MATH 4044 – Statistics for Data Science

Practical Week 3

Exercise 1

Data file for this exercise is based on a sample of 50 emails stored in a SAS data file called `email50.sas7bdat` located in `mydata` library on the SAS OnDemand server. The data statement to access this file is `data=mydata.email50`

Some of the variables in that file are as follows:

Variable	Description
spam	Specifies whether the message was spam; 0 = no, 1 = yes
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables or links, which would indicate the message is in html format; 1 = html, 0 = text
number	Indicates whether the email contained no number, a small number (under one million) or a large number; none = no number, small = number under one million, big = large number

(a) Use PROC MEANS to obtain 95% confidence intervals for the population mean number of characters in emails, overall and by format (text or html). Interpret those confidence intervals in words. Were the conditions for inference satisfied? Explain briefly. [You can use Tasks or write your own code.]

(b) We wish to test the hypothesis that an average email contains 10,000 characters. Set up the hypotheses and nominate the significance level. Use PROC UNIVARIATE to obtain appropriate output. Interpret and report your results. Were the conditions for inference satisfied? Explain briefly. [You can use Tasks or write your own code.]

(c) Repeat part (c) for plain text and html format emails separately using PROC TTEST. [You can use Tasks or write your own code.]

(d) Consider the variable *num_char*. Carry out the log transformation to get a new variable *log_char* = log(*num_char*) and discuss the Normal goodness of fit. Compare the untransformed and transformed distributions and discuss the impact of the transformation. Repeat the above comparisons using a square root transformation to create a new variable *sqrt_char* = sqrt(*num_char*). Which transformation seems more appropriate?

You can use the following code to create the required new variables:

```
data work.email50_transf;       /* Define new data set */
    set mydata.email50;
    log_char=log(num_char);     /* Define new variable */
    sqrt_char=sqrt(num_char);   /* Define new variable */
run;
```

Use the new data set to generate output needed to make the requested comparisons.