INTRODUCTION

Continuing from the previous assignment, we are tasked with using the powerful statistical analysis program SAS and its proprietary language, SAS Language, to employ several statistical techniques to obtain a deeper understanding of some dataset.

In Assignment Three we are tasked with finding out the 95% confidence intervals, a type of interval estimate computed from the statistics of the observed data which produces a range of potential values of the unknown population parameter. A 95% Confidence Interval doesn't mean that there is a 95% probability that the population parameter lies within the interval, however, it relates to how reliable the estimate is.

WORK STRESS IMPACT ON PERSONAL LIFE

In the first question, we have been told that 75% of restaurant employees said that work stress had a negative impact on their personal lives from a national survey. We, subsequently, have procured a Sample of 100 employees of some restaurant chain – containing the answer to the question "Does work stress have a negative impact on your personal life?". The sample found that 68 employees said yes.

We have been tasked with finding the 95% CL of the Sample data. As the nature of the question is yes or no, we are unable to use the UNIVARIATE procedure but instead must use the FREQ procedure which we can specify that the data is Binomial – consisting of two terms. Within the FREQ procedure we also specify the order of the variables, as we want the 95% CL to be constructed using the "YES" variable.

Binomial Proportion	
Stress = YES	
Proportion	0.6800
ASE	0.0466
95% Lower Conf Limit	0.5886
95% Upper Conf Limit	0.7714
Exact Conf Limits	
95% Lower Conf Limit	0.5792
95% Upper Conf Limit	0.7698

Table 1 - 95% CL for Impact on Personal Life from Work Stress

As we can see from the results SAS has produced, there are two 95% CL. The first, ASE, is the Normal Approximation and the second, Exact Conf, which is calculated numerically. Looking at both 95% CL's, we can see the range roughly lies between 0.58 and 0.77 which is interpreted as with a reliability of 95% we can say that between 58% and 77% of employees feels that work stress has a negative impact on their personal lives.

In part b of the question we are asked if there is any evidence that the proportion differs from the national proportion. We have been told that from the national survey; 75% of employees feel that work stress has a negative impact on their personal lives. Comparing this to the ranges produced by SAS we can see that 75% lies in both 95% CL's, suggesting that the sample obtained from the restaurant chain is somewhat similar to the national survey. Having said that, the national surveys 75% is closer to both upper limits but does lie within both. Looking at the ASE range, 58.68% - 77.14% feel work stress has a negative impact on their personal lives whilst the Exact Conf range states, 57.92% - 76.98% of employees feel work stress has a negative impact on their personal lives.

CODE

```
/*Create Data Set for Analysis*/
DATA StressEmp;
      /*Create 2 Variables - Stress (Y/N) & Percentage*/
      INPUT Stress$ Per;
      /*Put Values in Variables*/
      DATALINES;
      YES 68
      NO 32
      ;
RUN;
/*Call the FREQ Procedure */
/*Using ORDER = DATA to get the YES value*/
PROC FREQ DATA=StressEmp ORDER=DATA;
      /*Specifying the data is Binomial*/
      WEIGHT Per;
      TABLES Stress / BINOMIAL;
RUN;
```

AMOUNT SPENT IN CIMEMA COMPLEX

In question 2, we have been given 12 randomly selected values which represent the amount spent (in £) by a customer in a cinema complex on one night. We have been tasked with finding the 95% CI for the amount spent by customer.

Using SAS, we can use several procedures to compute the 95% CI, such as the CIBASIC command on the UNIVARIATE procedure and the TTEST procedure. I will employ the TTEST procedure to compute the 95% CI as this procedure produces a series of plots to analyse several features.

Looking at the results, we are first shown a statistical summary of the data:

N	Mean	Std Dev	Std Err	Min	Max
12	5.5667	2.2519	0.6501	1.9600	10.5000

Table 2 - Statistical Summary of Data

As expected, there are 12 entries with a Mean value of 5.5667, a Standard deviation of 2.2519 and ranges from 1.96 – 10.50. From this data we can conclude that the majority of values are clustered around 3 – 7 but the data set's range extends that showing either a large distribution or some outliers.

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
5.5667	4.1359	6.9975	2.2519	1.5952	3.8235

Table 3 - 95% CL for Mean and Std Dev

The next table produced by SAS's TTEST procedure is the 95% CI computation results. The Ttest produces 95% CL's for both the Mean value and Standard deviation. To interpret this table, we can say that the average amount of money spent at the cinema complex is 5.5667 as discovered before, however, we can add to the statement suggesting with a 95% reliability that the average spending at the complex is in-between £4.14 and £7.0. Furthermore, the table states that the Standard deviation is 2.2519 and has a 95% CL of £1.6 – £3.8

The next output from the TTEST procedure is a plot of a histogram and a boxplot to show the distribution of the sample data along with the 95% confidence interval for the mean value:

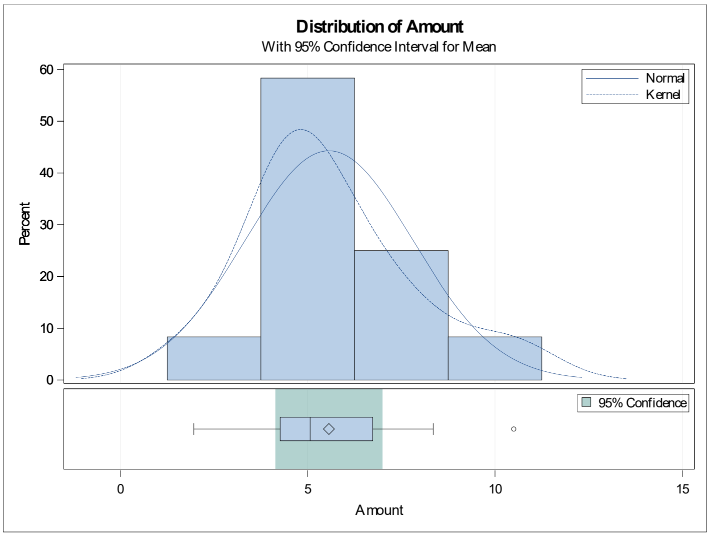

Figure 1 - TTEST Analysis of the Distribution of the Data

Looking at the distribution, as expressed before, the majority of values are clustered around £3-£7 and the dataset contains an outlier. Below the histogram, plotted on the same axis, is a box plot which shows where the 95% CI lies in the distribution. We can see that where the majority of values are is where the 95% CI is, with the whole IQR being encompassed by the 95% CI.

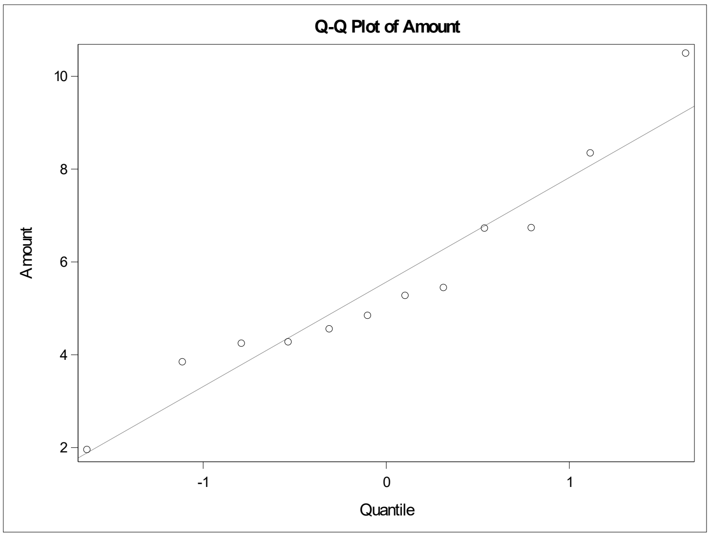

Figure 2 - QQPlot of the Amount with Normal

The final plot produced by the TTEST procedure is the QQPlot, Quantile-Quantile Plot, which plots the estimated quantile to the theoretical quantiles of the Normal Distribution – when the plot lies on the line there is an agreement with the normal distribution.

Looking at the QQPlot produced, we can see that two point lie on the line thus showing some agreement with the normal distribution. This suggests that the data follows somewhat of a normal distribution, with some outliers.

The second part of the question asks whether the sample data provides evidence that the amount spent by customers is £5. If we look at the mean, £5.57, it suggests that this is the case. Furthermore, if we look at the 95% CI for the mean value, we can see that it ranges from £4.14 - £7 which also encompasses the £5 value. This suggests that the amount spent by customers at the cinema complex is £5.

CODE

```
/*Create Data Set for Analysis*/
DATA CinemaSpend;
	/*Create a variable - Amount*/
	INPUT Amount;
	/*Put Values in Variables*/
	DATALINES;
	3.85
	5.28
	6.74
	1.96
	4.85
	4.28
	6.73
	4.56
	5.45
	8.35
	10.50
	4.25
	;
RUN;
/*Call the TTEST Procedure */
PROC TTEST DATA=CinemaSpend;
/*Specifying the Variable*/
	VAR Amount;
RUN;
```

CONCLUSION

In this assignment, we were tasked with finding the Confidence Interval of two different types of datasets. In question one, we were given the percentage of employees who felt that work stress negatively impacted their personal lives from a national survey and wanted to see if a sample collected from a restaurant differed. Using SAS's FREQ procedure, using the order key word to focus on the yes values as well as stating that the data was binomial - we found that the sample was similar to the national survey with the 75% lying in the 95% CI.

In the second question, we were given a dataset of amounts spent at a cinema complex and wanted to find the 95% CI of the mean value. Using SAS's TTEST procedure, we were able to produce several plots to show the distribution of the data along with statistical summaries and the 95% CI for both the mean and Standard Deviation.