OEIT6, Data Analytics. Spring 2022

ISE: Assignment

Note: Any relevant dataset can be selected for this assignment, Emdodo link will be broadcasted as per given due date. Your file name should be in UID_Part(I/II/III)_OEIT6.pdf format. Late submission will not be accepted. Read assignment very carefully and see the relevant deliverables. Good luck!

Part 1: Exploratory Data Analysis (5M)

Due date: 5 Mar 22'@5p

Introduction

Data exploration (aka exploratory data analysis, or EDA) and display is a fundamental process of data analysis. EDA is used to summarize your data, to visualize patterns in your data, and to refine your hypotheses, while data display presents those patterns to others. While EDA is often done 'on the fly', and with low-resolution graphics or print-outs, data display is 'presentation-quality' graphics, analogous to what you'd read in a standard scientific presentation, and is what should end up in your thesis or independent project. For this assignment, you will do some background reading in EDA and data display, and you will begin to summarize and illustrate your data.

Perhaps the most common and convenient way to summarize data is to report measures of location, spread (error), and confidence. Examples of measures of location are the mean, trimmed mean, median, and mode; examples of measures of spread are the standard deviation, standard error, variance, percentiles, range, and coefficient of variation; and confidence is usually expressed as a k% confidence interval or k% prediction interval. If your data fall into obvious groups (treatments), then summaries are usually reported for each group.

Summary statistics can be reported in tabular form ('Tables') or graphic form ('Figures'). While tables are more precise, figures are usually more compelling.

Assignment (bring it to class)

1. Begin to explore and illustrate your data. You should prepare as many graphs as you think appropriate to illustrate your hypotheses, in rough (EDA) form.

2. Compute summary statistics for your variables of interest. If appropriate, compute them 'by' categories of interest. Use Statistics --> Data Summaries --> Summary Statistics for computation. Write a one-page summary of your summary statistics, that shows that you understand the meaning and differences between different measures of location, spread, and confidence.

3. Plot your summary statistics in (a) way(s) that enables rapid comparison between or among groups of interest. Produce at least three different types of plots illustrating your summary statistics (example: box plots, bar charts, and category plots). Remember: pie charts are not allowed! Write a one or two paragraph description of your plots, in standard scientific style, drawing the reader's attention to the results that you think are most relevant.

4. Using the confidence intervals that you computed in part A, discuss, in 1-2 paragraphs the apparent similarities or differences among your different treatment groups.

During class time, you will each present your EDA graphics and data summaries for critique. Be prepared to explain why you chose the graphic types that you did, and why they illustrate your hypotheses. Be

prepared to help each other improve the clarity of your graphics. About a third of the class time will be devoted to presentation and critique, while the remainder will be allotted to improving your graphics.

Part II. Probability distributions and hypothesis testing (10M)

Due date: 12 Mar 22'@5p

1. Goodness of fit test

Introduction

One of the first steps in analyzing data is assessing the underlying distribution(s) of the variable(s) of interest. For example, coin flips can yield two possible outcomes: 'heads' or 'tails'; a long run of coin flips (of fair coins) gives rise to a binomial distribution of data. There are many other distributions that underlay common phenomena: the Poisson distribution and the Gaussian (or 'normal') distribution are two of the more common. Many basic statistical calculations, as well as most statistical tests used are based on the assumption that the sampled population (not the sample itself) has a known probability distribution (usually normal) that can be parameterized (hence the use of the term 'parametric' statistics). In this work, you will explore the distribution(s) of your variable(s).

We will also use your data distribution to introduce you to (or re-aquaint you with) hypothesis testing. Most of you are (or should be) familiar with standard hypothesis testing and P-values; these give rise to the oft-asserted (and routinely mis-used) 'significance' of your data. You will use the distributions, and measures of location and spread that you developed in Assignment part I to test formal hypotheses about the shape of your data distributions. The type of statistical test that we will use for this assignment is a goodness-of-fit test.

The familiar Chi-square test is an example of a goodness-of-fit test.

Assignment (bring it to class)

1. Plot your variables in a way that you can visualize their distribution. Useful graphs for visualizing data distributions include: histograms, box-plots, dot-plots, and stem-and-leaf plots. S-Plus will do all of these (Graph --> 2D Plots), although stem-and-leaf plots can only be done from the command prompt (use the command: stem(variable)).

2. Describe (in 1-2 paragraphs) what distribution(s) ought to be the best fit(s) for your variable(s). Explain why you expect these distributions to be the appropriate ones.

3. Use a one-sample goodness-of-fit test to determine if your data actually fit the distribution you predicted in (2.).

4. Generate a simulated dataset with the same number of observations as your raw data, whose values, for each simulated variable, come from a simulated distribution that you predicted in part I.

5. Plot your simulated variables, and visually compare the two plots. Do you have a good match? If not, why not? Use a two-sample goodness of fit test to determine if the simulated data and your real data are statistically indistinguishable (use Statistics --> Compare Samples --> Two Samples}

6. If your data are not 'normally' distributed, can you transform them so that they fit a normal distribution? An example is the logarithmic transformation: new variable = ln(old variable) for data that are left-skewed.

7. Write a one-page exposition describing this first attempt at hypothesis testing. Use accurate language in describing your null and alternative hypotheses and state your conclusions in appropriate statistical terminology. Refer to your figures when appropriate.

2. Statistical Power

1. Determine the statistical power of one of the goodness-of-9.fit tests you conducted in above work.

2. Write 1-2 paragraphs that illustrate that you understand the difference between statistical significance and statistical power, and their relationships to your data distributions.

Part III: Correlation and Regression (10M)

Due date: 20 Mar 22'@5p

Introduction

Probably the most common statistical procedures used are correlation and linear regression (for data measured on a continuous scale), and analysis of variance (ANOVA) for comparing mean responses among more than two treatment groups (for two treatment groups, use the familiar t-test or its non-parametric equivalent). Regression and ANOVA are usually discussed together, because Fisher demonstrated that all degrees of freedom and sums of squares (i.e., deviations from overall or within-group means) in an ANOVA problem are reducible to single-degree-of-freedom contrasts analyzable by regression.

For both regression and ANOVA, you must specify the independent variable (continuous in regression, discrete [categorical] in ANOVA). The independent variables are assumed to be measured without error. For correlation, the assumption is that both variables were measured with error, and that there is no obvious 'independent' variable. Before setting out to do one of these statistical procedures, make sure that the procedure is appropriate for the question being asked, the data are structured appropriately, and the data conform to the necessary assumptions.

Assignment (bring it to class)

1. Calculate pair-wise correlations (Statistics --> Data Summaries --> Correlations) and simple linear regression statistics (Statistics --> Regression --> Linear) for any pair of variables. Note your data are amenable to correlation or regression analysis,

a) choose a pair of variables and compute the correlation or regression statistics

b) check that the data meet the assumptions of correlation/regression. Illustrate that you have, in fact, checked the data.

c) if the data do not meet the assumptions, transform them appropriately, and do (a) again.

d) if you can't find an appropriate transformation, compute the Spearman's correlation coefficient based on

2. Determine if your regression or correlation statistics are 'significant'.

3. Write up what you've done. Be sure that you not only present your results and show that you've checked assumptions, but also that you interpret the meaning of all output parameters. One or two pages ought to be sufficient.