Using Non-Parametric Methods to Test for Salary Discrepancies among Genders and Races

- Srishti Lakhotia



B. Sc. Statistics (H) | St. Xavier’s College (Autonomous), Kolkata

Session: 2021-2024

Roll Number: 21-300-4-07-0446 | Registration Number: A01-2112-0859-21

Supervised by : Dr. Mausumi Bose

*I affirm that I have identified all my sources, and no part of my dissertation paper uses unacknowledged materials.*

Department of Statistics

6th April, 2024

A screenshot of a document

Description automatically generated

# Contents

|  |  |  |
| --- | --- | --- |
| 1 | Acknowledgement | Page 4 |
| 2 | Introduction | Page 5 |
| 3 | Data | Pages 6-11 |
| 4 | Kolmogorov-Smirnov Test | Pages 12-14 |
| 5 | Mann-Whitney -Test | Pages 15-17 |
| 6 | Analysing the Data (1) | Pages 18-20 |
| 7 | Kruskal-Wallis Test | Pages 21-23 |
| 8 | Analysing the Data (2) | Pages 24-25 |
| 9 | Analysing the Data (3) | Pages 26-27 |
| 10 | Analysing Outliers | Pages 28-30 |
| 11 | Analysing the Majority | Pages 31-32 |
| 12 | Attempting the Model The Data | Pages 33-34 |
| 13 | The Pay Gap | Page 35 |
| 14 | R-Code | Page 36 |
| 15 | Bibliography | Page 36 |

# Acknowledgement

I would like to thank my supervisor, Dr. Mausumi Bose for her invaluable guidance throughout this entire project, from filtering my ideas to completing this task. I am deeply grateful to have received the opportunity to work under her.

I would also like to extend my thanks to all my professors during my B. Sc. Statistics for helping me reach where I am today.

I also wish to acknowledge the help provided by my batchmates in successfully completing this dissertation.

# Introduction

***“We must work together to ensure the equitable distribution of wealth, opportunity, and power in our society.”***

*–Nelson Mandela*

One might have come across the terms “*Wage Gap*”, or, “*Pay Disparity*”, etc. This refers to the socio-economic phenomenon where certain groups of people earn considerably lesser than other groups, for reasons seeming to originate from the very difference that divides the groups. It is generally associated with salary disparities between genders or among races.

It is not uncommon to observe that women and people of colour, on average, get paid significantly less than white men in the United States, even if they have the same qualifications. This paper aims to study this issue on the basis of a 2007 dataset.

A majority of statistical techniques rely on the assumption that the true population distribution for our data is normal, and the errors of our sample data are independently distributed and homoscedastic. But what if our data glaringly violates one or more of these assumptions?

Non-parametric methods of statistical analysis make minimal assumptions about the underlying distributions of the data at hand and thus can be used to robustly test such data.

All work is accomplished using the language R and the IDE Rstudio.

# Data

The data has been procured from Kaggle: <https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset>

This link leads to two .csv files: *Panel Study of Income Dynamics* (PSID) and *Current Population Survey* (CPS). We have chosen to work with the PSID dataset which details microdata of people employed in the US over the 1980-2010 period. It contains 33398 observations across 274 columns. We further subset this file to analyse data from only the year 2007 and drop all unnecessary columns giving up 5959 observations over 47 variables.

These 47 variables include:

* sex
* age
* 4 dummy variables to indicate race (white, black, hispanic, others)
* no. of years of schooling (schupd)
* years of experience (yrsexp)
* hourly wage in 2007 (hrwage)
* 38 dummy variables to indicate occupation of the surveyee

## Exploratory Data Analysis (EDA)

The age distribution of our sample seems to be roughly uniformly distributed across the working ages, with the number decreasing gradually age 50 onwards—as shown in the following histogram:



Figure 1

The following table specifies the sex and race distribution of the sample:

Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender\Race | White | Black | Hispanic | Others | Total |
| Male |  |  |  |  |  |
| Female |  |  |  |  |  |
| Total |  |  |  |  |  |

The sex distribution of the sample is quite even. There are too few observations of people of the Hispanic and “Other” race, together making up only 4.54% of the sample population, whereas people of the White race consist of 64.3%of the sample. In subsequent analyses we may club categories together to make the group sizes more comparable.

The wages of the sample range from $2.03/hr to $572.25/hr with 5655 (out of 5959) observations being below $50/hr.

Since we are interested in comparing salaries, it would help to visualise the hourly wages of our sample via histograms and boxplots. We plot hourly wages for our entire sample, for gender and racial subgroups, and according to occupation as well.

Let us consider the following broad categories of occupation:

* Education (734 observations)
* Finance (986 observations)
* Medicine: doctors, nurses, healthcare technicians (865 observations)
* Trade: consisting of wholesale and retail traders (709 observations)
* Scientist/Computer/Architect (SCA) (330 observations)
* Other (2756 observations)

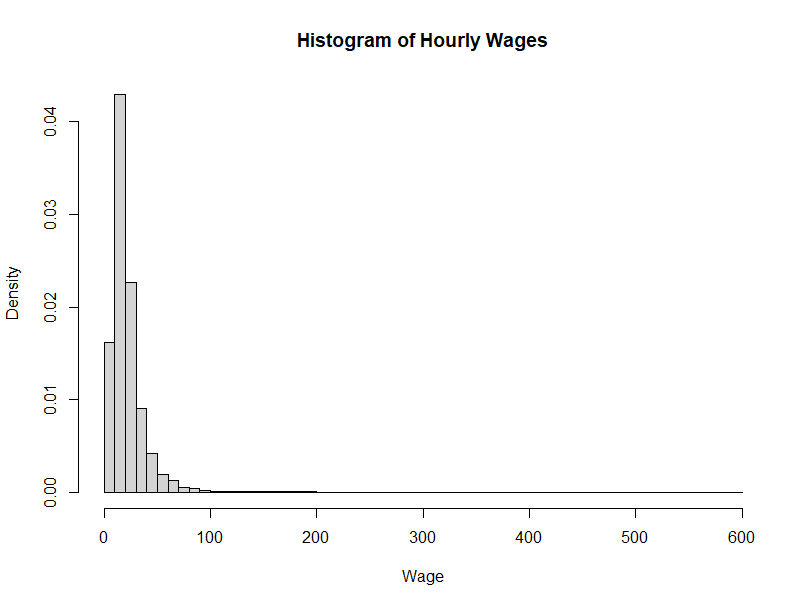


Figure 2

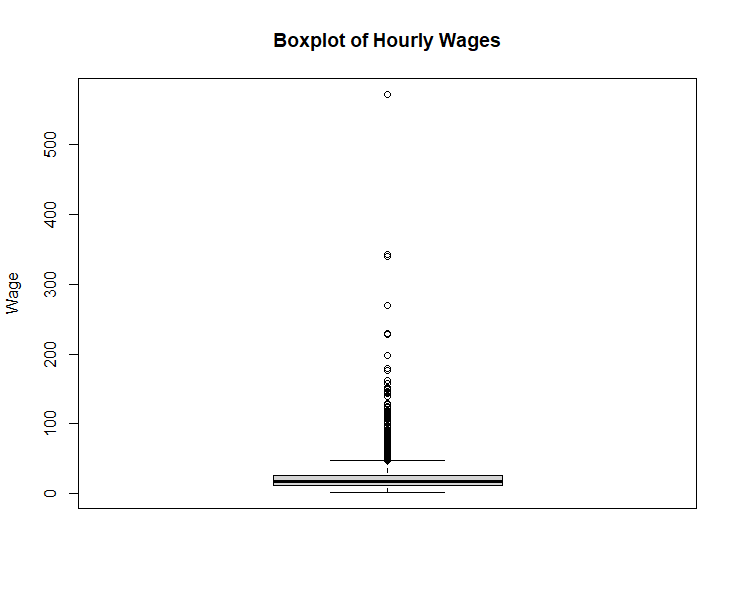


Figure 3

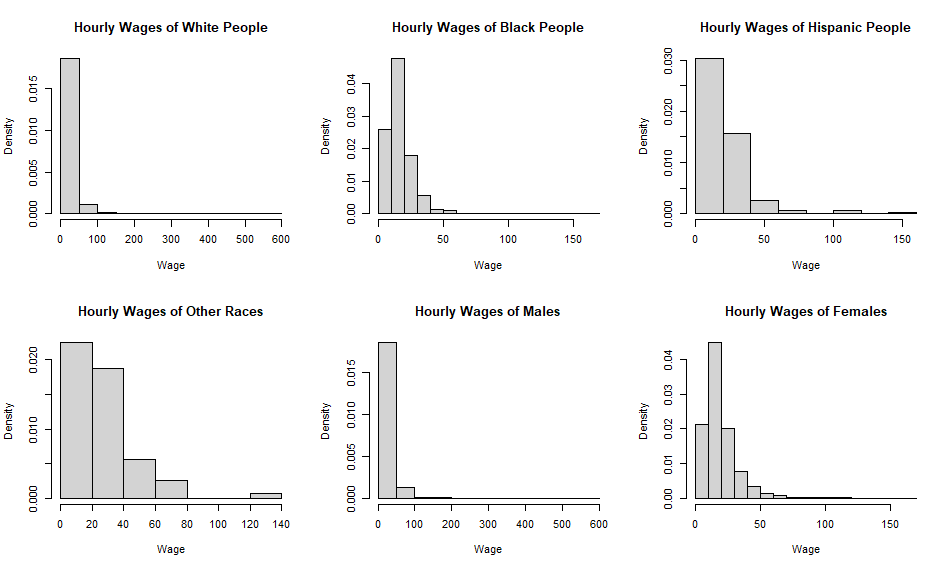
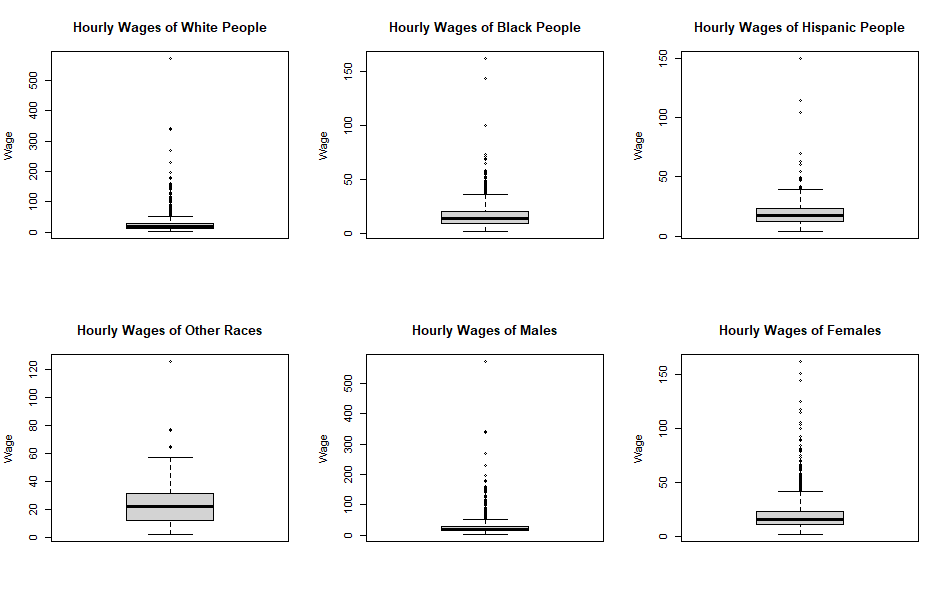


Figure 5

Figure 4

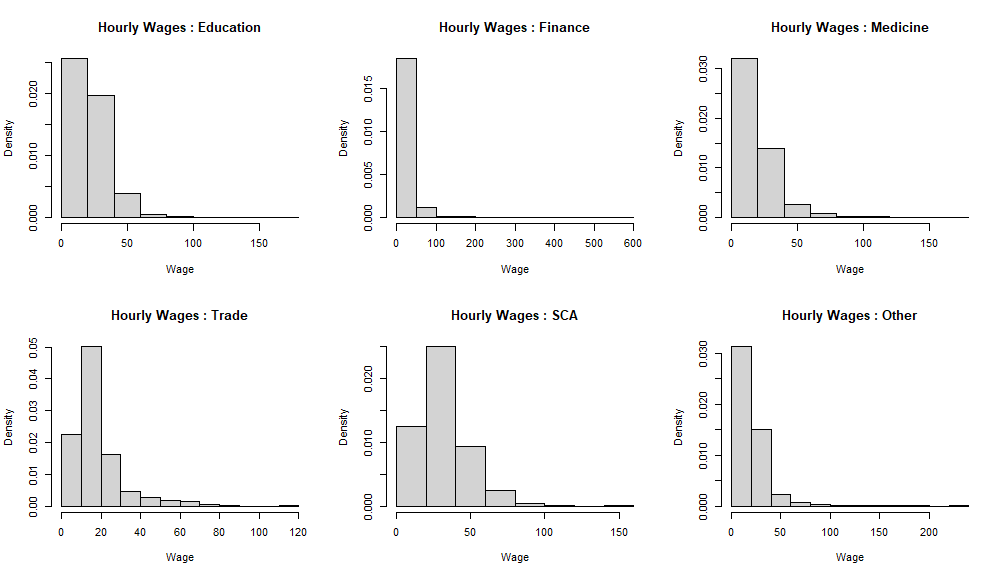
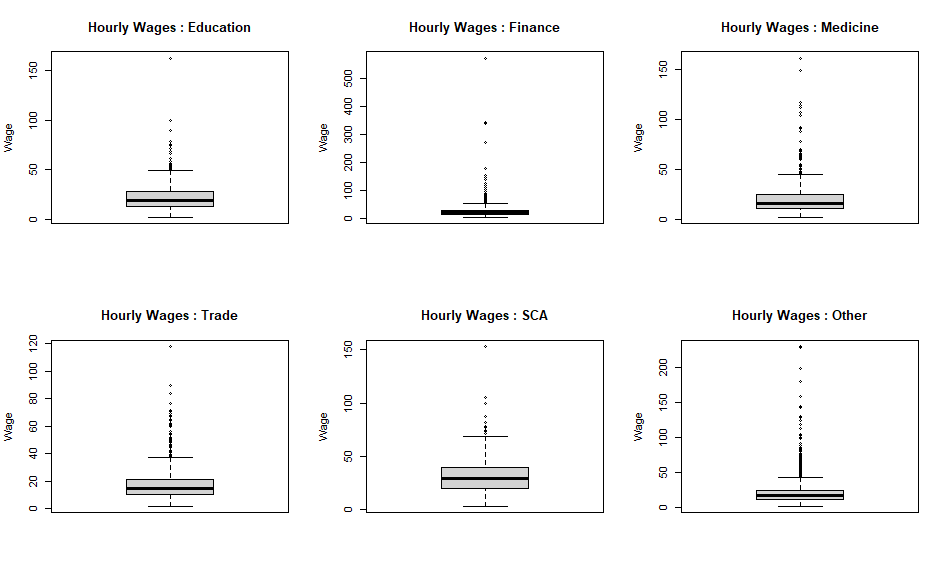


Figure 7

Figure 6

The distribution of wages seems to be heavily skewed, strongly indicating non-normality. There seems to be a fair number of outliers present for each group of people as well.

We can test for normality via a Kolmogorov-Smirnov test for goodness of fit.

# Kolmogorov-Smirnov Test [1]

The Kolmogorov-Smirnov (K-S) one sample test is a non-parametric test used to analyse the goodness of fit of a sample to any specified distribution.

Say, we have a random sample: {from the continuous distribution

We are to test:

**against**

,where is a completely specified continuous distribution function.

Let be the empirical distribution function (edf) of the sample for some , i.e.

Since is the statistical image of the population distribution , the differences between and should be small except for sampling variation, if the null hypothesis is true. Large absolute values of deviations of fromtend to discredit the hypothesis.

Our test statistic is thus:

Critical Region at level of significance (l.o.s.) is :

}

where s.t.

Clearly, does not depend on the distribution of our population i.e. provided it is continuous. Thus, it is a distribution-free statistic.

## Testing for Normality / Lilliefors Test [2]

Now, we consider the problem of goodness of fit test for a normal distribution without any specified mean and variance.

In general, K-S tests can be applied in the case of composite goodness-of-fit hypotheses after estimating the unknown parameters [ will then be replaced by ]. Unfortunately, the null distribution of the K-S test statistic with estimated parameters is far more complicated. This affects the -value calculations.

In the absence of any additional information, one approach could be to use the tables of the K-S test to approximate the -value or to find the approximate critical value. For the normal distribution, Lilliefors (1967) showed that using the usual critical points developed for the K-S test gives extremely conservative results. He then used Monte Carlo simulations to develop a table for the Kolmogorov-Smirnov statistic that gives accurate critical values. #Gibbons

The function LillieTest() in R under the package {DescTools} uses these values when testing for goodness of a normal fit.

The test statistic is:

where is the CDF of a distribution.

are estimated by (sample mean) and (sample variance with divisor respectively. Thus,

## Testing the sample

Upon performing the Kolmogorov-Smirnov test on the hourly wages of the entire sample, and also, every subgroup created earlier (according to gender, race, and occupation), it is found that in every case, the -values are extremely small i.e. our observations deviate very heavily from normality. The tables below summarise the testing:

Table 2

|  |  |  |
| --- | --- | --- |
| Category | -statistic value | -value |
| Entire sample | 0.16655 |  |
| Males | 0.1879037 |  |
| Females | 0.1366724 |  |
| White People | 0.1743357 |  |
| Black People | 0.1276494 |  |
| Hispanic People | 0.2145619 |  |
| Other Races | 0.1655665 |  |

Table 3

|  |  |  |
| --- | --- | --- |
| Occupational Group | -statistic value | p-value |
| Education | 0.1076366 |  |
| Finance | 0.2383868 |  |
| Medicine | 0.1621016 |  |
| Trade | 0.1759288 |  |
| SCA | 0.1041092 |  |
| Other | 0.1640843 |  |

These results provide substantial evidence against normality within the sample. This tells us that using statistical procedures that require normality assumptions likely would not give accurate results in analysis of this dataset. Thus, in this paper, we rely on non-parametric tests to test for salary disparities. Specifically, we use the Mann-Whitney test and the Kruskal-Wallis test.

# Mann-Whitney -Test [3]

The Mann-Whitney (MWU) test or the Wilcoxon Rank Sum Test, is a non-parametric statistical test used to compare two independent samples or groups. It can be considered a non-parametric counterpart to the two-sample -test.

This test is based on the idea that the pattern exhibited when the and random variables are arranged together in increasing order of magnitude provides information about the relationship between their populations. A sample pattern of arrangement where most of the ’s are greater than most of the 's, or vice versa, would be evidence against a random mixing and thus tend to discredit the null hypothesis of identical distributions.

Let and be two independent random variables with CDFs and respectively. We test:

**against**

corresponds to the hypothesis that is stochastically larger than , while corresponds to when is stochastically larger than .

Let and be random samples from of sizes respectively. We define:

The Mann-Whitney statistic is given by:

would tend to take:

larger values if is stochastically larger than .

smaller values if is stochastically larger than.

moderate values if neither are stochastically larger or smaller than the other.

Clearly, does not depend on either or , only on the relative positions of the sample values. Thus, is a distribution-free statistic.

Let us also define

takes the value when .

Critical Region at l.o.s. for :

}

where

Critical Region at l.o.s. for :

where

Critical Region at l.o.s. for :

where

## Large Sample Approximation

It can be shown that, under

Thus, we can use the Central Limit Theorem to approximate a null distribution of our test statistic as follows:

For large under ,

or

The critical regions can then be redefined as:

Critical Region at l.o.s. for :

}

Critical Region at l.o.s. for :

Critical Region at l.o.s. for :

,where is the upperpoint of the standard Normal Distribution.

The function wilcox.test() [with an argument paired = FALSE], performs a Mann-Whitney test in R. We use this test to check for salary discrepancies between independent groups in our sample.

# Analysing the Data (1)

We aim to answer this question on basis of our data:

***Do hourly wages differ significantly across genders and races?***

We can test for stochastic equality of two independent groups via the MWU test. Genders already exist in two categories : Male (M) and Female (F), and thus can be compared using the -test. Races can also be divided into two groups: “White” (W) and “Non-White” (NW) and hence compared. This division also makes the sizes of the respective samples of the two levels of “Race” more comparable.

For given random variables X, Y with CDFsrespectively, we test:

**against**

The findings are summarised in the tables below:

Table 4 : Gender

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Occupational Category |  |  | -val | A/R | -val | A/R | -val | A/R |
| All | M | F |  | R |  | R |  | A |
| Education | M | F |  | R |  | R |  | A |
| Finance | M | F |  | R |  | R |  | A |
| Medicine | M | F |  | R |  | R |  | A |
| Trade | M | F |  | R |  | R |  | A |
| SCA | M | F |  | R |  | R |  | A |
| Other | M | F |  | R |  | R |  | A |

A key to the symbols/abbreviations used in the table:

|  |  |
| --- | --- |
| M | Hourly wages of males |
| F | Hourly wages of females |
| R | We reject for the corresponding alternate hypothesis |
| A | We accept for the corresponding alternate hypothesis |
| -val | -value computed for the test |

Table 5 : Race

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Occupational Category |  |  | -val | A/R | -val | A/R | -val | A/R |
| All | W | NW |  | R |  | R |  | A |
| Education | W | NW |  | R |  | R |  | A |
| Finance | W | NW |  | R |  | R |  | A |
| Medicine | W | NW |  | R |  | R |  | A |
| Trade | W | NW |  | R |  | R |  | A |
| SCA | W | NW |  | R |  | R |  | A |
| Other | W | NW |  | R |  | R |  | A |

A key to the symbols/abbreviations used in the table:

|  |  |
| --- | --- |
| W | Hourly wages of White People |
| NW | Hourly Wages of Non-White People |
| R | We reject for the corresponding alternate hypothesis |
| A | We accept for the corresponding alternate hypothesis |
| -val | -value computed for the test |

## Interpretation

### (1) : Table 4

We see that when our alternative hypothesis is i.e. when we check for , where Y: denotes the hourly wages of males and X is strongly rejected in every case.

There exists a significant salary disparity between males and females in the entire sample and also within the different occupational groups.

Again, when our alternative hypothesis is i.e. when we check for , is strongly rejected in every case.

Hourly wages for men is stochastically larger than that of women i.e. we can say that, on an average, men earn more than women.

We check again, this time keeping our alternative hypothesis as and see that is accepted in every case.

Hourly wages for men is not stochastically smaller than that of women.

Thus, the results of our tests lead us to the conclusion that:

***Hourly wages do vary across gender. Men tend to earn more than women.***

### (2) : Table 5

We see that when our alternative hypothesis is i.e. when we check for , is strongly rejected in every case.

There exists a significant salary disparity between white people and non-white people in the entire sample and also within the different occupational groups.

Again, when our alternative hypothesis is i.e. when we check for , is strongly rejected in every case.

Hourly wages for white people is stochastically larger than that of non-white people i.e. we can say that, on an average, men earn more than women.

We check again, this time keeping our alternative hypothesis as and see that is accepted in every case.

Hourly wages for white people is not stochastically smaller than that of non-white people.

Thus, the results of our tests lead us to a similar conclusion as in the previous case:

***Hourly wages do vary across race. White people tend to earn more than Non-white people.***

We can also analyse whether wages vary significantly for combinations of sexes and races. We employ the Kruskal Wallis Test for this purpose.

# Kruskal-Wallis Test [4]

The Kruskal-Wallis (K-W) test is a non-parametric approach to the one-way ANOVA problem; it deals with the -sample location problem.

Let be independently distributed continuous RVs with CDFs respectively.

We test for:

**against**

We assume that the CDFs differ, if at all, with respect to their location only i.e.

where is some continuous distribution, is the location parameter of the RV.

Our testing problem can be then re-written as:

**against**

Let be random samples of sizes from respectively. Also let .

Thus, we have:

We combine the observations into one ordered sequence from smallest to largest and assign ranks to each observation as follows:

ranges from to.

Under , we essentially have one single sample of size from a single continuous population. The sample ranks may be considered as an SRS (simple random sample) of size drawn from the first natural no.s.

If adjacent ranks are well distributed among the samples, which would be true for a random sample from a single population, the total sum of ranks, , would be divided proportionally according to sample size among the k samples. For the sample which contains observations, the expected sum of ranks would be:

Let denote the actual sum of ranks of the sample i.e.

A reasonable test statistic could be based on a function of the deviations between these observed and expected rank sums. Higher deviations would indicate towards falsehood of the null hypothesis.

We use the following test statistic, due to Kruskal and Wallis (1952):

where mean of the ranks of the sample.

Clearly, does not depend on the distributions of , and only on the relative positions of the sample values. Thus, is a distribution-free statistic.

Critical Region at l.o.s. is :

}

where s.t.

## Large Sample Approximation

It can be shown that, under

Thus, we can use the Central Limit Theorem to approximate a null distribution of as follows:

Under , for large

or

We define as:

can thus be rewritten as:

By properties of the chi-square distribution, we can say that,

as

in distribution

The critical region at l.o.s. can then be redefined as:

}

where is the upperpoint of the distribution.

The function kruskal.test() in R performs a Kruskal-Wallis-sample test.

# Analysing the Data (2)

We aim to answer this question on basis of our data:

***Do hourly wages differ significantly for combinations of genders and races?***

We have two categories for Gender: Male (M) and Female (F) and two for Race: White (W) and Non-White (NW). Thus, there can be 4 combinations:

White Male (WM) White Female (WF)

Non-White Male (NWM) Non-White Female (NWF)

Let be our 4 independently distributed random variables denoting the hourly wages for the sub-groups WM, WF, NWM, NWF respectively.

We test for the following using the Kruskal-Wallis test:

**against**

Again, we assume that the CDFs differ, if at all, with respect to their location only.

We test againstfor the entire sample and then for each occupational group.

The following table summarises our findings:

Table 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Occupational Category | statisticvalue | df | -value | A/R |
| All |  | 3 |  | R |
| Education |  | 3 |  | R |
| Finance |  | 3 |  | R |
| Medicine |  | 3 |  | R |
| Trade |  | 3 |  | R |
| SCA |  | 3 |  | R |
| Other |  | 3 |  | R |

A key to the symbols/abbreviations used in the table:

|  |  |
| --- | --- |
| df | Degrees of freedom |
| R | We reject |
| A | We accept |

## Interpretation

is strongly rejected in every case. This leads us to conclude that:

***There does exist significant differential effects between wages for different combinations of genders and races.***

We can perform tests for each pair of combinations to figure out where the differential effects really exist.

# Analysing the Data (3)

There are 4 combinations of race and gender and thus possible pairings:

{(WM, WF), (WM, NWM), (WM, NWF), (WF, NWM), (WF, NWF), (NWM, NWF)}

For a given pair of combinations ( we can perform a Mann-Whitney tests to check differential effects between and as:

**against**

However, we have already performed these tests for the pairs (White people, Non-White people) and (Males, Females) and have reached unanimous conclusions across all occupational groups. Thus, testing for the pairings {(WM, WF), (WM, NWM), (WF, NWF), (NWM, NWF)} would just give us the same conclusions reiterated, since in each of the pairings, only 1 factor out of sex and race is changed.

Here, our pairings of interest would be: {(WM, NWF), (WF, NWM)}.

Thus, we conduct tests for these two pairings across occupational groups.

The following table summarises our findings:

Table 7

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Occupational Group |  |  | -val | A/R | -val | A/R | -val | A/R |
| All | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | R |  | R |  | A |
| Education | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | A |  | A |  | A |
| Finance | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | A |  | A |  | A |
| Medicine | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | R |  | R |  | A |
| Trade | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | A |  | A |  | A |
| SCA | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | A |  | A |  | A |
| Other | WM | NWF |  | R |  | R |  | A |
| WF | NWM |  | A |  | A |  | A |

## Interpretation

We can interpret values of this table in a similar manner as we did tables 4 and 5.

Unsurprisingly, the -values indicate that there exists a significant difference in the hourly wages of White Males (WM) and Non-White Females (NWF); the former, on average, earns more than the latter.

When we consider the (WF, NWM) pair, we see that in most cases, is accepted.

For the overall sample and the “Medicine” subgroup the -values indicate that there is a significant differential effect between the components of the pair; White Females tend to earn more than Non-White Males.

However, the -values for the other subgroups, namely: “Education”, “ Finance”, “Trade”, and “SCA”, do not indicate any significant difference in the hourly wages between White Females and Non-White Males.

# Analysing Outliers

“*Outliers*” are generally considered to be unusual observations in the sample, characterised by having values of the concerned variable(s) significantly different from most other. As a common rule of thumb, outliers can be considered those observations that lie outside the interval , where are the first and third quartiles of the sample and is the interquartile range (

The quantile behaviour of the hourly wages in our sample can be depicted as:

Table 8

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Entire Sample |  |  |  |  |  |

There is a huge gap between the third quantile and the maximum value in the dataset; while 75% of the wages are less than $26/hour, the remaining observations, constituting 25% of the data span a range of (572.25-26.04 =) $546.21/hr. This indicates the presence of large outlier(s).

.

the interval to consider is

Since hourly wages cannot be negative, there are no observations that are “unusually” low.

However, we find that there are 363 observations in our sample that can be deemed “unusually” high. These 363 observations constitute our “Outliers”.

The presence of outliers in a dataset makes it difficult to work with and one may be tempted to drop any and all outliers present. However, these “unusual” data points may contain valuable information that can help make keen inference about the population under study.

### Race/Gender

The following table summarise the occurrence of outliers within sexes and races.

Table 9

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender\Race | White | Black | Hispanic | Others | Total |
| Male |  |  |  |  |  |
| Female |  |  |  |  |  |
| Total |  |  |  |  |  |

We see that outliers are not distributed proportionally among the racial and sexual groups.

In our sample, the no. of males and females was nearly equal but here, we see that the no. of men who earn unusually high salaries are much higher (about 2.5 times) than that for women.

Also, while white people did amount to a much greater part of the sample than the other races, the contribution to the outlier category is disproportionately large. In the original sample, we have 64.3% white people whereas here, the percentage jumps to 84.85%.

### Occupation

We can also look at which occupations possibly lead to a pay that can be considered usually high. Let us use the actual 38 variables for occupation present in the original dataset for this purpose.

Table 10

|  |  |  |  |
| --- | --- | --- | --- |
| Occupation Variables | Frequency in Sample | Frequency in Outliers | Proportion |
| lawyerphysician | 55 | 28 | 0.509090909 |
| manager | 544 | 109 | 0.200367647 |
| scientist | 82 | 15 | 0.182926829 |
| architect | 112 | 20 | 0.178571429 |
| professional | 459 | 63 | 0.137254902 |
| financialop | 127 | 17 | 0.133858268 |
| computer | 136 | 16 | 0.117647059 |
| Utilities | 88 | 10 | 0.113636364 |
| Communications | 162 | 17 | 0.104938272 |
| artist | 89 | 9 | 0.101123596 |
| finance | 400 | 40 | 0.1 |
| postseceduc | 42 | 4 | 0.095238095 |
| nondurables | 319 | 27 | 0.084639498 |
| sales | 468 | 37 | 0.079059829 |
| durables | 510 | 38 | 0.074509804 |
| wholesaletrade | 202 | 15 | 0.074257426 |
| business | 124 | 9 | 0.072580645 |
| publicadmin | 425 | 29 | 0.068235294 |
| healthcare | 295 | 18 | 0.061016949 |
| legaleduc | 430 | 23 | 0.053488372 |
| Education | 673 | 35 | 0.052005944 |
| Medical | 800 | 34 | 0.0425 |
| miningconstruction | 348 | 13 | 0.037356322 |
| protective | 189 | 7 | 0.037037037 |
| retailtrade | 507 | 17 | 0.033530572 |
| Transport | 299 | 10 | 0.033444816 |
| farmer | 32 | 1 | 0.03125 |
| SocArtOther | 438 | 12 | 0.02739726 |
| transport | 393 | 10 | 0.025445293 |
| production | 431 | 9 | 0.020881671 |
| healthsupport | 203 | 4 | 0.019704433 |
| Agriculture | 52 | 1 | 0.019230769 |
| constructextractinstall | 489 | 9 | 0.018404908 |
| socialworker | 130 | 2 | 0.015384615 |
| officeadmin | 1041 | 12 | 0.011527378 |
| foodcare | 360 | 3 | 0.008333333 |
| hotelsrestaurants | 277 | 2 | 0.007220217 |
| building | 187 | 1 | 0.005347594 |

This table compares the list the frequency of each occupation within the original sample, the outliers and the proportion of observations corresponding to that occupation that belong to the “Outliers”. The table is sorted by proportion in decreasing order.

Clearly, outliers are not distributed uniformly in the occupations. This suggests that the probability of a person earning unusually high salaries is highly dependent on their occupation.

The table is topped by the “lawyerphysician” variable which, according to the data description on Kaggle, refers to physicians and dentists[\*]. Thus, if our sample is assumed to be truly representative of the 2007 US working population, the probability of the hourly wage of a physician or dentist to be unusually high is more than half! The probability is relatively high for managerial positions as well.

[\*] This, however, is likely inaccurate since it makes little sense for a variable named “lawyerphysician” to not deal with lawyers at all, but this presumption cannot be said to be true conclusively.

### Other variables

It is also likely that the amount of formal education and experience of a person also play a role in being able to earn higher wages. We can verify this rudimentarily by comparing median years of schooling and years of experience for observations that are outliers and observations that are not. Let us call the dataset without outliers “Majority”/ “Majority Sample”.

The findings are given as:

Table 11

|  |  |  |
| --- | --- | --- |
|  | Outliers | Majority |
| Median years of schooling | 16 | 14 |
| Median years of experience | 23 | 17 |

We see that the table concedes with the expected outcome. People earning high salaries tend to have more formal education and/or practical experience.

We can see that the “outliers” do not exist purely by chance; they contain valuable information that can be used to make inferences about the population. However, it may still help to observe if and how our conclusions change on removal of outliers.

# Analysing the Majority

We redo tables 4, 5, and 6, testing the same hypotheses but considering the “Majority Sample” as our input.

Table 12 Table 4

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Occupational Category |  |  | -val | A/R | -val | A/R | -val | A/R |
| All | M | F |  | R |  | R |  | A |
| Education | M | F |  | A |  | A |  | A |
| Finance | M | F |  | R |  | R |  | A |
| Medicine | M | F |  | R |  | R |  | A |
| Trade | M | F |  | R |  | R |  | A |
| SCA | M | F |  | R |  | R |  | A |
| Other | M | F |  | R |  | R |  | A |

Table 13⟵ Table 5

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Occupational Category |  |  | -val | A/R | -val | A/R | -val | A/R |
| All | W | NW |  | R |  | R |  | A |
| Education | W | NW |  | R |  | R |  | A |
| Finance | W | NW |  | R |  | R |  | A |
| Medicine | W | NW |  | R |  | R |  | A |
| Trade | W | NW |  | R |  | R |  | A |
| SCA | W | NW |  | R |  | R |  | A |
| Other | W | NW |  | R |  | R |  | A |

Table 14 ⟵ Table 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Occupational Category | statisticvalue | df | -value | A/R |
| All |  | 3 |  | R |
| Education |  | 3 |  | R |
| Finance |  | 3 |  | R |
| Medicine |  | 3 |  | R |
| Trade |  | 3 |  | R |
| SCA |  | 3 |  | R |
| Other |  | 3 |  | R |

## Interpretation

Largely, our conclusions remain the same. When testing between genders, except for in the medical field, the data reiterates:

***Hourly wages do vary across gender. Men tend to earn more than women.***

When testing between races, the data sings the same song:

***Hourly wages do vary across race. White people tend to earn more than Non-white people.***

However, we note that, the -values for this sample, without outliers, are much higher for and when looking at Table 12 and 13, and when considering Table 14. For the Medical Field in Table 12, the -values exceed 0.05.

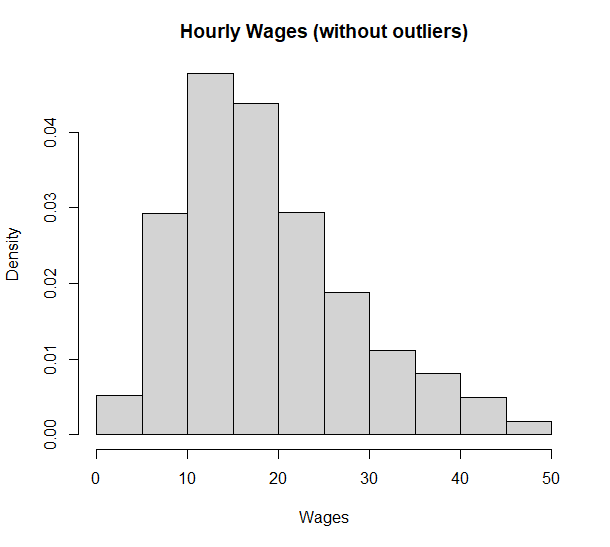
This implies that while the wage gap does persist, the discrepancy between genders and races is smaller for the Majority Sample than the Entire Sample.

# Attempting to Model the Data

A Classical Multiple Linear Regression can be performed to model hourly wages on covariates such as sex, age, highest level of schooling, years of experience etc. However, this is method will yield good results only if our model assumptions hold true.

Among the model assumptions of a Classical Linear Regression Model (CLRM) are that the errors are distributed independently and identically, and they come from a Normal Distribution with parameters: (0, ). Our original dataset did not comply with normality and thus is unsuitable for such a Regression Model. However, we can check if we removal of outliers from the data makes our data tend more strongly towards coming from a normal distribution.

We plot a histogram of hourly wages of the Majority:



The graph is less skewed than before, yet not symmetric.

Again, we perform a Lilliefors’ Test to check goodness of a Normal Fit:

statistic: ; p-value:

We can see that our data still rejects normality heavily. Hence, it is not possible to fit a CLRM to this data.

# The Pay Gap

The analysis of this 2007 dataset provides vital information about the socio-economic state of society. We have established that there exists a prominent pay gap between the sexes and among different races. This raises important questions: W*hy is there a wage gap between the sexes?* *Why is it that White People tend to earn more than People of Colour?*

Some might be tempted to play Devil’s Advocate and try to hypothesize the possibility of there existing an actual divide between men and women, or white people and people of colour, on basis of capability and which hence justifies the existence of the pay gap. However, the existence of any difference in capability itself would result from a society that has failed to nurture a portion of its people. One can dream of equality in economic success only when there first exists an equality in opportunity available to every individual regardless of their gender or skin colour.

Prejudices of employers, imbalanced expectations of society, unequal access to opportunity, all work hand in hand to create a sort of economic hierarchy that our analysis points towards.

“Family caregiving responsibilities bring different pressures for working women and men, and research has shown that being a mother can reduce women’s earnings, while fatherhood can increase men’s earnings.”[5]

“Because women tend to work fewer hours to accommodate caregiving and other unpaid obligations, they are also more likely to work part time, which means lower hourly wages and fewer benefits compared with full-time workers…Gender-based pay discrimination has been illegal since 1963 but is still a frequent, widespread practice—particularly for women of color. It can thrive especially in workplaces that discourage open discussion of wages and where employees fear retaliation. Beyond explicit decisions to pay women less than men, employers may discriminate in pay when they rely on prior salary history in hiring and compensation decisions; this can enable pay decisions that could have been influenced by discrimination to follow women from job to job.”[6]

“Some (racial) pay gaps can be explained by educational differences and geographic location factors. However, even when you control for such factors, wage disparities remain prevalent. Intentional, system-wide laws and practices underlie these differences and have played a role in creating them.”[7]

# R-Code

The codes used in analysis of the dataset can be accessed via: [https://github.com/creationisme/dissrt](https://github.com/creationisme/dissrt%20)

# Bibliography

<https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset>

[1] Gibbons, Chakraborti, Nonparametric Statistical Inference, 2003, Marcel Dekker, Inc. page 120

[2] Gibbons, Chakraborti, Nonparametric Statistical Inference, 2003, Marcel Dekker, Inc. page 133

[3] Gibbons, Chakraborti, Nonparametric Statistical Inference, 2003, Marcel Dekker, Inc. page 268

[4] Gibbons, Chakraborti, Nonparametric Statistical Inference, 2003, Marcel Dekker, Inc. page 363

[5] [https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/#:~:text=Women%20are%20much%20more%20likely,(40%25%20say%20this)](https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/#:~:text=Women%20are%20much%20more%20likely,(40%25%20say%20this).)

[6] <https://www.americanprogress.org/article/quick-facts-gender-wage-gap/>

[7] <https://www.investopedia.com/wage-gaps-by-race-5073258>