MATH 4044 – Statistics for Data Science

Practical Week 3 Solutions

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.]

The MEANS Procedure

Analysis Variable : num_char					
N Obs	N	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Std Error
50	50	11.598	7.868	15.328	1.856

Analysis Variable : num_char						
format	N Obs	N	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Std Error
text	13	13	2.308	0.117	4.499	1.006
html	37	37	14.862	10.291	19.434	2.254

Figure 1: PROC MEANS confidence limits output for variable *num_char*

Consider first the number of characters overall.

We are 95% confident that the true population mean number of characters in an email is between 7.87 and 15.33 thousand, or 11.6 thousand ± 3.73 thousand if we choose to quote the margin of error.

Assuming a simple random sample, we had large enough sample size ($n = 50 > 30$) to proceed with constructing a confidence interval for the population mean.

Consider now plain text emails. Based on the SAS output in Figure 1, we are 95% confident that the true population mean number of characters in a plain text email is between 117 and 4.5 thousand.

As the sample size is small (n = 13 < 30), we need to check Normality. Skewness and kurtosis measures in Figure 2 indicate lack of symmetry and heavy tails. The P-P plot in Figure 4 shows a curved pattern consistent with a right skewed distribution. The Kolmogorov-Smirnov test results in Figure 3 indicate that the number of characters in plain text emails , D(13) = 0.29, P-value < 0.01, is significantly non-Normal.

The UNIVARIATE Procedure
Variable: num_char
format = text

Moments			
N	13	Sum Weights	13
Mean	2.30815385	Sum Observations	30.006
Std Deviation	3.62568062	Variance	13.14556
Skewness	2.80522766	Kurtosis	8.70427736
Uncorrected SS	227.005184	Corrected SS	157.74672
Coeff Variation	157.081411	Std Error Mean	1.00558288

Figure 2: Moments for variable *num_char* when *format* = *text*

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.29339989	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.30028063	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.76200692	Pr > A-Sq	<0.005

Figure 3: Results of goodness-of-fit tests for variable *num_char* when *format* = *text*

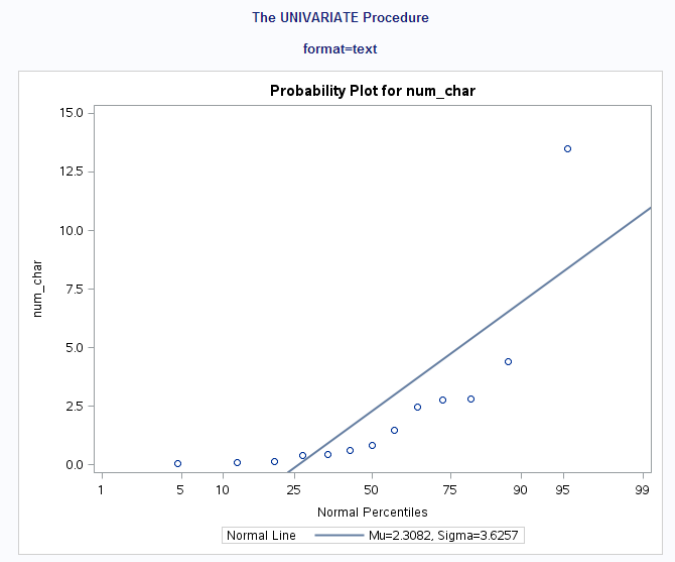

Figure 4: P-P plot for variable *num_char* when *format* = *text*

For the html emails, we are 95% confident that the true population mean number of characters is between 10.29 and 19.43 thousand. In contrast to plain-text emails, we

had a large enough sample (n = 37 > 30) to proceed with constructing a confidence interval for the population mean.

Moments			
N	37	Sum Weights	37
Mean	14.8622973	Sum Observations	549.905
Std Deviation	13.7107318	Variance	187.984166
Skewness	1.74176912	Kurtosis	3.7039182
Uncorrected SS	14940.2816	Corrected SS	6767.42996
Coeff Variation	92.2517663	Std Error Mean	2.25403042

Figure 5: Moments for variable *num_char* when *format* = *html*

It should be however be noted that the distribution of the number of characters in html emails is quite severely skewed (see sample skewness in Figure 5) and non-Normal (see goodness-of-fit results in Figure 6 and P-P plot in Figure 7). While the Central Limit Theorem says that as the sample size n increases (approaches infinity in fact), the sampling distribution of the mean becomes more Normal regardless of the shape of the population distribution, a sample of size much larger than 30 may be required to achieve a close approximation to Normality. For strongly skewed distributions, inference results therefore need to be treated with caution.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.21755082	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.33175758	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.91285379	Pr > A-Sq	<0.005

Figure 6: Results of goodness-of-fit tests for variable *num_char* when *format* = *html*

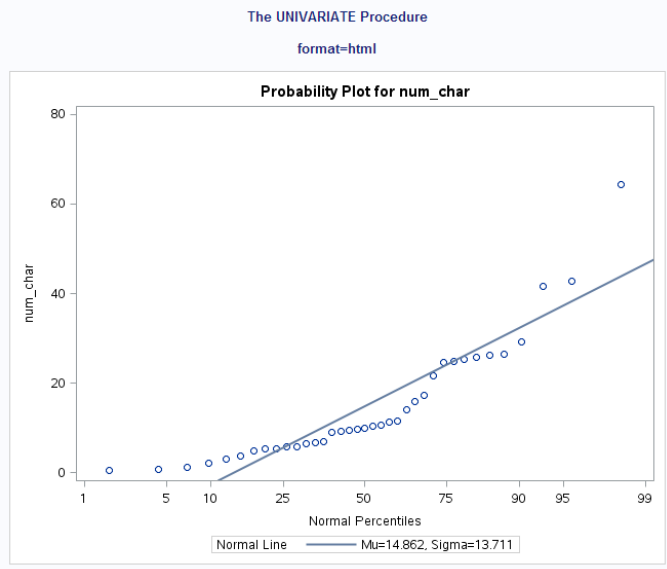

Figure 7: P-P plot for variable *num_char* when *format* = *html*

(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.]

We shall use the two sided alternative and 5% significance level:

$H_0 : \mu = 10$

$H_1 : \mu \neq 10$

$\alpha = 0.05$

Assuming a simple random sample, the sample size is large enough to proceed with a one-sample *t*-test ($n = 50 > 30$).

The output of SAS UNIVARIATE procedure is shown below in Figure 8. The test statistic is $t = 0.8610$ with 49 degrees of freedom. Since the *P*-value = 0.3934 > 0.05, H_0 can't be rejected.

Tests for Location: Mu0=10				
Test	Statistic		p Value	
Student's t	t	0.861021	Pr > |t|	0.3934
Sign	M	-6	Pr >= |M|	0.1189
Signed Rank	S	-44.5	Pr >= |S|	0.6719

Figure 8: *t*-test output for variable *num_char*

Conclusion: At 5% significance level, there is not enough statistical evidence to conclude that the population mean number of characters in emails is different from 10 thousand.

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

Consider first plain text emails. We shall again use the two sided alternative and 5% significance level:

$H_0 : \mu_{text} = 10$

$H_1 : \mu_{text} \neq 10$

$\alpha = 0.05$

The output of SAS TTEST procedure is shown below in Figure 9. The test statistic is $t = -7.65$ with 12 degrees of freedom. Since the *P*-value < 0.0001 < 0.05, H_0 is rejected.

Conclusion: At 5% significance level, there is enough statistical evidence to conclude that the population mean number of characters in text emails is different from 10 thousand.

The TTEST Procedure

Variable: num_char

format=text

N	Mean	Std Dev	Std Err	Minimum	Maximum
13	2.3082	3.6257	1.0056	0.0570	13.5020

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2.3082	0.1172	4.4991	3.6257	2.5999	5.9850

DF	t Value	Pr > |t|
12	-7.65	<.0001

Figure 9: *t*-test output for variable *num_char* when *format = text*

However, the sample size is small (n = 13 < 30) and the Q-Q plot in Figure 10 below indicates a curved pattern which suggests that the data does not come from a population that follows a Normal distribution. While our conclusion is probably correct – text emails typically have fewer than 10 thousand characters – a *t*-test was not appropriate for this data. A non-parametric alternative should have been considered instead.

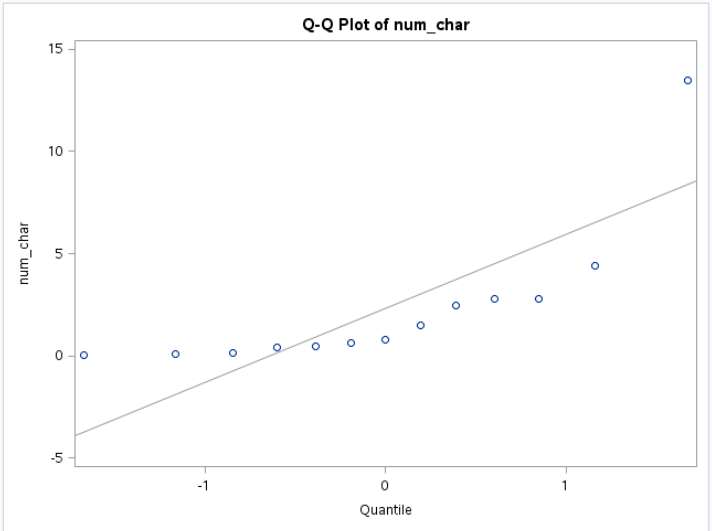

Figure 10: Q-Q plot for variable *num_char* when *format = text*

Consider now html emails. We shall use the two sided alternative and 5% significance level:

$H_0 : \mu_{html} = 10$

$H_1 : \mu_{html} \neq 10$

$\alpha = 0.05$

Assuming a simple random sample, the sample size is large enough to proceed with a one-sample *t*-test (n = 37 > 30).

The output of SAS TTEST procedure is shown below in Figure 11. The test statistic is $t = 2.16$ with 36 degrees of freedom. Since the P-value = 0.0377 < 0.05, H_0 is be rejected.

Conclusion: At 5% significance level, there is enough statistical evidence to conclude that the population mean number of characters in html emails is different from 10 thousand. The 95% confidence limits in Figure 11 suggest in fact that the population mean number of characters in html emails may be higher than 10 thousand.

The TTEST Procedure

Variable: num_char

format=html

N	Mean	Std Dev	Std Err	Minimum	Maximum
37	14.8623	13.7107	2.2540	0.4930	64.4010

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
14.8623	10.2909	19.4337	13.7107	11.1497	17.8097

DF	t Value	Pr > |t|
36	2.16	0.0377

Figure 11: *t*-test output for variable *num_char* when *format = html*

(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.

Consider first the log transformation. Skewness and kurtosis measures shown in Figure 12 indicate a distribution that is left-skewed and with heavier tails than a Normal distribution. This is confirmed by the histogram in Figure 13. The P-P plot in Figure 15 shows a curved pattern consistent with a left skewed distribution with heavy tails. The Kolmogorov-Smirnov test results in Figure 14 indicate that the log-transformed number of characters is still non-Normal, D(50) = 0.14, P-value = 0.021 < 0.05.

The UNIVARIATE Procedure
Variable: log_char

Moments			
N	50	Sum Weights	50
Mean	1.61515165	Sum Observations	80.7575823
Std Deviation	1.63659038	Variance	2.67842808
Skewness	-0.965218	Kurtosis	0.53362394
Uncorrected SS	261.678718	Corrected SS	131.242976
Coeff Variation	101.327351	Std Error Mean	0.23144883

Figure 12: Moments of log-transformed variable *num_char*

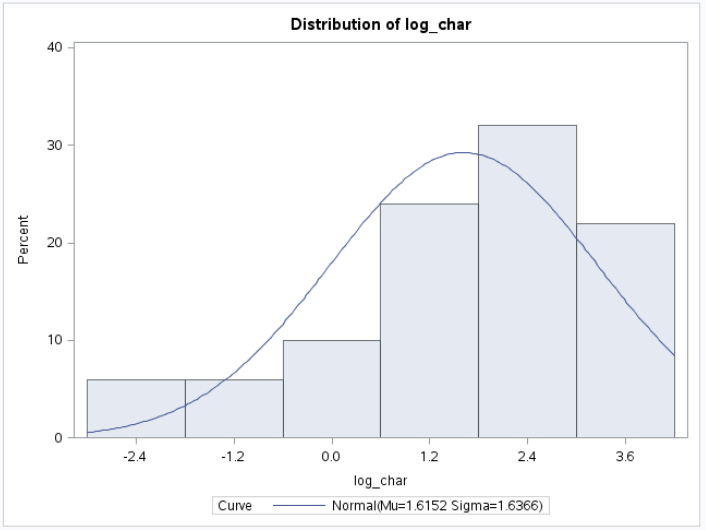

Figure 13: Histogram of log-transformed variable *num_char*

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.13591100	Pr > D	0.021
Cramer-von Mises	W-Sq	0.20141991	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.19014525	Pr > A-Sq	<0.005

Figure 14: Results of goodness-of-fit tests for log-transformed variable *num_char*

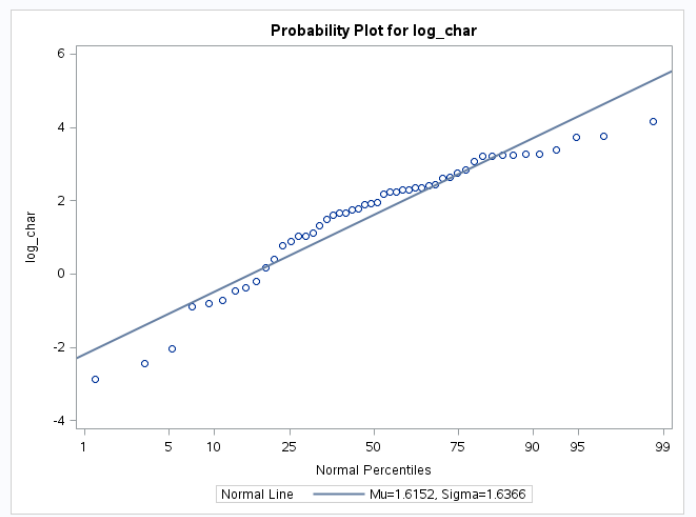

Figure 15: P-P plot of log-transformed variable *num_char*

Consider now the square root transformation. Skewness and kurtosis measures shown in Figure 16 indicate a distribution that is slightly right-skewed. This is confirmed by the histogram in Figure 17. The P-P plot in Figure 19 shows a reasonably linear pattern. The Kolmogorov-Smirnov test results in Figure 18 indicate that the square root-transformed number of characters can be assumed to be Normal, D(50) = 0.09, P-value > 0.15.

The UNIVARIATE Procedure
Variable: sqrt_char

Moments			
N	50	Sum Weights	50
Mean	2.89779544	Sum Observations	144.889772
Std Deviation	1.80729857	Variance	3.26632813
Skewness	0.64225655	Kurtosis	0.07314552
Uncorrected SS	579.911	Corrected SS	160.050078
Coeff Variation	62.3680521	Std Error Mean	0.25559062

Figure 16: Moments of square root-transformed variable *num_char*

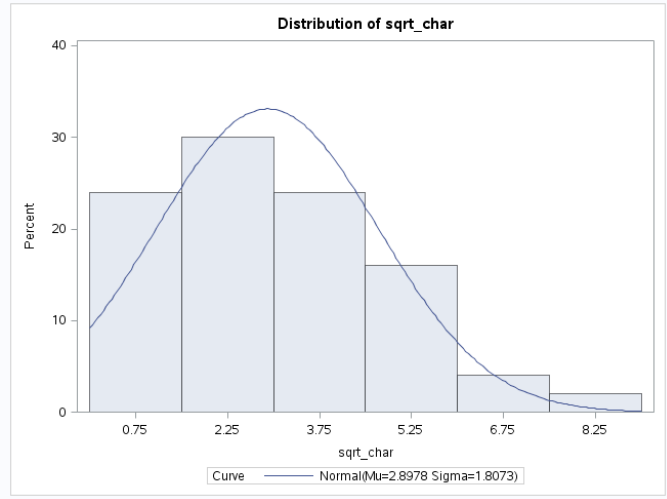

Figure 17: Histogram of square root-transformed variable *num_char*

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.09139787	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.07801241	Pr > W-Sq	0.222
Anderson-Darling	A-Sq	0.55548686	Pr > A-Sq	0.148

Figure 18: Results of goodness-of-fit tests for square-root-transformed variable *num_char*

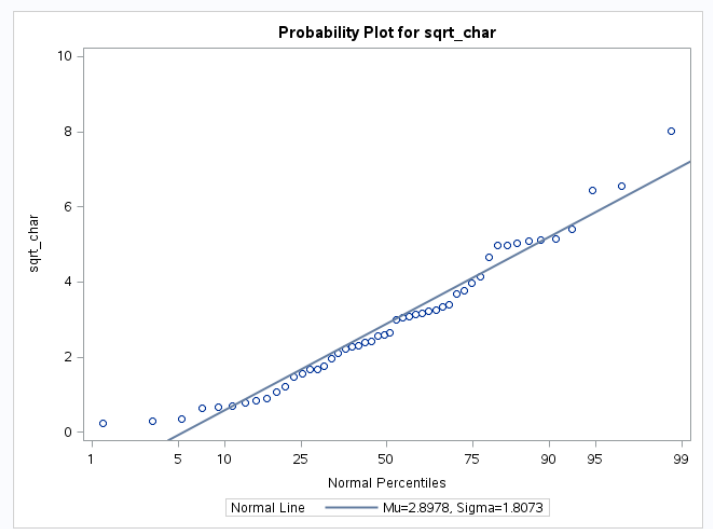

Figure 19: P-P plot of square root-transformed variable *num_char*

Conclusion:
Square-root transformation was more appropriate; the log-transformation was too strong for this data.

Appendix – SAS code

```
/* Defining formats in order to make output easier to read */

proc format ;
    value TypeF 0='text' 1='html';
run;

/* Part (a) */

proc means data=mydata.email50 n mean clm stderr maxdec=3
printalltypes;
    format format TypeF.;
           /* Note overuse of 'format'! It is a variable name and
also a statement */
    var num_char;
    class format;
run;

/* Sorting the data in column num_char by format, in order to
use 'by' statement in PROC UNIVARIATE and TTEST */

PROC SORT DATA=mydata.email50(KEEP=num_char format)
     OUT=WORK.email50_sorted; /* Creating an output file with
sorted data */
    BY format;
RUN;

proc univariate data=work.email50_sorted;
                /* Using the temporary file with sorted data */
    var num_char;
    format format TypeF.;
     by format;   /* Generating results by email type */
    histogram / normal;
    probplot / normal(mu=est sigma=est);
run;

/* Part (b) */

proc univariate data=mydata.email50 mu0=10;
    var num_char;
run;

/* Part (c) */

PROC TTEST DATA=WORK.email50_sorted alpha=0.05 H0=10 sides=2;
                      /* Using the temporary file with sorted data
*/
    VAR num_char;
    format format TypeF.;
    BY format;  /* This statement requires the data to be already
sorted
                           according to email type (text or html) */
RUN;
```

```
/* Part (d) */

data work.email50_transf;  /* Creating transformed data */
    set mydata.email50;
    log_char=log(num_char);
    sqrt_char=sqrt(num_char);
run;

proc univariate data=email50_transf;
    var log_char sqrt_char;
    histogram / normal;
    probplot / normal (mu=est sigma=est);
run;
```