 6: Descriptive statistics by year for likely males only.

	2013	2014	Test Statistic
	$N = 156473$	$N = 158968$	
Total Earnings	55269 **92842** 137102 (99249± 65152)	55082 **94654** 141429 (101754± 68375)	$F_{1,315439} = 63,\ P < 0.001$

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD.Test used:Wilcoxon test

Table 7: Descriptive statistics by year for likely females only.

	2013	2014	Test Statistic
	$N = 203899$	$N = 208682$	
Total Earnings	45527 **70053** 100256 (75199± 49008)	45065 **72289** 103626 (77296± 51546)	$F_{1,412579} = 156,\ P < 0.001$

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD.Test used:Wilcoxon test

---

The analysis was completed on Friday Jun 10 3:25:01 PM 2016 in 0 seconds.