End of Term Report Instructions

Your End of Term Report should be written in a style similar to a report that you would write for an employer, that is, it should be well organized and well-written in complete sentences. More marks will be awarded to submissions that are consistent with the style of a written report. As a reminder, you should not include the questions in your report, nor should you answer in a 'question and answer' format unless instructed to do otherwise. Your report should be structured in paragraphs and sentences.

Your report must be typed. You may create your document in Word, Google Docs, LaTeX or any other word processor. Numerical and graphical summaries must be obtained using R.

Uploading your End of Term Report

(1) You must upload your End of Term Report according to the instructions given on Crowdmark to facilitate marking. Please review the Crowdmark instructions now and budget an appropriate amount of time for the final upload of your work. Do not leave uploading your work to Crowdmark to the final hour before it is due, as the upload may take some time.

(2) You must upload your End of Term Report to the appropriate LEARN Dropbox as **one pdf file** so that your submission may be evaluated using plagiarism detection software.

(3) You must upload a file of the R code (including comments) that you used to do the report in the appropriate LEARN Dropbox.

Important Note: The End of Term Report must use the same dataset as your Midterm Report. Please ensure that the dataset you used for the Midterm Report is in the Midterm Dataset Dropbox and that it is in csv format.

Penalties will be applied if the above instructions are not followed. Reports which are not typed will not be marked.

What to Do

For your End of Term Report you will continue to analyse the dataset that you analysed as part of your Midterm Report. Please check the Midterm Report Instructions which explained the context of this study. Your manager has now asked you to complete a second report, following the one you completed for the Midterm. Note that this report will be shared with colleagues who did not see your Midterm Report, and so you should not presume someone reading this is familiar with your Midterm Report. **Note: if you believe it is appropriate, you may re-use passages from your Midterm Report in writing your End of Term Report.**

Recall that the information regarding the dataset is summarized in the LEARN documents Assignment Dataset Information (Coursework Submission 1 folder) and Midterm Dataset Information (Midterm Report folder). You will need these documents, which include the descriptions of the variates, to complete the End of Term Report.

Important: Your report should be structured with sections for Problem, Plan, Data, Analysis, and Conclusion. Each section should contain the items specified below. These items represent the minimum set of details and information that your report should include. As a reminder, you should not include the questions or your R code in your report. Instructions in Crowdmark indicate how the steps are to be uploaded to Crowdmark for marking purposes.

Problem

Define the target population for the study.
Note: This is the same as the target population for the Midterm Report. You may re-use or include a corrected version of your answer from the Midterm Report.

In this study you are asked to analyse the following variates: `urls.binary`, `username`, `time.of.day`, `likes`, `retweets`, `long.words`, `hashtags.binary`, `media.binary`, `is.retweet`.

For each variate specify its type and give an attribute in the *target population* for this variate. (The attribute must refer to the *target population*.)
Note: Some of these variates were used in the Midterm Report. You may re-use or include a corrected version of your answer for these variates from the Midterm Report.

Example:
The variate `media` is a *discrete* variate. An attribute of interest is the mean number of media items (images or videos) used in tweets in the *target population*.

The Problem step includes the motivating questions for the study which are stated in terms of the attributes of the *target population*.

By looking at the analyses you are asked to do in the Analysis step below:

- construct **one** motivating question for the analysis described in (1)
- construct **two** motivating questions for the analysis described in (2)
- construct **two** motivating questions for the analysis described in (3)
- construct **one** motivating question for the analysis described in (4)
- construct **one** motivating question for the analyses described in (5)
- construct **one** motivating question for the analysis described in (6)

These questions must be stated in terms of attributes in the *target population*. Specify the type of problem (descriptive, predictive) for each of these 8 motivating questions. Your motivating questions must include **both** descriptive and predictive problems. At least two of the questions must be predictive problems.

In addition, explain why these data may not be used to examine any causative problems.

Note: You may present your motivating questions in a list as shown below. **This is only being allowed in this step of your report for ease of marking.**

Example:
Two motivating questions for analysis (_) are:
What is the mean number of media items in tweets in the *target population*? This is a descriptive problem.
If a tweet is chosen at random from the *target population* what is the probability it contains no tweets? This is a predictive problem.

Plan

Define the study population for the study.
Note: This is the same as the study population for the Midterm Report. You may re-use or include a corrected version of your answer from the Midterm Report.

Give one example of **study error** related to the variate `time.of.day`.
Note: Be sure to reference an attribute when giving your example of study error.

Describe the sampling protocol for this study in as much detail as possible. Be sure to indicate the sample size for your dataset.
Note: This is the same as the sampling protocol for the Midterm Report. You may re-use or include a corrected version of your answer from the Midterm Report.

Identify one possible source of measurement error related to the variate `long.words`.

Data

In $1-2$ sentences, discuss any concerns or issues that you have in obtaining your sample (dataset).
Note: This is the same Data step as for the Midterm Report. You may re-use or include a corrected version of your answer from the Midterm Report.

Analysis

Complete the analyses given below. You should provide as much explanation as you think necessary for your manager to understand your report. Be sure to write in complete sentences. Since this is a report, your analyses should not contain any detailed mathematical derivations or calculations. All tables and plots require titles/labels as appropriate.

Note that none of the variates used in the End of Term Report should have missing values (which may show as NA). If your dataset contains missing values in these variates then please review the instructions posted to LEARN on October 18 regarding using the read.table() function in R. If you continue to encounter missing values then please ask for help on Piazza.

Note: The numbering below is only used to help with instructions for uploading the parts to Crowdmark for ease of marking. You should not include this numbering in your report.

(1) Your dataset contains the variate `urls.binary`.

Let Y = the number of tweets which contain at least one url. Assume Y has a Binomial(n, θ) model where n is the number of observations.

Explain why the Binomial model is a reasonable model to use for this variate.

The parameter θ corresponds to what attribute in the study population?

Give the maximum likelihood estimate and 15% likelihood interval for θ.

A recent study estimated that approximately 19% of tweets on Twitter contained urls. Test the hypothesis $H_0 : \theta = 0.19$ using the likelihood ratio test statistic. **(Include the observed value of the test statistic, the approximate distribution of the test statistic, and the p-value.)**

(2) Your dataset contains the variates `username` and `time.of.day`. The variate `time.of.day` is measured in seconds. Use this variate to create a new variate called `time.of.day.hour` which is the time of day the tweet was published on Twitter, expressed as hours after midnight.

If a $G(\mu_O, \sigma_O)$ model is assumed for the variate `time.of.day.hour` for tweets for @ONThealth, what attribute does the parameter μ_O correspond to in the *study population*?

If a $G(\mu_A, \sigma_A)$ model is assumed for the variate `time.of.day.hour` for tweets for @GoAHealth, what attribute does the parameter μ_A correspond to in the *study population*?

Use numerical and graphical summaries to decide how well the Gaussian model fits the data for @ONThealth. **(You must include at least one graphical summary.)**

Use numerical and graphical summaries to decide how well the Gaussian model fits the data for @GoAHealth. **(You must include at least one graphical summary.)**

Give a 99% confidence interval for σ_O. Give a 99% confidence interval for σ_A. Based on these intervals, is it reasonable to assume $\sigma_O = \sigma_A$?

Based on your answer to whether it is reasonable to assume $\sigma_O = \sigma_A$, test the hypothesis $H_0 : \mu_O = \mu_A$. Your answer should include a point estimate of the mean difference, a 95% confidence interval for $\mu_O - \mu_A$, and a $p - value$ for the test of the hypothesis $H_0 : \mu_O = \mu_A$. You should also explain what assumptions you have made in carrying out these steps.

Your manager will be presenting the results of this study to a group of colleagues who are not statisticians. As part of this analysis, your manager has asked you to include an explanation of the meaning of the hypothesis $H_0 : \mu_O = \mu_A$ in the context of this study. Your manager has also asked you to include an explanation of the evidence based on the data regarding the hypothesis. Both of your explanations should be phrased in terms that non-statisticians would understand.

(3) Your dataset contains the variates `retweets` and `likes`. For this analysis you will need to create the transformed variate `retweets.log=log(retweets+1)` and `likes.log=log(likes+1)`.

What are the least squares estimates of α and β if you fit a simple linear regression model $y = \alpha + \beta x$ to your data where $x =$ `retweets.log` is the explanatory variate and $y =$ `likes.log` is the response variate?

Check the adequacy of the simple linear regression model by using three plots: (1) a scatterplot of the data including the fitted line, (2) a plot of the standardized residuals versus the fitted values, and (3) a qqplot of the standardized residuals. (**Include your plots with your analysis.)**

The parameter β corresponds to what attribute in the *study population*?

Give a 95% confidence interval for the parameter β.

Using only the 95% confidence interval for β, what can you say about the $p - value$ for testing the hypothesis $H_0 : \beta = 1$?

Briefly explain why the hypothesis $H_0 : \beta = 1$ might be of interest.

For a randomly chosen tweet in the *study population* with 30 retweets, give a point estimate and a 90% prediction interval for the number of likes.

(4) Your dataset contains the variate `long.words`.

Assuming a Poisson(θ) model for the variate `long.words`, give the maximum likelihood estimate of θ and an approximate 95% confidence interval for θ based on the asymptotic Gaussian pivotal quantity.

Use the likelihood ratio goodness of fit test to test the hypothesis that a Poisson model is reasonable for these data. (**Include a table of observed and expected values, the observed value of the test statistic, the approximate distribution of the test statistic, and the p-value.)**

Give a grouped barplot of the observed and expected frequencies that you used for testing the goodness of fit.

For a randomly chosen tweet in the study population, give a point estimate of the probability that this tweet contains at least 2 long words. (You may assume the Poisson model fits these data to answer this question.)

(5) Your dataset contains the variates `hashtags.binary` and `media.binary`.

Test the hypothesis of independence between these two variates using the likelihood ratio test statistic. **(Include a two-way table of observed and expected frequencies, the observed value of the test statistic, the approximate distribution of the test statistic, and the p-value.)**

As mentioned previously, your manager will be presenting the results of this study to a group of colleagues who are not statisticians. As part of this analysis, your manager has asked you to include an explanation of the meaning of the hypothesis of independence in the context of this study. Your manager has also asked you to describe the relationship between these two variates based on the observed data. Both of your explanations should be phrased in terms that non-statisticians would understand.

(6) Your dataset contains the variates `username` and `is.retweet.`

Let Y_j be the number of tweets which are retweets from other accounts for provincial health agency i, $i = 1, 2, \ldots, 8$. Assume that Y_i has a Binomial(n_j, θ_j) distribution, $j = 1, 2, \ldots, 8$ where n_j is the number of tweets sampled from provincial health agency j and θ_j is the probability a randomly chosen tweet from the tweets in the study population for provincial health agency j is a retweet. Assume the order @GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth for the health agencies

As part of this analysis your manager has asked you to include an explanation of the meaning of the hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_8$ that they can use in their presentation to a group of colleagues who are not statisticians. Your explanation should be phrased in terms that non-statisticians would understand.

Give the maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_8$.

If you assume $\theta_1 = \theta_2 = \cdots = \theta_8 = \theta$, what is the maximum likelihood estimate for θ?

Use the Pearson goodness of fit test to test the hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_8$. (**Include a table of observed and expected values, the observed value of the test statistic, the approximate distribution of the test statistic, and the p-value.)**

Conclusion

The conclusions you make should be based on the Analysis step and should address the motivating questions you created in the Problem step.

Conclusions can only be made in relation to the *study population*.

Your conclusion should address any statistical uncertainty in your answers to the motivating questions.

Important Note: Since different markers mark different sections, we are asking you to include your motivating questions from the Problem step in the Conclusion step. You may present the motivating questions and the conclusions in a list as shown below. **This is only being allowed in this step of your report for ease of marking.**

Example:

What is the mean number of media items in tweets in the *target population*?
If a tweet is chosen at random from the *target population* what is the probability it contains no tweets?

The mean number of media items in tweets in the *study population* is estimated to be 1.273 which is the sample mean. The uncertainty in this estimate is reasonably small since the interval estimate for the mean is $[0.773, 1.773]$ which has width equal to 1 which is quite narrow.
An estimate of the probability that a tweet, chosen at random from the *study population*, contains no tweets is 0.514. The uncertainty in the estimate is reasonably small since the sample size for this study is $n = 924$ is large.

Give two limitations of the study.
Note: These are the same limitations as for the Midterm Report. You may re-use or include a corrected version of your answer from the Midterm Report.