Exploratory Data Analysis and Design of Experiment

Exploratory Data Analysis (EDA) is the first, and crucial, part of the data science/data analysis process. Plots, graphs and summary statistics are the fundamental EDA tools. John Tukey, a mathematician who worked at Bell Labs among other places, pioneered EDA techniques (he literally wrote a book called "Exploratory Data Analysis" in 1977). Hypotheses about relationships within the data are not explicitly specified at the beginning of the EDA process; we are just exploring the data and finding anything interesting or useful. EDA is primarily utilized to generate summary statistics, from which we can get an idea of the general trends in the data. It can be used for verifying/disproving assumptions, examining the relationships among variables, detecting mistakes, and looking for outliers, among other things.

Normally, the data is in the form of tables such as a spreadsheet or database. The row represents an experimental subject, whereas the columns signify subject identifiers, and values are explanatory and outcome variables. Nevertheless, eyeballing the spreadsheet to search for the important patterns in the data is usually impossible. EDA techniques have been developed to facilitate this tedious situation.

Seltman [2014] emphasizes on understanding the type of a variable. His reasons are 1) types of the information being collected should be recognized 2) a variable type restricts the appropriate statistical models 3) which analysis is applicable for the data given the restricted models. The data are classified into two main types, each with two sub-types as follows:

- Quantitative variables
    - Discrete variables
    - Continuous variables
- Categorical variables
    - Nominal variables
    - Ordinal variables

EDA can be displayed as non-graphical and graphical methods. The non-graphical methods are based on calculations of statistical summaries. While, the graphical methods represent data using pictorial or diagrammatic approaches. The non-graphical and graphical methods can be further divided into univariate (one variable) and multivariate (two or more variables) methods. It is recommended to perform the univariate EDA before multivariate EDA.

In summary, EDA consist of 4 types: univariate non-graphical, multivariate non-graphical, univariate graphical and multivariate graphical [Seltman, 2014].

Sampling

We must be wary of any sampling bias that we may incur in our data. "Sampling bias" means we have collected data from a sample of our population that isn't representative of the full population. Drawing a marble from a jar of 100 marbles randomly is an example

of sampling. If the marbles in the jar are equal parts red, green, and blue, and we draw 100 blue marbles, that would be an extremely baised sample.

If we keep doing this many times (drawing 100 marbles), this can reveal the underlying distributions in the population. Population refers to the total number of things we are measuring (in the marble example, it would be all marbles in the jar), from which a sample is selected – a sample is a subset of a population. A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population.

Univariate non-graphical EDA

Univariate non-graphical EDA is just like it sounds – one variable, and no graphics. This type of EDA focuses on sample distribution, which is the implication of the population distribution. Outlier detection plays a role in this analysis.

For categorical data, the suitable EDA technique is a table of frequency with each category percentages (proportion or percentage of the data).

For quantitative data, it is worth to look at sample statistics such as sample mean, sample variance, sample skewness, and sample kurtosis. These values may be different from sample to sample. Nevertheless, they represent some uncertain information about the underlying population and its parameters.

Central tendency signifies the location of a distribution, which is related to the middle values such as mean, median, and mode. This covers geometric, harmonic, and other means. In any symmetrically shaped distribution, the means is the point where the symmetry holds. In the case of non-symmetric distributions, the mean is the balance point. Median is the middle value of an ordered list. If there are an even numbers of values, median is the average of the two middle values. Median has a property called robustness since changing the upper or lower half of the data may have no affect on changing the median. Mode is the most frequently occuring value. Typically, it is used to describe a distribution with a single peak (unimodal) or two or more peaks (bimodal or multi-modal). In skewed distribution or outlier cases, the median is a preferred measurement. Otherwise, mean is the common used measurement.

Spread signifies how far are the data values from the central such as standard deviation, variance, and interquartile range. Additive is the important property of variances. For example, the sum of variances is derived from adding independent sources of variation together. The standard deviation is the square root of the variance. This additive property, however, is not found in standard deviation. The quartiles divide the distribution or observed data into one fourth evenly. Q1 (first quartile) signifies one quarter of the data is below Q1. Q2 (second quartile) means two fourths or half of the data is below Q2. Q3 (third quartile) is the three-fourths data below Q3. The interquartile range (IQR) is calculated from IQR = Q3-Q1. For normal distribution, the standard deviation is approximately 3/4 of the IQR.

Skewness is a measure of asymmetry. Kurtosis is a measure of peakedness, which usually compares with the shape of normal distribution.

Univariate graphical EDA

Again, just like it sounds – one variable, and this time we make graphics to represent the statistics (graphs, plots).

The histogram is the most basic graph. It offers a general idea about the shape of the distribution including central tendency, spread, modality and outliers. The size of the bin may affect the shape of the distribution.

A pie chart is similar to a histogram, but is not often used.

Stem and leaf plot can be used to substitute for a histogram. However, the histogram is considered better for showing the data and shape of the distribution.

Boxplot is a rectangular box bounded above and below. It usually displayed in the vertical format, although it is also possible for horizontal format. Information about central tendency, symmetry, skew and outliers are represented well with boxplots. The data values beyond 1.5 IQRs in either direction may signify the outlier. Whereas, data beyond 3 IQRs may indicate extreme outliers. Be aware that outliers do not always imply problems. The perfect normal distribution is expected to be boxplot outliers 0.70%. In addition, boxplots depend on robust statistics such median and IQR instead of mean and standard deviation.

Quantile-normal plots or quantile-quantile (QQ plots) is typically used to investigate whether a sample follows any underlying distribution especially Gaussian distribution. Moreover, the plot can be used to examine left/right skew, positive/negative kurtosis and bimodality. Points should appear randomly on or around the diagonal line if the assumption about normality is not violated. If the points display the curved pattern instead of along the diagonal line, it is not normality. In the latter case, the values such as confidence interval, p-values are not valid.

Multivariate non-graphical EDA

Cross-tabulation is a two-way table with levels of one variable on the column and levels of another variable on the row. It is the basic form of non-graphical EDA for bivariate data.

Correlation and covariance are the basic statistics for two quantitative variables. In simple words, a sample covariance is a co-vary measurement (vary together) of two variables. Positive covariance means two variables vary in the same direction such as one measurement is above the average, the other will also above the average. For negative covariance, one variable varies in a different direction from the other variable such as when one variable is above the average, the other variable is below the average. When the values of covariance are close to zero, this indicates that two variables vary independently. Nevertheless, covariance is complicated to apprehend. Therefore, the correlation is often used. The positive correlation means it is positive linear correlation, while, the negative correlation means it is negative linear correlation. A correlation coefficient measures both strength and direction of the relationship. In addition, the correlation coefficient is standardized. This means that the perfect linear relationship has a coefficient of 1 (or -1). Covariance is not standardized, the perfect linear relationship depends on the data.

For two random variables *X* and *Y*, with means μ_x and μ_y , and standard deviation σ_x and σ_y respectively.

Covariance $\sigma_{xy} = E[(X - E[X])(Y - E[Y])]$

Correlation $\rho_{xy} = \frac{(¿\sigma_x \sigma_y)}{E[(X - E[X])(Y - E[Y])]/¿}$

Where E is the expected value.

Multivariate graphical EDA

Scatter plots are useful for two quantitative variables. Traditionally, the outcome variable (dependent variable) is placed on the y-axis, and the explanatory variable (independent variable) is on the x-axis. Side-by-side box plots are the recommended graphical EDA to explore the relationship between a categorical variable and a quantitative variable.

It is important to perform EDA so that you become familiar with the data (i.e. distribution, relationship between variables) before further analysis on that data.

Design of Experiment or Experimental Design (DOE)

The experimental design is important because it helps us to design the experiments to maximize information from the fewest experiments. So, the time needed for gathering the data is minimized. In addition, it helps to quantify the effect of particular factors on the dependent variable. For example, machines used in manufacturing can be adjusted to various settings. These settings have huge impact on the production quality. The concepts from experimental design can be used to optimize the manufacturing process. DOE methods have numerous applications in research, business, and industry.

The principle of experimental design is based on ‘induction’ where inference of the general conclusion comes from particular instances. For example, an instance can be a group of 2500 high school students who apply for a computer science program in the year 2015 of a particular state. The inference might be the average GPA of those who get accepted to the program.

Experimental design can be modeled as a process where inputs feed into the process and may affect the output. The inputs consist of controllable factors, uncontrollable factors, and noise factors. The controllable factors are measured and determined by the experimenter. The uncontrollable factors are measured but not determined by the experimenter. Finally, noise or error is unmeasured and uncontrolled factors.

In general, designing an experiment can be defined as the following steps [Hoff, 2009]

1. Identify the research hypotheses or objectives to achieve
2. Choose a set of experimental units
3. Determine the response or output variable. What do you want to observe?
4. Determine potential sources of variation in the response
5. Decide which variables (factors, levels) to measure and control
6. Decide how treatments are to be randomly assigned

In summary, DOE main purposes are: to design an optimal experiment and to analyze the results from the experiment.

Terminology Reviews:

- *Response variable* is the output variable. It is measurement of the outcome of the experiment.
- *Factor variable* is the input variable, also known as predictor variable (independent variable). It is the variable that affects the response variable.
- *Levels* are subdivision (different levels) of a factor that can be controlled.
- *Treatment* is a combination of factor levels.
- *Primary factors* the factors, which their effects need to be measured.
- *Secondary factors* are the factors in which we are not interested.
- *Replication* repeated of the experiments.
- *Experimental design* plan for experimentation such as number of experiments, number of factors, etc.
- *Experimental unit* refers to any entity (i.e. person, animal, plant, time period) employed in the experiments.
- *Interaction* is the change of one level effecting the change in the other level.

Replication is essential in reducing the effect of uncontrolled variation so that the precision increases. Randomization is needed to avoid bias and allows later use of probability theory and statistical analysis.

Three major types of Experimental design

1. Completely randomized design
2. Randomized block design
3. Factorial design

1. Completely randomized design is a design, in which one primary factor is randomly assigned to the experimental units. So for example, we may want to test if different levels of a fertilizer affect plant growth. We would vary the amount of fertilizer (primary factor), and measure the plant growth for different groups of plants (experimental units). The experiment objective is to compare means of *k* treatments. Is there a significant difference between the *k* treatments? When k > 2, this is also known as one-way ANOVA.

This experiment considers a single factor. Its goal is to compare the treatment. The response variable (y) is assumed to be a quantitative number. Comparing 2 treatments is a special case of completely randomized design. For example, comparing the effect of two fertilizers on the yield of blueberries.

The assumptions are need for the test to be valid include 1) samples are randomly and independently selected from the *k* treatment populations 2) All populations are approximately normal distribution 3) The population variances are equal. If the assumptions are not valid, use a non-parametric statistical method: Kruskall-Wallis.

Brief note on hypothesis testing

A hypothesis is a proposed explanation for some effect or trend, based on limited evidence or intuition. In statistics, we typically have a null hypothesis (H_0), and an alternative hypothesis (H_a). Often we are comparing the mean (average) of some

measurement of groups, for example, the average height of men and women in the US. Typically, the null is that the means are equal, and the alternative is that the means are not equal. Statisticians will usually speak in terms of rejecting or accepting the null. If we reject the null hypothesis, it means the alternative must be true. If we accept the null, this is usually referred to as either rejecting or not rejecting the null. Hard-core statisticians would never say 'we accept the alternative hypothesis'; instead they would say 'we cannot reject the null hypothesis'. To see if we can reject the null, we usually use a test statistic, which is calculated from the data. If the statistic meets a certain threshold, we can then reject the null.

The hypothesis for one-way ANOVA can be written as:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$
H_a: *at least two treatment means are different*

The test statistic is $F = \frac{MST}{MSE}$
(MST is the mean square treatments; MSE is the mean square error)

In other words, the null hypothesis is that the means of all the groups are equal.

Another way to write this is $F = \frac{MS_B}{MS_W}$
Where MS_B is the mean-square between groups, and MS_W is the mean-square within groups. If we have 3 groups, MS_B can be calculated as:

$$MS_B = \frac{n(\acute{Y}1 - \acute{Y}) + n(\acute{Y}2 - \acute{Y}) + n(\acute{Y}3 - \acute{Y})}{f_b}$$

Where n is the number of samples per group, $\acute{Y}1$ is the mean value of measurements from group 1, and f_b is the between-groups degrees of freedom (number of groups minus 1).

The mean-square within groups is the sum of squares of every data point divided by the within-group degrees of freedom, $f_w = a(n-1)$, where a is the number of groups.

$$MS_w = \frac{\sum_i d_i}{f_w}$$

The rejection region is $F > F_\alpha$. Fa is the critical value of the F-statistic for our degrees of freedom, $F_{alpha}(f_b, f_w)$. We typically look this up in a F-table. We choose a significance level (usually alpha=0.05).

If the null hypothesis is rejected, a post-test may be necessary to compare which treatment pairs are ***different*** using methods such as Tukey, Bonferroni, and Scheffe'. The criteria for choosing the methods are based on treatment sample sizes (equal or unequal), and types of comparisons (pairwise, general contrasts). For k treatments, there are $k(k-1)/2$ pairs of means to compare.

For example [adapted from: Yau, 2014], A national chain restaurant wants to investigate 3 new menu items for marketing purposes. 21-chained restaurants are randomly selected in the study, in which, 7 of these restaurants are randomly selected to test the first menu. Another 7 restaurants are tested using the second menu, and the rest are tested on the third menu.

Suppose the upper case letters (A, B) are factors and lower case (y) is the numeric variables. Use *aov* function in R to perform completely randomized design as follows:
>result= aov(y~A , data=dataframe)
>summary(result)

Randomization in the experiment is important in reducing or eliminating the influence of the factors not considered in the experiments. Furthermore, it also validate the statistical analysis under the normality assumptions [Yu, 2006]

2. Randomized block design is a design, in which one primary factor is considered in the experiment. In addition, the experimental units are grouped into blocks (b). These blocks should contain experimental units that are very similar. Each block consists of k units, where k is the number of treatments. Each treatment is assigned to the experimental unit in each block in random order. So, there are altogether $n=bk$ responses.

Factors can be divided into 2 types: the first type is the effects, that are the primary interest to the experimenter and the second type is the effects that are desired to be removed. Typically, a blocking factor is considered when there is heterogeneity among the experimental units. In the same block, however, experimental units are homogeneous. This means that the variability between blocks is eliminated using block as an explicit factor. For example, 2 different types of fertilizers with 3 types of seeds are investigated. Fertilizer factor is the primary concern, whereas seed types are treated as a blocking factor.

The required assumptions for the tests include 1) the b blocks and k are randomly selected and applied 2) the distributions of the block-treatment bk are normal approximation 3) the variances of the block-treatment bk are equal.

The hypothesis, typically, can be written as:

$H_0: \mu_1=\mu_2=\ldots=\mu_k$
H_a: *at least two treatment means are different*

Test statistics is $F = \frac{MST}{MSE}$

The rejection region is $F > F_{\alpha}$ with DF= $(k-1), (n-b-k+1)$

For example [Adapted from: Yau, 2014], A national chain restaurant wants to investigate 3 new menu items for marketing purposes. Seven restaurants are randomly selected in the study. Each restaurant is test using all 3-menu items. Each week, only one menu item will be tested. Therefore, it takes 3 weeks to complete the testing. The order of the menu items for each restaurant is completely random.

The following is the example of using aov () function in R for randomized block design.
>result=aov (y~A+B, data=dataframe)
>summary(result)

If the assumptions are not valid for this randomized block design, use a non-parametric statistical method such as Friedman F.

3. Factorial Design is a design, in which more than one factors are considered in the experiment. The treatment levels are randomly assigned to the experimental units for every factor combination. This design is an effective way to study the interaction effects between factors. The experiment with two factors is known as two-way factorial design.
It is more efficient that doing only single factor experiments.

The main focus of two previous designs is on the effect of a single factor to the response variable. The factorial design concerns the effects of two or more factors and the interactions between factors. For example, a 3 x 4 factorial design means there are 2 factors. The first factor has 3 levels, the other factor has 4 levels. In this case, there 12 different treatment groups.

The required assumptions for the tests include 1) the distribution of the response on each factor level combinations is normal 2) the variance of the response is constant for all treatments 3) the experimental units are random and independently assigned for each treatment.

For example [adapted from: Yau, 2014], A national chain restaurant wants to investigate 3 new menu items for marketing purposes in the East and West coasts of the US. Twelve restaurants each from the East and West coasts are randomly chosen for the study. From twelve restaurants of the East coast, four of them are randomly assigned for testing the first menu, another four for testing the second menu and the rest for the third menu. Another twelve restaurants on the West coast repeat the same procedure as the East cost.

The hypothesis testing for the treatment means:

H_0: *There is no difference among all ab treatment means*
H_a: *at least two treatment means are different*

Test statistics is $F = \frac{MST}{MSE}$

The rejection region is $F > F_{\alpha}$ with DF= $(ab-1),(n-ab)$

The following is the example of using aov () function in R for factorial design.

```
> result = aov(y~A*B,data=dataframe) # interaction included
> summary(result)
```

Besides the treatment means testing, we can also test for factor interactions (factors A and B do not have an interaction affect for the response), test for the main effect of factor A (mean response is the same for each level of factor A), and test for the main affect of factor B (mean response is the same for each level of factor B).

Reference:

Hoff, D.P. (Dec 9, 2009), Statistics 502 Lecture Notes. University of Washington. Retrieved from: http://www.stat.washington.edu/hoff/courses/stat421-502/LectureNotes/notes.pdf

Miller, R. (2014), Chapter Ten: Analysis of Variance (Lecture Notes). Retrieved from: http://www2.fiu.edu/~millerr/Chapter%20Ten.pdf

Quick-R (2014), *ANOVA*. Retrieved from: http://www.statmethods.net/stats/anova.html

Seltman, J. H. (Nov 11, 2014). *Chapter 4: Exploratory Data Analysis*. Carnegie Mellon University. Retrieved from: http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

Trochim, M.K.W. (Oct 20,2006). *Experimental Design (Research Methods Knowledge Base)* Department of Policy Analysis and Management, Cornell University. Retrieved from: http://www.socialresearchmethods.net/kb/expfact.php

Wu, C & Chen, J. (2006) *Sampling and Experimental Design*. University of Waterloo. Retrieved from: http://sas.uwaterloo.ca/~jhchen/stat332/total.pdf

Yau, C. (2014) *Analysis of Variance* (R Tutorial. An R Introduction to Statistics) Retrieved from: http://www.r-tutor.com/elementary-statistics/analysis-variance