**Week 3 – In2research placement**

Descriptive statistics using : Dataset 2: <https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete> (CC0: Public Domain)

**Boxplots, Skewness and kurtosis** Chart, box and whisker chart

Description automatically generated

***Figure 1****: Boxplot demonstrating global life expectancy from the years 2000-2019 for males and females.*

Boxplot’s can be analysed by skewness which describes the asymmetry of the data whereas kurtosis describes distributions, leptokurtic indicates the data peaking higher than that of the usual bell curve (usual bellcurve can be described as mesokurtic, value=3), while a lower and flatter peak than the normal bell curve is known as platykurtic.

Chart, diagram

Description automatically generated

***Figure 2****: shows expected trends for analysing data by kurtosis*

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

***Figure 5*** *: > skewness(vec3) [1] -0.9983953 > kurtosis(vec3) [1] 2.560413, the*

*skewness is negative here is negative suggesting the distribution is left skewed, the kurtosis is less than 3 indicating the data is platykurtic. Performing the jarque.test(), this is the* ***Jarque-Bera Normality Test*** *and it tells us whether the sample data has a skewness and kurtosis that matches normal distribution data, results:*

*JB = 32.747, p-value = 7.748e-08, alternative hypothesis: greater. Therefore we can reject the null hypothesis as the sample does not match normal distribution as there is sufficient evidence to suggest that the data has a skewness and kurtosis different from normal distribution.This data is representing the globe, I chose not to facet based on country as there is not enough sample data per country to provide sufficient conclusions.*

**Performing T-tests**

**Using the one sample t-Test** – This type of statistical test is done to compare the means of two different groups of data, this can be done in the format of a one-sample, two-sample or paired t-test.

I have been looking at the WHO records for life expectancy data from the years of 2000-2019 and began with performing a one sample t-Test to see whether the life expectancy of citizens within Africa deviates far from an estimation of the mean of 60 years. Using the t-Test, I want to see if my mean differs significantly from this value.

Text, letter

Description automatically generated

***Figure 3:*** *These figures tell us 1) we can say with 95% confidence that our mean for the dataset lies between 55.45007 and 64.18743 . 2) The p value is very large and is larger than the 5% threshold of 0.0250 as we are performing a two tailed test so we accept the null hypothesis as our estimated mean is very close to our calculated value. 3) the t value here is very small indicating there is very insignificant difference between the estimated and calculated mean value. Changing my estimation for life expectancy to Africa to 40 years lead to a drastic decrease in p value and a great increase in t value consequently.*

**Using the two sample t-Test**: this statistical test is used to determine if two population means are equal, our null hypothesis here assumes that the difference in mean value for life expectancy between years 2000-2019 of Africa vs South-East Asia is zero.

Text, letter

Description automatically generated

***Figure 4:*** *Our p value here is less than 2.5% therefore we can reject the null hypothesis, therefore our mean values for each region are statistically different as p values indicate the likelihood that the difference is due to chance.*

**Plotting Correlation graphs**

Chart, scatter chart

Description automatically generated

A picture containing table

Description automatically generated

***Figure 6:*** *This scatter plot is explaining the correlation between the doctors available per 10,000 citizens and the percentage probability of being diagnosed with cancer between the ages of 30-70. I performed a correlation test in R to obtain Pearson’s product-moment corelation and I obtained a value of -0.470689 indicating negative correlation between the two variables, they are not dependent on each other, there are however a few anomalies that could suggest otherwise.*

*A picture containing chart

Description automatically generated*

*Chart

Description automatically generated*

*Figure 7 and 8: The correlation matricies above demonstrate the correlation between many variable pairs including doctors available per 10000 individuals, road traffic deaths, cancer cases and basic sanitation available. The legand explains the correlation coefficients calculated between these variable sets. The porsitive values demonstrate positive correlation and the negative values demonstrate negative correlation. The correlation values have been calculated using global data.*