Correlation Analysis

1. Data Summary

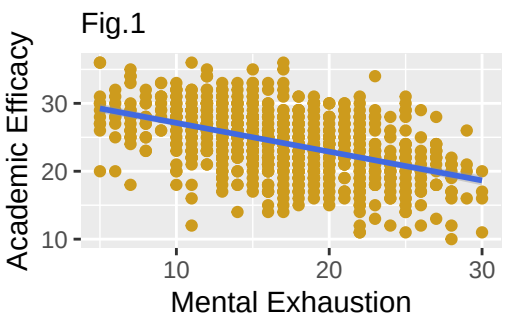
The dataset has information regarding mental health, burnout, empathy and academic efficacy of students studying medical in Switzerland. The dataset also includes factors such as job and health satisfaction, mental exhaustion. It has 20 unique variables and 886 observations. There are continuous and categorical variables in this dataset. This report will mainly focus on 3 variables namely: ***mbi_ex*** (int) : MBI Emotional Exhaustion(ranging from 5 to 30), ***mbi_ea*** (int): MBI Academic Efficacy (ranging from 10 to 36), ***health*** (int) : Satisfaction with health(ranging from 1 to 5, 1 being least satisfied).

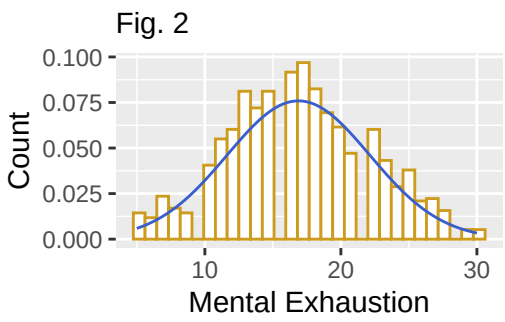
1.1 Objective

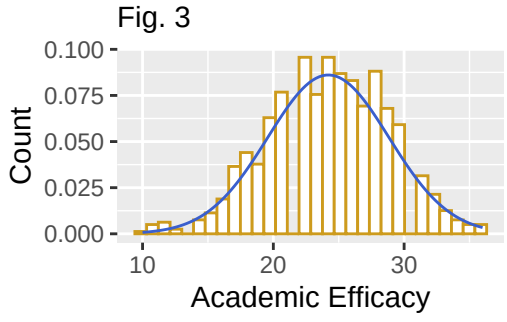
There is ought to be a link between the mental exhaustion and cognitive abilities in humans. The objective of this report is to find the correlation between MBI Emotional Exhaustion and MBI Academic Efficacy and how the academic efficacy is affected with emotional exhaustion in medical students. This report analyses the notion of mental exhaustion affecting the cognitive and academic efficacy in humans.

1.2 Data Cleaning and Preprocessing

This section deals with cleaning and transforming data. The initial step is to drop the redundant data and only keep the variables useful for this analysis. Hence, mbi_ex, mbi_ea and health variables are selected from the tibble, rest are forwent. The scaled down data is then tested for any missing values, it turns out that there are no missing values in dataset. Furthermore, the dataset is tested for any outliers by plotting scatterplots and describing the dataset, resulting with no outliers or anomalies.







2. Planning

The first objective is to check for assumptions. The assumptions being tested in this report are normality, homoscedasticity and interval data. After checking the assumptions, the variables are tested for correlation using correlation coefficients . The correlation coefficient selected depends upon assumptions as well as degree of correctness of assumptions.

2.1 Test for Normality

To test for normality, this analysis checks for 1) **skewness** and **kurtosis**, 2) plotting the **density histograms** with dnorm distribution and 3) **Q-Q plots**. After observing these visualizations, the data appears to be near normal. In order to test further, this report incorporates normality testing method **Shapiro Wilk**[1], which resulted in p value < 0.05, that indicates significance and therefore resulting in data to be non-normal. However, as per Q-Q plots, the data looked near normal. The log and square root data transformation techniques are applied to normalize the data, but these transformations skewed the data even further, thus the data transformation techniques are not used for the scope of this analysis. To remove the ambiguity, the results from QQ plots and descriptive analysis are considered by invoking the **Central Limit Theorem**[2] because datapoints are > 30. Consequently, the sample is considered to be normal.

2.2 Test for Homoscedasticity

Homoscedasticity[3] is tested by Levene Test[4]. The test results in p value > 0.05, confirms that the dataset has same finite variance and is similar. Hence, this assumption is true and sample is considered to have homoscedasticity.

[1] Test used to check whether the sample is normal

[2] Central Limit Theorem states that if the random sample size is sufficiently large than the sample distribution means follow an approximate normal distribution

[3] Homogeneity of variance

[4] Test used to check if samples have same variance

2.3 Interval Data

The variables to be tested out are **mbi_ex**(int ranging from 5 to 30) and **mbi_ea**(int ranging from 10 to 36). Both these variables fall under the category of interval data. However, the third variable **health**(int ranges from 1 to 5, 1 being least satisfied) seems to be ordinal data as it is rated from high to low, but since the values are considered to be equidistant and average ratings can be calculated, indicating some arithmetic operations can be performed, this report considers health variable to be an interval.

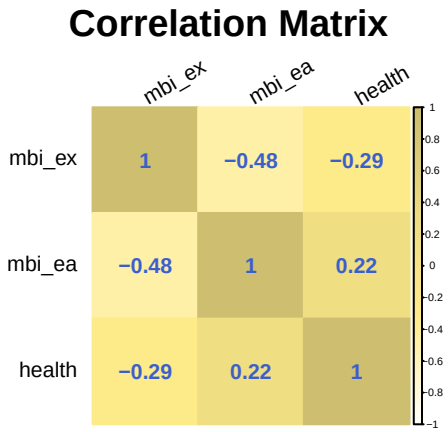
2.4 Choosing Correlation Coefficient

For **Pearson Correlation Coefficient**[5] the assumptions are : 1) Interval Data 2) If Confidence Intervals[6] are to be calculated, then the distribution should be normal. As tested above, both assumptions are true for the given dataset. Thus, Pearson Correlation Coefficient is used to calculate the correlation in this analysis.

3. Analysis

This section tests for the correlation between **Mental Exhaustion** and **Academic Efficacy**. While conducting the test to calculate Pearson Coefficient, results are $r = -0.4808207$ with 95% CIs (-0.5299022, -0.4285288). Negative correlation states that the variables are inversely proportional, in other words, when Mental Exhaustion increases the Academic Efficacy decreases. The effect size of this correlation ranges from medium to large. The CIs range from -0.529 to -0.428, this range does not include 0, stating that the correlation is negative throughout and does not change. To calculate R^2 [7], Pearson Coefficient(r) is squared, thus $R^2 = 0.23118$. To calculate the correlation percentage Multiply R^2 by 100. Hence, there is a 23.11% correlation between Mental Exhaustion and Academic Efficacy.

In addition, to calculate the pure relationship between Mental Exhaustion and Academic Efficacy, the influence of health satisfaction is taken into account which results in **Partial Correlation**. The partial correlation coefficient changed from r = -0.4808207 to r = -0.4462688 with p value < 0.001 when effect of health satisfaction is held constant, decreasing the correlation percentage from 23.11% to 19.91%.



Conclusion

After primitively scaling down the data, visualizing using graphs, checking for assumptions, and choosing the best possible method to find the correlation between Mental Exhaustion and Academic Efficacy, the Pearson Correlation Coefficient is found to be the most feasible option to calculate correlation. It is examined that Academic Efficacy is negatively related to Mental Exhaustion but positively related to health satisfaction, and health satisfaction itself is negatively related to Mental Exhaustion(see Correlation Matrix). The correlation percentage between Mental Exhaustion and Academic Efficacy is 23.11 %, between Mental Exhaustion and Health is 8.15%, and between Academic Efficacy and Health is 5.02%. The partial correlation for variables Mental Exhaustion and Academic Efficacy when controlling the effect of health is 19.91%, proving that the relationship is diminished. This analysis showcased that medical students' **Academic performance is negatively affected by Mental Exhaustion**.

[5] Measures a linear correlation between 2 variables
[6] Confidence Interval states the probability of output parameter to fall between set of two values
[7] coefficient of determination, R^2