Treating the incomes of a city residents

University of Tehran
School of Mathematics, Statistics and Computer Science

January 2021

Hossein Zandinejad

Treating the incomes of a city residents

Hossein Zandinejad
University of Tehran
School of Mathematics, Statistics and Computer Science
January 2021

Introduction

Making economic and political determinations is one of the duties of a government. To help the government achieve that goal, I analyzed our dataset with the help of Central tendencies, Index of Dispersion and Parametric tests.

Data

The dataset that is given to us contains the statistics of people's sex, age, education, and salary.
In the beginning, I classified the dataset by people's sex and, then I compared the Mean and Standard deviation of each group's salary. After that, I calculated the percentage of people with different degrees divided by groups labeled male and female.

sex	age	education	Salary 1	Salary 2
F	18	B	$ 2,500	$ 600
M	16	M	$ 2,600	$ 400
M	21	U	$ 2,000	$ 600
F	35	B	$ 3,000	$ 700
M	54	U	$ 2,400	Missing
F	45	M	$ 1,500	$ 600
M	36	M	$ 4,000	$ 4,660
M	22	M	$ 2,600	$ 2,300
F	28	U	$ 1,200	$ 1,000
F	36	B	$ 2,600	$ 2,500
M	26	M	$ 3,600	$ 1,000
M	25	B	$ 1,000	$ 2,000
F	24	U	$ 900	$ 600
F	60	M	$ 1,500	$ 2,500
F	56	B	$ 800	$ 650
M	58	M	Missing	$ 800
M	52	U	$ 700	$ 600
F	51	B	$ 800	$ 4,500
F	46	B	$ 900	$ 600
M	49	B	$ 2,000	$ 1,200
F	38	U	$ 1,600	$ 300
M	34	U	$ 1,700	$ 350
F	33	U	$ 1,800	$ 400
M	36	M	$ 1,600	$ 1,000
F	59	M	$ 1,200	$ 950
M	52	M	$ 5,000	$ 960
F	51	M	$ 6,000	$ 360
M	53	B	$ 7,000	$ 654
F	25	B	$ 2,600	Missing

Methodology

In our dataset, we mostly have two separate salaries for each person, but some of them, are missed.
First, to find the missed salaries, I calculated the Mean of that column, which contains the missed value we were looking for and, then I replaced the missed data with that Mean.

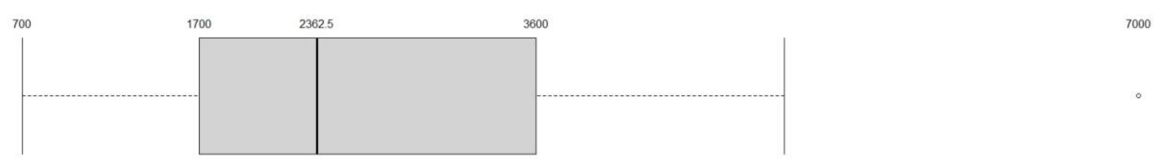
To compare the Means of salary 1 and salary 2, I used the Independent t-test, Paired t-test, and ANOVA (for education levels) with the help of Python and R languages. For using those tests, I needed to check these assumptions:

- Residuals (experimental error) are normally distributed
- Homogeneity of variances (variances are equal between treatment groups)
- Observations are sampled independently from each other

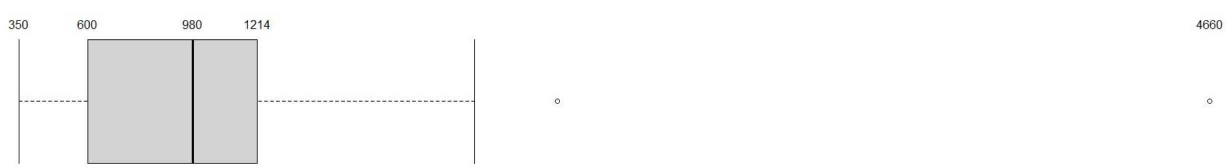
Analysis

➢ I collected these facts from the dataset:

I. The population made up of 52% female and 48% male.
II. Adults (41-60), young adults (25-40), and youths (15-24) made 45%, 38% and, 17% of the sample, respectively.
III. 38% of the people have master's, 34% have bachelor's, and the rest of them are undergraduates.
IV. The number of women with a master's degree (7) is more than the number of men with master's (3).
V. The number of men and women with no special education degrees are equal.
VI. The age average is 39.89 years, and the age standard deviation (SD) is 13.77 years.
VII. The mean and SD of salary 1 are $2325 and $1543.10 per month when these two elements for salary 2 are $1214 and $1118.27 per month.
VIII. As we can see in the last fact, salary 1 is more scattered than salary 2.
IX. Women's salary 1's mean is $1927, and men's salary 1 is $2752 on average.
X. The mean of salary 2 is $1165 for women and $1267 for men.
XI. Salary 1's boxplot for men:



XII. Salary 2's boxplot for men:



XIII. Salary 1's boxplot for women:

XIV. Salary 2's boxplot for women:

- **And these are the tests for salary 1 & salary 2:**

  By the use of Central Limit Theorem (CLT), we can do parametric tests!

  - **The sample size is large (>25).**

- **Independent t-tests: (Classified by Sex)**

  - We want to do the test at $\alpha = 0.05$ significance level.

I. Salary 1:

- Homogeneity of variances:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

We define M for men and W for women

$$M_1, \dots, M_{14}, W_1, \dots, W_{15} \sim N$$

$\bar{M} = 2751.78$

$\bar{W} = 1926.66$

$$S_1^2 = \frac{1}{13}\sum_{i=1}^{14}(M_i - \bar{M})^2 = 1677.29$$

$$S_2^2 = \frac{1}{14}\sum_{i=1}^{15}(W_i - \bar{W})^2 = 1340.82$$

i. $F = \frac{S_1^2}{S_2^2} = 1.251$

ii. $F_{(13,14,0.975)} = 0.324$

$F_{(13,14,0,025)} = 3.011$

iii. $0.324 < 1.251 < 3.011 \quad \rightarrow \quad don't\ reject\ H_0$

$\Longrightarrow \sigma_1^2 = \sigma_2^2$

Now we test: $\begin{cases} H_{0:}\, \mu_1 = \mu_2 \\ H_{a:}\, \mu_1 \neq \mu_2 \end{cases}$

i. $S_p^2 = \frac{13(1677.29)+14(1340.82)}{27} = 1502.82$

ii. $t = \frac{2751.78-1926.66}{38.76\sqrt{\frac{1}{14}+\frac{1}{15}}} = \frac{825.12}{38.76(0.371)} = \frac{825.12}{14.379} = 57.38$

iii. $t_{(27;0,975)} = 2.052$

iv. $57.38 > 2.052 \quad \rightarrow \mathit{reject}\ H_0$

$\Longrightarrow\ \mu_1 \neq \mu_2$

II. Salary 2:

- Homogeneity of variances:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

We define M for men and W for women

$$M_1, \dots, M_{14}, W_1, \dots, W_{15} \sim N$$

$\bar{M} = 1267$

$\bar{W} = 1165$

$$S_1^2 = \frac{1}{13}\sum_{i=1}^{14}(M_i - \bar{M})^2 = 1124.454$$

$$S_2^2 = \frac{1}{14}\sum_{i=1}^{15}(W_i - \bar{W})^2 = 1149.607$$

i. $F = \frac{S_1^2}{S_2^2} = 0.978$

ii. $F_{(13,14,0.975)} = 0.324$

$F_{(13,14,0,025)} = 3.011$

iii. $0.324 < 0.978 < 3.011 \quad \rightarrow \quad \mathit{don't\ reject}\ H_0$

$\Longrightarrow\ \sigma_1^2 = \sigma_2^2$

Now we test: $\begin{cases} H_{0:}\, \mu_1 = \mu_2 \\ H_{a:}\, \mu_1 \neq \mu_2 \end{cases}$

i. $S_p^2 = \frac{13(1267)+14(1165)}{27} = 1214.11$

ii. $t = \frac{1267-1165}{34.84\sqrt{\frac{1}{14}+\frac{1}{15}}} = \frac{102}{34.84(0.371)} = \frac{102}{12.92} = 7.89$

iii. $t_{(27;0,975)} = 2.052$

iv. $7.89 > 2.052 \quad \rightarrow \mathit{reject}\ H_0$

$\Longrightarrow\ \mu_1 \neq \mu_2$

- **Paired t-test: (Salary 1 & Salary 2)**

  - We run the test at $\alpha = 0.05$ significance level.

$$d_i = X_i - Y_i \;; i = 1, \dots, 29 \qquad , d_1, \dots, d_{29} \sim N(\mu_d, \sigma_d^2)$$

$$Note: X \coloneqq Salary1 \;\&\; Y \coloneqq Salary2$$

$$\begin{cases} H_0: \mu_d = 0 \\ H_1: \mu_d \neq 0 \end{cases}$$

i. $\bar{d} = \frac{1}{29}\sum_{i=1}^{29} d_i = 1110.793$

ii. $S_d^2 = \frac{1}{28}\sum_{i=1}^{29}(d_i - \bar{d})^2 = 1932.405$

iii. $T = \frac{1110.793 - 0}{\frac{43.959}{\sqrt{29}}} = \frac{1110.793}{8.163} = 136.076$

iv. $t_{(28,0.975)} = 2.048$

v. $136.076 > 2.048 \quad \rightarrow reject\ H_0$
$\Longrightarrow \mu_d \neq 0$

- **ANOVA (ANalysis Of VAriance): (Classified by Education)**

- **Assumptions:**

  - **The responses for each factor level have a normal population distribution**
  - **Homogeneity of variances (check below)**
  - **The data are independent**

I. Salary 1:

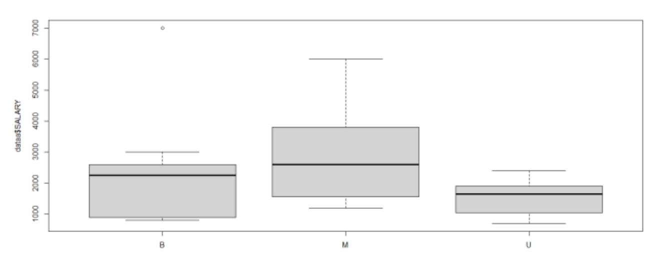
- We want to do the test at $\alpha = 0.01$ significance level.

  - Homogeneity of variances:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \\ H_a: \quad o.w \end{cases}$$

Note: We use "1" for the undergraduates, "2" for bachelor degree and "3" for master degree.

U	B	M
2000	2500	2600
2400	3000	1500
1200	2600	4000
900	1000	2600
700	800	3600
1600	800	1500
1700	900	2325
1800	2000	1600
	7000	1200
	2600	5000
		6000



i. $SSE = \sum_{j=1}^{3}(n_j - 1)S_j^2 = 58045318$

ii. $MSE = \frac{SSE}{\sum_{j=1}^{3}(n_j)-3} = \frac{55774555.99}{26} = 2232512$

iii. $M = \sum_{j=1}^{3}(n_j - 1)\ln\ MSE - \sum_{j=1}^{3}(n_j - 1)\ln S_j^2 = 380.08458 - 371.543152 = 8.5414$

iv. $C^{-1} = 1 - \frac{1}{3(k-1)}(\sum_{j=1}^{3}\frac{1}{n_j-1} - \frac{1}{\sum(n_j-1)}) = 1 - \frac{1}{6}(0.31549) = 0.94741$

v. $MC^{-1} = 8.09229102$

vi. $\chi^2_{(2,0.99)} = 9.210$

vii. $MC^{-1} < 9.210 \ \rightarrow don't\ reject\ H_0\ \ (at\ \alpha = 0.01)$
$\Rightarrow \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

viii. And also, we can use R:

```
        Bartlett test of homogeneity of variances

data:  SALARY by EDU
Bartlett's K-squared = 8.1147, df = 2, p-value = 0.01729
```

Now we test: $\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_{1:} \qquad o.w \end{cases}$

i. $SST = \sum_{j=1}^{3} n_j\,(\overline{x}_{.j} - \overline{x}_{..})^2 = 8627181.82$

ii. $MST = \frac{SST}{3-1} = 4313590.91$

iii. $F = \frac{MST}{MSE} = \frac{4313590.91}{2232512.238} = 1.932169$

iv. $F_{(2,26,0.99)} = 5.53$

v. $F < 5.53\ \rightarrow\ \ don't\ reject\ H_0$

```
            Df   Sum Sq Mean Sq F value Pr(>F)
EDU          2  8627182 4313591   1.932  0.165
Residuals   26 58045318 2232512
```

$\Longrightarrow \mu_1 = \mu_2 = \mu_3$

II. Salary 2:

- We want to do the test at $\alpha = 0.05$ significance level.

Homogeneity of variances:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \\ H_a: \qquad o.w \end{cases}$$

Note: We use "1" for the undergraduates, "2" for bachelor degree and "3" for master degree.

U	B	M
600	600	400
1214	700	600
1000	2500	4660
600	2000	2300
600	650	1000
300	4500	2500
350	600	800
400	1200	1000
	654	950
	1214	960
		360

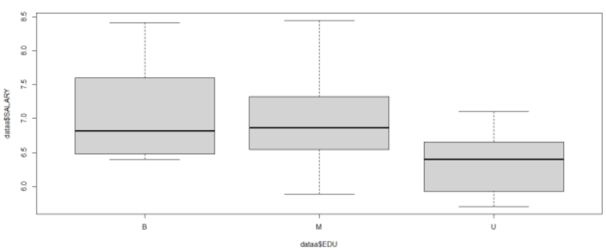
i. $SSE = \sum_{j=1}^{3}(n_j - 1)S_j^2 = 31269967$
ii. $MSE = \frac{SSE}{\sum_{j=1}^{3}(n_j)-3} = \frac{31269967}{26} = 1202691.05$
iii. $M = \sum_{j=1}^{3}(n_j - 1)\ln\ MSE - \sum_{j=1}^{3}(n_j - 1)\ln S_j^2 = 364.001876 - 352.307616 = 11.69426$
iv. $C^{-1} = 1 - \frac{1}{3(3-1)}(\sum_{j=1}^{3}\frac{1}{n_j-1} - \frac{1}{\sum(n_j-1)}) = 1 - \frac{1}{6}(0.31549) = 0.9474155$
v. $MC^{-1} = 11.07932$
vi. $\chi^2_{(2,0.95)} = 5.991$
vii. $MC^{-1} > 5.991\ \rightarrow\ reject\ H_0$
viii. And by the use of R:

```
        Bartlett test of homogeneity of variances

data:  SALARY by EDU
Bartlett's K-squared = 11.11, df = 2, p-value = 0.003868
```

- In this case, I transformed the dataset, because I found a significant effect on "log" of data.

U	B	M
6.39693	6.39693	5.991465
7.101676	6.55108	6.39693
6.907755	7.824046	8.446771
6.39693	7.600902	7.740664
6.39693	6.476972	6.907755
5.703782	8.411833	7.824046
5.857933	6.39693	6.684612
5.991465	7.090077	6.907755
	6.483107	6.856462
	7.101676	6.866933
		5.886104



Now we check homogeneity of variances again:

```
        Bartlett test of homogeneity of variances

data:  SALARY by EDU
Bartlett's K-squared = 1.5927, df = 2, p-value = 0.451
```

i. Since $p - value > \alpha$, we don't reject H_0 hypothesis.
$\Rightarrow \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

Now we test: $\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \quad o.w \end{cases}$

i. $SST = \sum_{j=1}^{3} n_j\,(\bar{x}_{.j} - \bar{x}_{..})^2 = 2.467$
ii. $MST = \frac{SST}{3-1} = 1.2335$
iii. $SSE = \sum_{j=1}^{3}(n_j - 1)S_j^2 = 12.256$
iv. $MSE = \frac{SSE}{\sum_{j=1}^{3}(n_j)-3} = \frac{12.256}{26} = 0.4714$
v. $F = \frac{MST}{MSE} = \frac{1.2335}{0.4714} = 2.617$
vi. $F_{(2,26,0.95)} = 3.37$
vii. $F < 3.37\ \rightarrow\ don't\ reject\ H_0$

```
            Df Sum Sq Mean Sq F value Pr(>F)  
EDU          2  2.467  1.2335   2.617 0.0922 .
Residuals   26 12.256  0.4714                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Longrightarrow \mu_1 = \mu_2 = \mu_3$

Summary:

From the analysis of the data with respect to classification by sex and education, I extracted general information and performed t-tests and ANOVA test to achieve some results about the population's mean and variance.

References:

1. Robert Kabacoff - R in Action. Data Analysis and Graphics with R (2015, Manning)

2. Dr. A. Parsian – Basic Concepts of Probability and Statistics for Science and Engineering Students (Third Edition)

3. Nader Nematollahi – Statistical Methods

4. Homogeneity of variances in R
https://www.datanovia.com/en/lessons/homogeneity-of-variance-test-in-r

5. Basic Inferential Data Analysis Instructions
https://rpubs.com/hmisaii/Statistical_Inference

6. One-Way ANOVA Test In R
http://www.sthda.com/english/wiki/one-way-anova-test-in-r