Salary Variance Analysis

Table of Contents

Contents

List of Figures

Salary Variance Analysis

Summary:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination

Note: Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Sample of the Dataset:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Figure 1. Salary_Dataset_Sample

Exploratory Data Analysis:

Let us check the basic info of the data frame

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Figure 2. Salary_Dataset_Info

- There are 40 rows and 3 columns
- 2 columns are of the 'object' (text) type and 1 of 'integer' type (non decimal values)

- There are no missing values in the dataset as denoted by 40 non-null in every column

Data breakdown

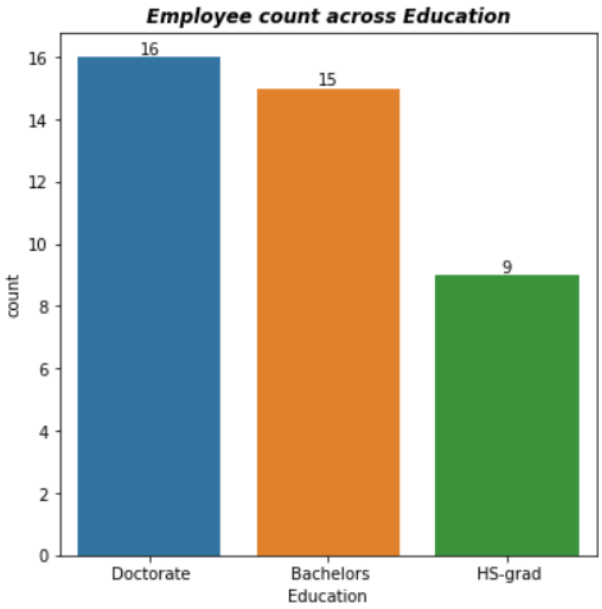

Figure 3. Employee Count_Education

F

Figure 4. Employee count across Occupation

- The above 2 charts give a breakdown of employees in the dataset
- The first chart shows a split of employees by their education levels

- The second chart breaks down the first chart even further by providing details on the various occupations for each educational level

Assumptions of Anova:--

- Independent Sample - Sample should be selected randomly (Equally likely events); there should not be any pattern in the selection of sample
- Normal Distribution - Distribution of each group should be normal
- Homogenous Group - Variance between the group should be the same
- The groups must have the same sample size

Questions:

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

1.1.1 Education

H_O: The mean Salary of an individual is same across all levels of Education
H_A: The mean Salary of an individual is different for at least one level of Education

1.1.2 Occupation

H_O: The mean Salary of an individual is the same across all levels of Occupation
H_A: The mean Salary of an individual is different for at least one level of Occupation

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Figure 5. Education_ANOVA

- The above figure shows the results of the one-way Anova performed for Education with respect to Salary
- We assume a significance level of 0.05
- As the p-value (1.257709e-08) is lesser than the significance level, we reject the Null

- We conclude that there is significant difference in the mean salaries for at least one level of education

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Figure 6. Occupation_ANOVA

- The above figure shows the results of the one-way Anova performed for Occupation with respect to Salary
- We assume a significance level of 0.05
- As the p-value (0.46) is greater than the significance level, we FAIL to reject the Null
- We conclude that the mean Salary is the same for all levels of Occupation

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

- The Null hypothesis was rejected for Education (1.2) and could not be rejected for Occupation (1.3)
- The class means are significantly different for Education
- We conclude the mean salary is different for different levels of Education and the same for different levels of Occupation

1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

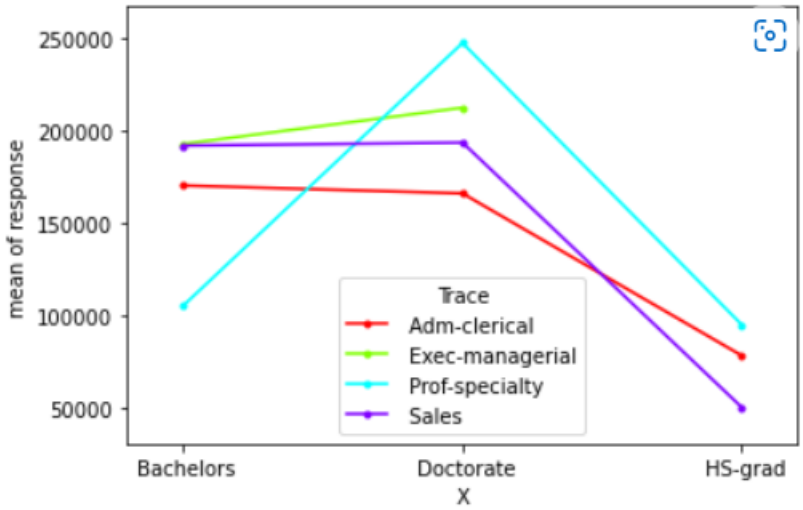

Figure 7. Interaction Plot_1

- We can see the salaries for the 3 educational levels (Doctorate, Bachelors and HS-grad) distributed across the 4 different occupations (Adm-clerical, Sales, Prof-specialty and Exec-managerial)
- Doctorate has the highest paying jobs across 3 occupations (Adm-clerical being the exception), followed by Bachelors and HS-grad being the least
- For Doctorate, the highest salary of 250,000 is for the Prof- specialty category and the least salary of 160,000 is for the Adm-clerical category
- Exec-managerial at 210,000 exceed Sales at 190,000 for Doctorates
- For Bachelors, the highest salary is 190,000 for Sales and Exec-managerial roles, this is followed by Adm-clerical at 170,000 and Prof- specialty at 100,000
- For HS-grad, the highest salary of 90,000 is for the Prof- specialty category, followed by Adm-clerical at 80,000 and Sales being the least at 50,000
- There are no roles in Exec-managerial for HS-grads

1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable 'Salary'. State the null and alternative hypotheses and state your results. How will you interpret this result?

H_O: The mean Salary of an individual is the same across all levels of Education
H_O: The mean Salary of an individual is the same across all levels of Occupation
H_O: There is no interaction between the 2 factors – Education and Occupation
H_A: The mean Salary of an individual is different for at least one level of Education or Occupation

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Figure 8. Education_Occupation_two-way ANOVA_1

- The above figure shows the results of the two-way Anova performed for Education & Occupation with respect to Salary
- We assume a significance level of 0.05
- The p-value for Education is very minimal and lesser than the significance level; hence we reject the Null
- As the p-value for Occupation is high and greater than the significance level, we FAIL to reject the Null
- We conclude that the mean Salary is the same for all levels of Occupation and different for at least one level of Education
- Also, Occupation alone doesn't contribute to explain the variance of Salaries in the dataset

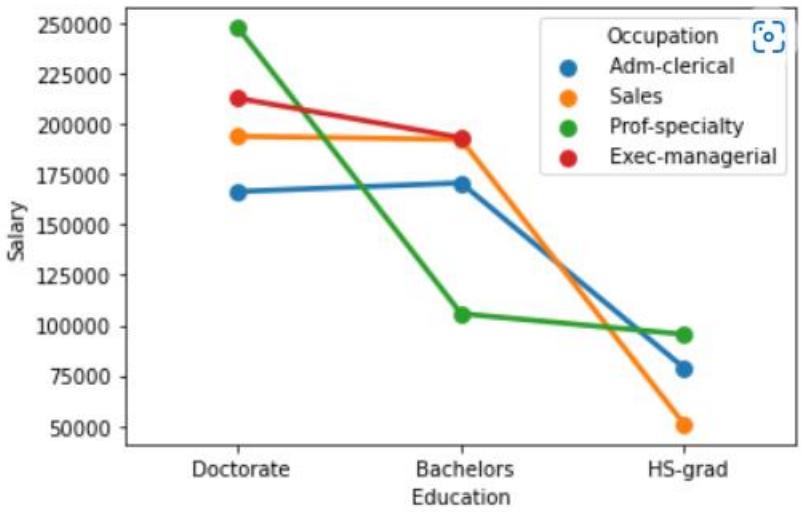

Figure 9. Interaction Plot_2

- We plot the above graph to check whether there is any interaction between the factors
- As there is significant overlap between the two, we'll introduce a new interaction term to perform the two-way Anova

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Figure 10. Education_Occupation_two-way ANOVA_2

- Due to the inclusion of the interaction effect term, we can see the p-value for Education is very minimal and lesser than the significance level, and Occupation p-value is higher than the significance level as in the Two-Way ANOVA without the interaction effect terms.
- The p-value of the interaction effect term of 'Education' and 'Occupation' suggests that the Null hypothesis is rejected in this case.
- Education and Occupation together explain the variance in the salaries

1.7 Explain the business implications of performing ANOVA for this particular case study

- ANOVA checks the impact of one or more factors by comparing the means of different samples (Education and Occupation being the factors in this scenario)
- The business will be able to understand the impact of education and designation on the salary levels amongst employees
- It can also focus on developing the career path of employees based on their education and long-term goals
- The business can also look at helping employees who'd like to switch roles based on their current occupation and long-term goals
- The Null hypothesis was rejected for Education and hence we concluded that there is significant difference in the mean salaries for at least one level of education
- The Null could not be rejected for Occupation; hence we concluded that there is no significant difference in the mean salaries across the 4 categories of occupation
- A larger data size can help us in getting more accurate results and help the business to make better plans and strategies