EGS 2405 Basic Statistics and Exploratory Data Analysis

This chapter seeks to ensure that readers have a sound understanding of the basic quantitative methods for obtaining and summarizing information on the environment.

1 Introduction

The data from an experiment are generally collected into a rectangular array (e.g., spreadsheet or database), most commonly with one row per experimental subject and one column for each subject identifier, outcome variable, and explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable. (Some more complicated experiments require a more complex data layout.)

People are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. They find looking at numbers to be tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation. Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either **non-graphical** or **graphical**. And second, each method is either **univariate** or **multivariate** (usually just bivariate).

Non-graphical methods generally involve the calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way. Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually, our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally, it will involve three or more variables. It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Beyond the four categories created by the above cross-classification, each of the categories of EDA have further divisions based on the role (outcome or explanatory) and type (categorical or quantitative) of the variable(s) being examined.

Although there are guidelines about which EDA techniques are useful in what circumstances, there is an important degree of looseness and art to EDA. Competence and confidence come with practice, experience, and close observation of others. Also, EDA need not be restricted to techniques you have seen before; sometimes, you need to invent a new way of looking at your data.

1.1 Univariate non-graphical EDA

The data that come from making a particular measurement on all of the subjects in a sample represent our observations for a single characteristic, such as age, gender, speed at a task, or response to a stimulus. We should think of these measurements as representing a “sample distribution” of the variable, which in turn more or less represents the “population distribution” of the variable. The usual goal of univariate non-graphical EDA is to better appreciate the “sample distribution” and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis.

Categorical data

The characteristics of interest for a categorical variable are simply the **range of values** and the frequency (or relative frequency) of occurrence for each value. (For ordinal variables it is sometimes appropriate to treat them as quantitative variables using the techniques in the second part of this section.) Therefore the only useful univariate non-graphical techniques for categorical variables is some form of **tabulation of the frequencies**, usually along with calculation of the fraction (or percent) of data that falls in each category.

For example categorizing subjects by College at Carnegie Mellon University as H&SS, MCS, SCS and "other" (Table 1.1.1). Taking a random sample of 20 students for the purposes of performing a memory experiment, the sample measurements can be listed as H&SS, H&SS, MCS, and other. The EDA will look as follows;

Table 1.1.1 Subjects categorization by College at Carnegie Mellon University

Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

Quantitative data

The most informative records are those for which the variables are measured fully quantitatively on continuous scales with equal intervals. For examples include the soils thickness, its pH, the cadmium content of rock, and the proportion of land covered by vegetation. Some such scales have an absolute zero, whereas for others the zero is arbitrary. Temperature may be recorded in kelvin (absolute zero) or in degrees Celsius (arbitrary zero). In most instances we need not distinguish between them. Some properties are recorded as counts, e.g. the number of roots in a given volume of soil, the pollen grains of a given species in a sample from a deposit, the number of plants of a particular type in an area. Such records can be analysed by many of the methods used for continuous variables if treated with care.

Properties measured on continuous scales are amenable to all kinds of mathematical operation and to many kinds of statistical analysis. They are the ones that we concentrate on because they are the most informative, and they provide the most precise estimates and predictions. The same statistical treatment can often be applied to binary data, though because the scale is so coarse the results may be crude and inference from them uncertain.

Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.
The characteristics of the population distribution of a quantitative variable are its center, spread, modality (number of peaks in the pdf), shape (including "heaviness of the tails"), and outliers. What we observe in the sample of measurements for a particular variable that we select for our particular experiment is the "sample distribution". We need to recognize that this would be different each time we might repeat the same experiment due to the selection of a different random sample, a different treatment

randomization, and different random (incompletely controlled) experimental conditions. In addition, we can calculate “sample statistics” from the data, such as the sample mean, sample variance, sample standard deviation, sample skewness and sample kurtosis. These again would vary for each repetition of the experiment, so they don't represent any deep truth, but rather represent some uncertain information about the underlying population distribution and its parameters, which are what we really care about. Many of the sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram. In most situations, it is worthwhile to think of univariate non-graphical EDA as telling you about aspects of the histogram of the distribution of the variable of interest. Again, these aspects are quantitative, but because they refer to just one of many possible samples from a population, they are best thought of as random (non-fixed) estimates of the fixed, unknown parameters of the distribution of the population of interest.

If the quantitative variable does not have too many distinct values, a tabulation, as we used for categorical data, will be a worthwhile univariate, non-graphical technique. But mostly, for quantitative variables we are concerned here with the quantitative numeric (non-graphical) measures which are the various sample statistics. In fact, sample statistics are generally thought of as estimates of the corresponding population parameters.

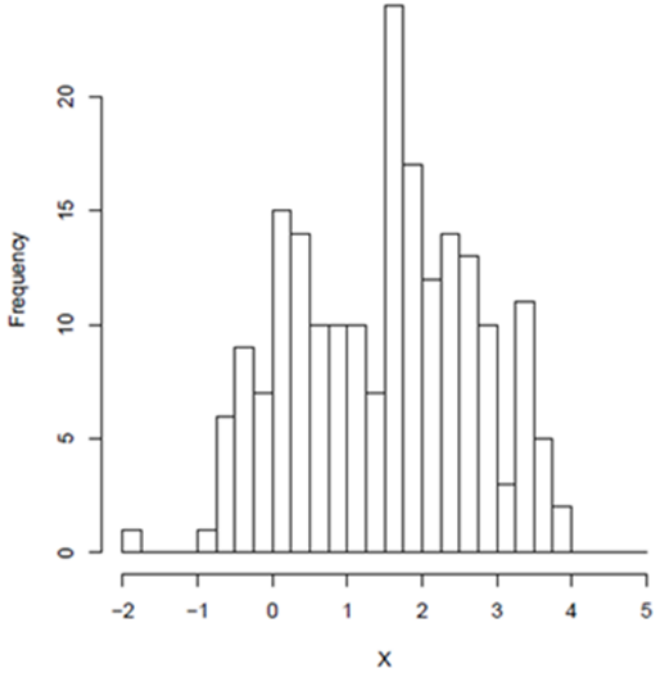

Figure 1.1.1: Histogram from distribution.

Figure 1.1.1 shows a histogram of a sample of size 200 from the infinite population characterized by distribution C. The bi-modality is visible, as is an outlier at X=-2. There is no generally recognized formal definition for outlier, but roughly, it means values that are outside of the areas of a distribution that would commonly occur. This can also be thought of as sample data values which correspond to areas of the population pdf (or pmf) with low density (or probability).
Therefore, set of measurements may be divided into several classes, and the number of individuals in each class counted. For a variable measured on a continuous scale, we can divide the measured range into classes of equal width and count the number of individuals falling into each. For quantitative variables (and possibly for ordinal variables), it is worthwhile looking at the central tendency, spread, skewness, and kurtosis of the data for a particular variable from an experiment.

1.1.1 The central tendency

Measures of central tendency provide information about where the center of a distribution is located. The most commonly used measures of center for numerical data are the **mean** and the **median** (mode is another measure of center and is the value that occurs most often in a sample).The mean is the simple arithmetic average: the sum of the values of a variable divided by the number of observations (calculated for interval data), as in (1.1.1):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Equation 1.1.1

where
n is the total number of observations
xi is the score of the ith observation
Σ is the symbol of summation (pronounced sigma)
$\bar{x}$ is the sample mean value
The mean takes account of all of the observations, it can be treated algebraically, and the sample mean is an unbiased estimate of the population mean.

The **median** is the value that divides the sorted scores form smaller to larger in half. It is a measure of center. There are as many values less than the median as there are greater than it. If a property has been recorded on a coarse scale then the median is a rough estimate of the true centre. Its principal advantage is that it unaffected by extreme values, i.e. it is insensitive to outliers, mistaken records, faulty measurements and exceptional individuals. Therefore, it is a robust summary statistic compare to the mean.
Relevant for interval/ratio and ordinal data. The median overcomes the outlier problem. When n is odd, we have a single median. When n is even, there are two "middle values," and in this case the average is taken.

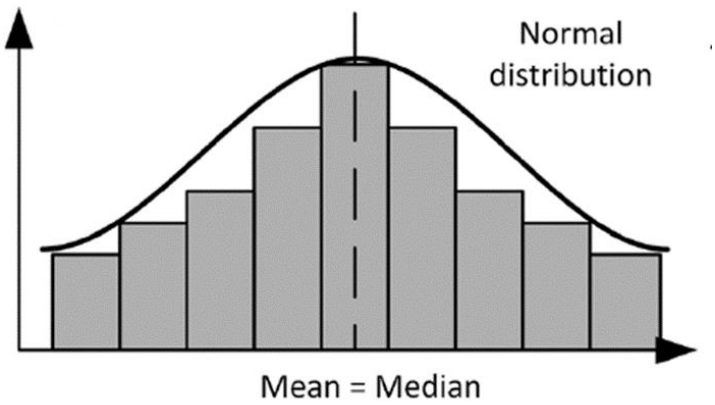

Figure 1.1.2 The median is located in the center of a normal distribution and coincides with the mean.

For symmetric distributions, the mean and the median coincide. For unimodal skewed (asymmetric) distributions, the mean is farther in the direction of the "pulled out tail" of the distribution than the

median is. Therefore, for many cases of skewed distributions, the median is preferred as a measure of central tendency.
The median has a very special property called robustness. A sample statistic is "**robust**" if moving some data tends not to change the value of the statistic. The median is highly robust, because you can move nearly all of the upper half and/or lower half of the data values any distance away from the median without changing the median. More practically, a few very high values or very low values usually have no effect on the median.
A rarely used measure of central tendency is the **mode**, which is the most likely or frequently occurring value. More commonly we simply use the term "mode" when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal). In symmetric, unimodal distributions (Figure 1.1.3), the mode equals both the mean and the median. In unimodal, skewed distributions, the mode is on the other side of the median from the mean. In multi-modal distributions, there is either no unique highest mode, or the highest mode may well be unrepresentative of the central tendency.

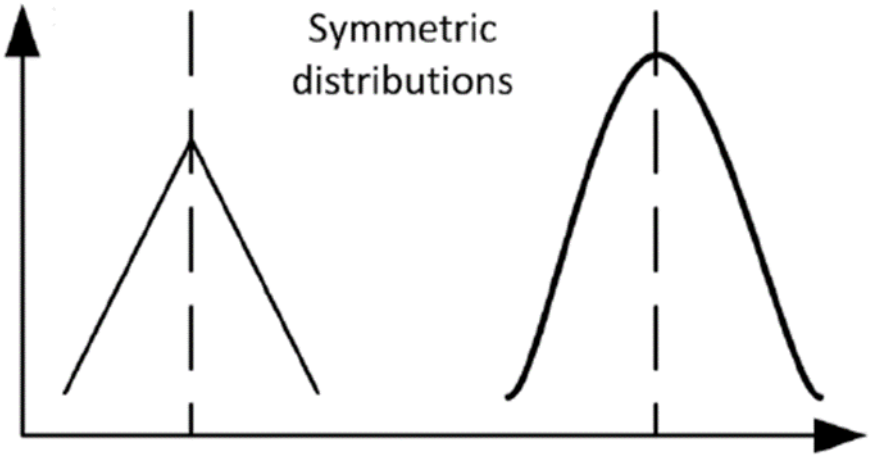

Figure 1.1.3 An example of a symmetric distribution

For a symmetric distribution, the mode, the mean and the median are, in principle, the same. For an asymmetric one.
(mode - median) ~ 2 * (median - mode)

Measures of dispersion

Several statistics are commonly used as a measure of the spread of a distribution, including variance, standard deviation, and interquartile range. Spread is an indicator of how far away from the center we are still likely to find data values.
The most common measures are as follows (de Smith 2018):

- Range
- Deviation from the mean
- Variance
- Standard deviation
- Coefficient of variation
- Percentiles and quartiles

The **range** is the difference between the largest and smallest values of the variable studied.
Range = xmax-xmin
Deviation from the mean is the subtraction of the mean from each score. Deviation=$(xi-\bar{x})$
The sum of all deviations is zero (sometimes, due to rounding up, the sum is very close to zero).

The **variance** is a standard measure of spread. The variance of a set of values, which we denote S^2, is by definition. The variance is the second moment about the mean. Like the mean, it is based on all of the observations, it can be treated algebraically, and it is little affected by sampling fluctuations. It is both additive and positive. To estimate population variance σ^2 from a sample, then N in Equation (1.1.2) is replaced by N-1.

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2$$

Equation 1.1.2

The **standard deviation** is simply the square root of the variance. Therefore, it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol S.

For the **coefficient of variation (CV)**, the SD is regarded in relative terms (relative variability). The CV is particularly useful when you want to compare results from two different surveys or tests that have different measures or values (hence different means)

$$CV = (SD/\bar{x}) * 100$$

Equation 1.1.3

The **quartiles** of a population or a sample are the three values which divide the distribution or observed data into even fourths. So one quarter of the data fall below the first quartile, usually written Q1; one half fall below the second quartile (Q2); and three fourths fall below the third quartile (Q3). The astute reader will realize that half of the values fall above Q2, one quarter fall above Q3, and also that Q2 is a synonym for the median. Once the quartiles are defined, it is easy to define the IQR as IQR = Q3-Q1. By definition, half of the values (and specifically the middle half) fall within an interval whose width equals the IQR. If the data are more spread out, then the IQR tends to increase, and vice versa.
The IQR is a more robust measure of spread than the variance or standard deviation. Any number of values in the top or bottom quarters of the data can be moved any distance from the median without affecting the IQR at all. More practically, a few extreme outliers have little or no effect on the IQR. In contrast to the IQR, the range of the data is not very robust at all.
If you collect repeated samples from a population, the minimum, maximum and range tend to change drastically from sample to sample, while the variance and standard deviation change less, and the IQR least of all. The minimum and maximum of a sample may be useful for detecting outliers, especially if you know something about the possible reasonable values for your variable. They often (but certainly not always) can detect data entry errors such as typing a digit twice or transposing digits (e.g., entering 211 instead of 21 and entering 19 instead of 91 for data that represents ages of senior citizens.) The IQR has one more property worth knowing: for normally distributed data only, the IQR approximately equals 4/3 times the standard deviation. This means that for Gaussian distributions, you can approximate the sd from the IQR by calculating 3/4 of the IQR.

Measures of shape

Two additional useful univariate descriptors are the skewness and kurtosis of a distribution. Skewness is a measure of asymmetry. Kurtosis is a measure of “peakedness” relative to a Gaussian shape.
Skewness

The skewness measures the asymmetry of the observations. It is defined formally from the third moment about the mean:

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^3$$

Equation 1.1.4

The coefficient of skewness is then

$$g_1 = \frac{m_3}{m_2\sqrt{m_2}} = \frac{m_3}{S^3}$$

Equation 1.1.5

Where m2 is the variance and S the standard deviation. Symmetric distributions have g1 = 0. N
Negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed. Positive skewness would indicate that the mean of the data values is larger than the median, and the data distribution is right-skewed.

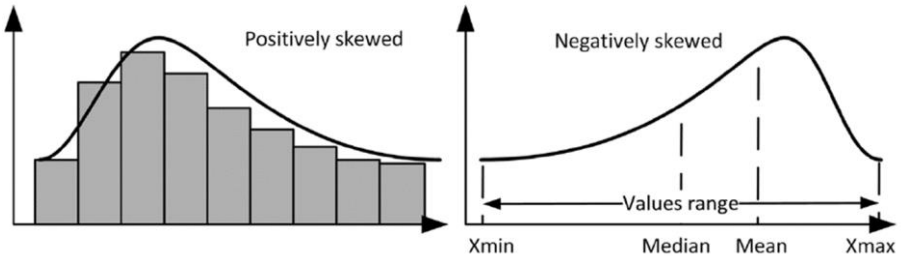

Figure 1.1.4 positively and negatively skewed distributions

Kurtosis
Kurtosis is obtained from the fourth moment about the mean

$$m_4 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^4$$

Equation 1.1.6

The coefficient of Kurtosis is given by

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{(S^2)^2} - 3$$

Equation 1.1.7

1.2 Univariate graphical EDA

If we are focusing on data from the observation of a single variable on n subjects, i.e., a sample of size n, then in addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data

1.2.1 Histograms

The only one of these techniques that makes sense for categorical data is the histogram (basically just a bar plot of the tabulation of the data). The resulting set of frequencies constitutes the frequency

distribution and its graph (with frequency on the ordinate and the variate values on the abscissa). The number of classes chosen depends on the number of individuals and the spread of values.

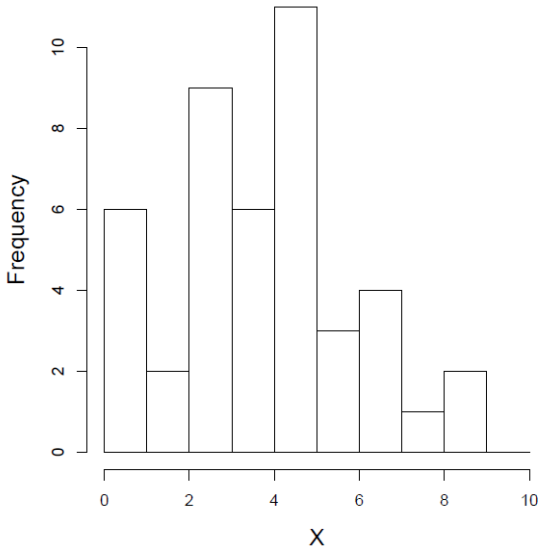

Figure 1.2.1 An example of Histogram

In general, the fewer the individuals, the fewer the classes needed or justified for representing them. Having equal class intervals ensures that the area under each bar is proportional to the frequency of the class. If the class intervals are not equal, then the heights of the bars should be calculated so that the areas of the bars are proportional to the frequencies.

1.2.2 Box plots

Another popular device for representing a frequency distribution is the **box plot** due to Tukey (1977). The boxplot will be described here in its vertical format, which is the most common, but a horizontal format is also possible. An example of a boxplot is shown in figure 1.2.2, of the same data as the histogram above.

Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers, although they can be misleading about aspects such as multimodality. The plain box and whisker diagram, e.g. in Figure 1.2.2, has a box enclosing the interquartile range, a line showing the median, and whiskers (lines) extending from the limits of the interquartile range to the extremes of the data, or to some other values such as the 90th percentiles.

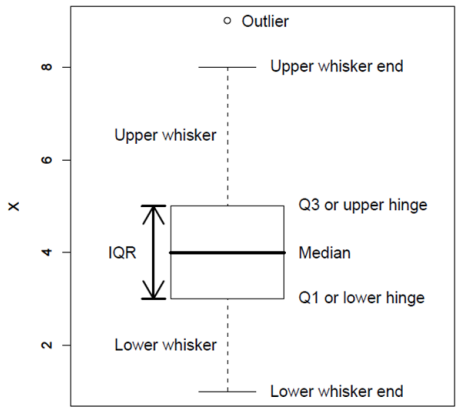

Figure 1.2.2 An example of Box plot

Both the **histogram** and the **box-plot** enable us to picture the distribution to see how it lies about the mean or median and to identify extreme values (outliers). Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

1.2.3 Quantile-normal plots

It is called the quantile-normal or QN plot or more generality the quantile-quantile or QQ plot. It is used to see how well a particular sample follows a particular theoretical distribution. Although it can be used for any theoretical distribution, we will limit our attention to seeing how well a sample of data of size n matches a **Gaussian distribution** with mean and variance equal to the sample mean and variance. By examining the quantile-normal plot we can detect left or right skew, positive or negative kurtosis, and bimodality.

The example shown in figure 1.2.3 shows 20 data points that are approximately normally distributed.

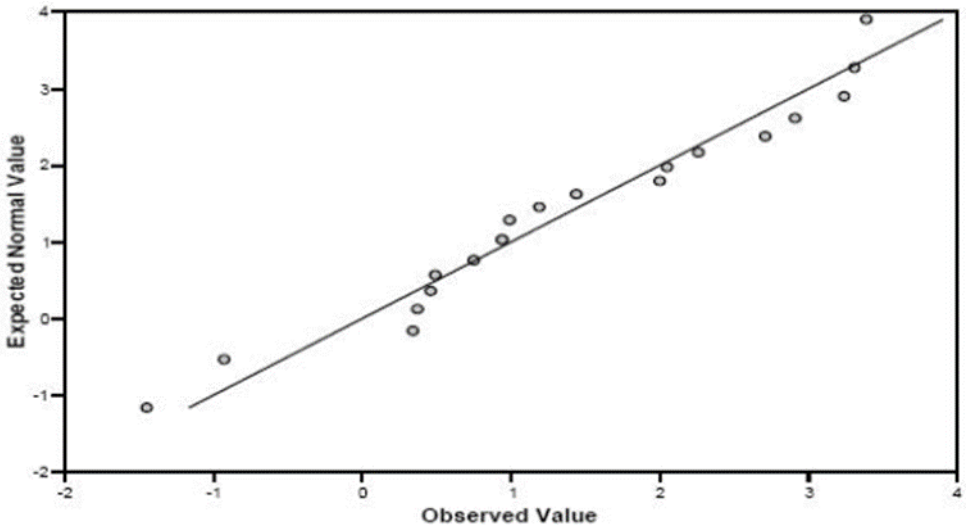

Figure 1.2.3 An example of QQ plot

NB: Do not confuse a quantile-normal plot with a simple scatter plot of two variables. The title and axis labels are strong indicators that this is a quantile-normal plot.

Many statistical tests have the assumption that the outcome for any fixed set of values of the explanatory variables is approximately normally distributed, and that is why QN plots are useful: if the assumption is grossly violated, the p-value and confidence intervals of those tests are wrong.

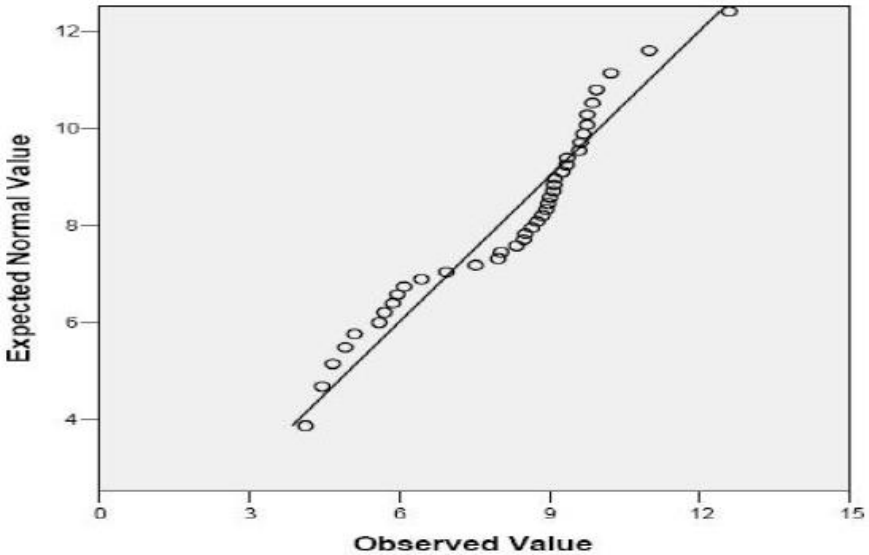

Figure 1.2.4 An example of Q-Q plot showing bimodality

1.3 Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

1.3.1 Cross-tabulation

For categorical data (and quantitative data with only a few different values) an extension of tabulation called cross-tabulation is very useful. For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels. The two variables might be both explanatory, both outcome and one of each. Depending on the goals, row percentages (which add to 100% for each row), column percentages (which add to 100% for each column) and/or cell percentages (Which adds to 100% of overall cells) are also useful. Table 1.3.2 shows the cross-tabulation.

We can easily see that the total number of young females is 2, and we can calculate, e.g., the corresponding cell percentage is 2/11 *100 = 18.2%, the row percentage is 2/5*100 = 40%, and the column percentage is 2/7*100 = 28.6%.

Cross-tabulation can be extended to three (and sometimes more) variables by making separate two-way tables for two variables at each level of a third variable.

Table 1.3.1 Sample Data for Cross-tabulation

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Table 1.3.2 Cross-tabulation for sample data

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

1.3.2 Correlation for categorical data

Another statistic that can be calculated for two categorical variables is their correlation. But there are many forms of correlation for categorical variables (find out more).

1.3.3 Univariate statistics by category

For one categorical variable (usually explanatory) and one quantitative variable (usually outcome), it is common to produce some of the standard univariate non-graphical statistics for the quantitative variables separately for each level of the categorical variable, and then compare the statistics across levels of the categorical variable. Comparing the means is an informal version of ANOVA. Comparing medians is a robust informal version of one-way ANOVA. Comparing measures of spread is a good informal test of the assumption of equal variances needed for valid analysis of variance.

1.3.4 Correlation and covariance

For two quantitative variables, the basic statistics of interest are the sample covariance and/or sample correlation, which correspond to and are estimates of the corresponding population parameters. The sample covariance is a measure of how much two variables "co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes. Covariance for a finite set of observations can be expressed as:

$$C_{1,2} = \frac{1}{N}\sum_{i=1}^{N}\{(z_1 - \bar{z_1})(z_2 - \bar{z_2})\}$$

Equation 1.3.1

Where $\bar{z_1}$ and $\bar{z_2}$ are the means of the two variables. This expression is analogous to the variance of a finite set of observations.

1.4 Multivariate graphical EDA

There are a few useful techniques for graphical EDA of two categorical random variables. The only one used commonly is a grouped bar plot, with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

1.4.1 Side-by-side boxplots.

When we have one categorical (usually explanatory) and one quantitative (usually outcome) variable, graphical EDA usually takes the form of "conditioning" on the categorical random variable. This simply indicates that we focus on all of the subjects with a particular level of the categorical random variable and then make plots of the quantitative variable for those subjects. We repeat this for each level of the categorical variable, then compare the plots. The most commonly used of these are side-by-side boxplots.

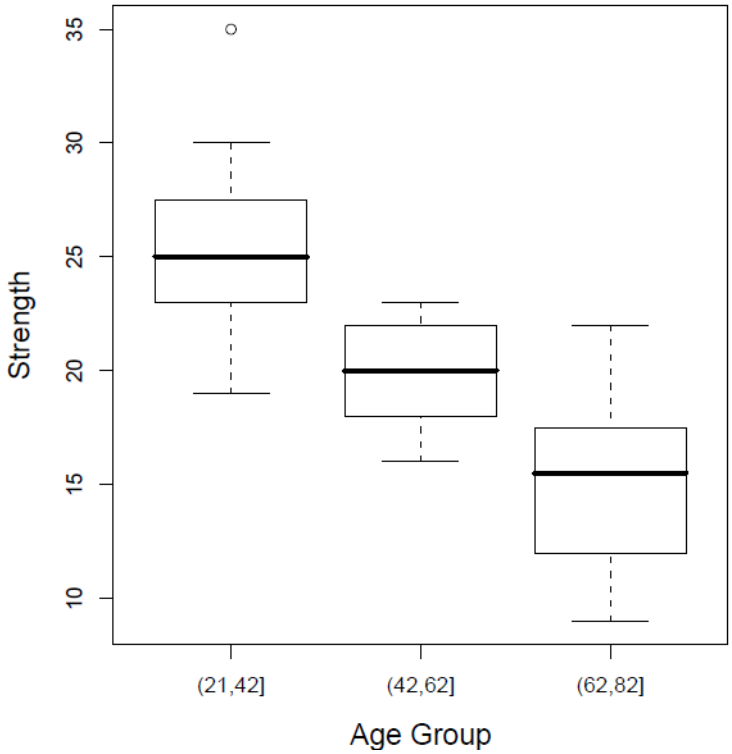

Figure 1.4.1 Side-by-side Boxplot

1.4.2 Scatterplots

For two quantitative variables, the basic graphical EDA technique is the scatterplot, which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. If one variable is explanatory and the other is outcome.

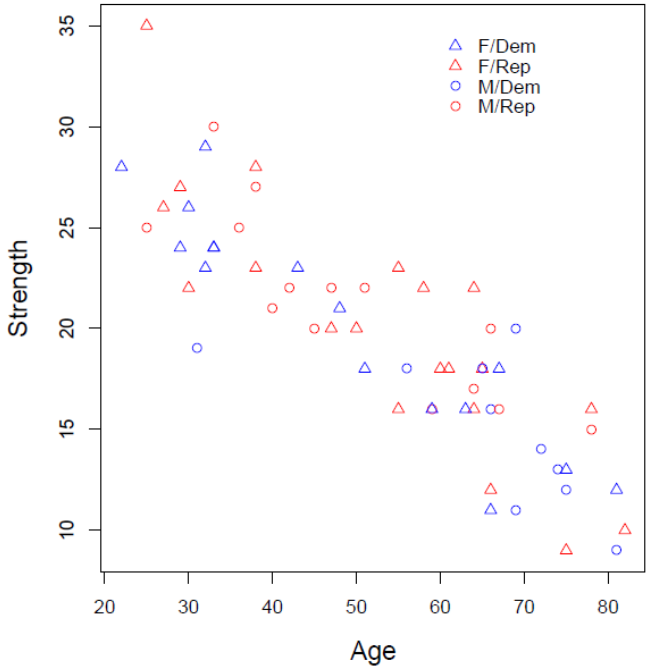

Figure 1.4.2 An example of Age vs. strength is shown, and different colours and symbols are used to code political party and gender.

In a nutshell: You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn

about variable distributions, and learn about relationships between variables. EDA is not an exact science; it is a very important art!

2 The normal distribution

The normal distribution is central to statistical theory. It has been found to describe remarkably well the errors of observation in physics. Many environmental variables, such as of the soil, are distributed in a way that approximates the normal distribution. The form of the distribution was discovered independently by De Moivre, Laplace and Gauss, but Gauss seems generally to take the credit for it, and the distribution is often called Gaussian. It is defined for a continuous random variable X in terms of the probability density function (pdf), f(x), as

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Equation 2.1

Where
μ is the mean of the population
σ is the population standard deviation
σ^2 is the population variance
x is the value of the variable

The normal distribution is continuous and symmetrical, with its peak at the mean of the distribution. The ordinate f(x) at any given value of x is the probability density at x. The total area under the curve is 1, the total probability of the distribution. The area under any portion of the curve, say between z1 and z2, represents the proportion of the distribution lying in that range

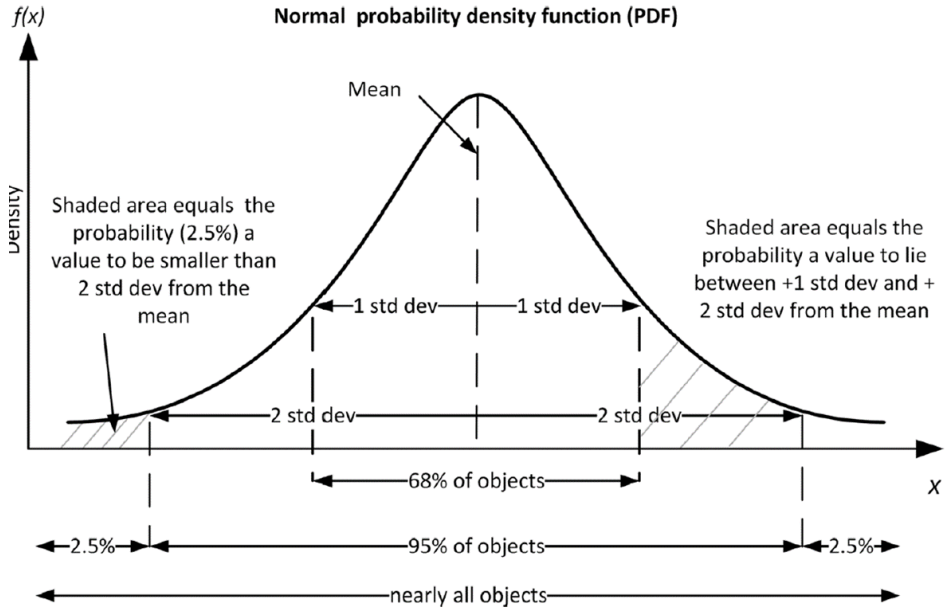

Figure 2.1 Illustration of the distribution of values based on standard deviations in a normal distribution

Slightly more than two-thirds of the distribution lies within one standard deviation of the mean, i.e. between μ-σ AND μ+ σ; About 95% lies in the range μ-2σ AND μ+ 2σ; About 99.73% lies within three standard deviations of the mean μ-3σ AND μ+ 3σ;

3 The Cumulative distribution function

Pdf can be represented as a cumulative distribution. In this representation the normal distribution is characteristically sigmoid. The main use of the cumulative distribution function is that the probability of a value being less than a specified amount can be deduced.

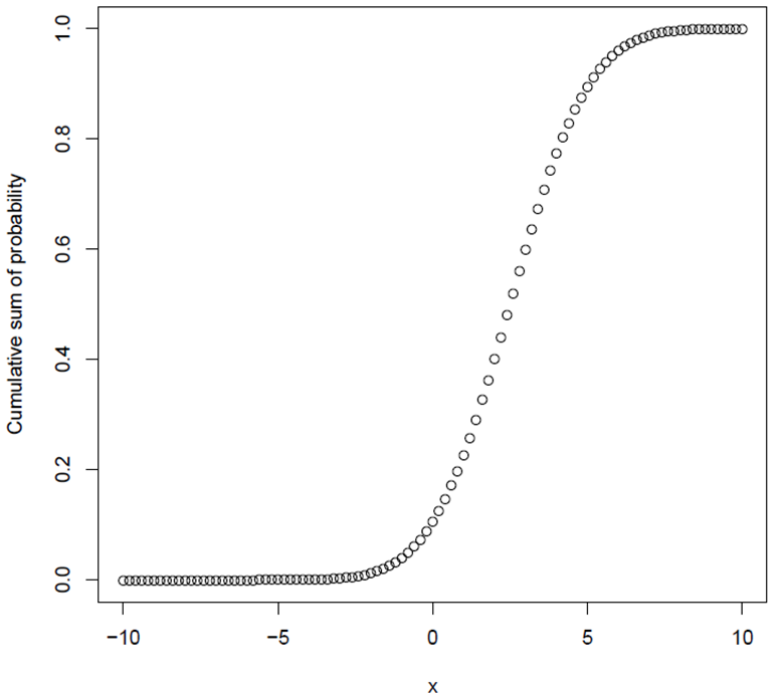

Figure 3.1 Illustration of distribution of values for a cumulative distribution function

4 Data transformation

To overcome the difficulties arising from departures from normality, observations can be transformed to a new scale on which the distribution is more nearly normal. All further analysis can be done on the transformed data, and if necessary, transform the results to the original scale at the end. The following are some of the commonly used transformations for measured data.

4.1 Logarithmic transformation

The geometric mean of a dataset is given as

$$\bar{g} = \left\{ \prod_{i=1}^{N} z_i \right\}^{\frac{1}{N}},$$

Equation 4.1

Therefore, the log transformation is

$$\log \bar{g} = \frac{1}{N} \sum_{i=1}^{N} \log z_i,$$

Equation 4.2

The logarithm may be either natural (ln) or common (log10). If by transforming the data Z_i= 1, 2, 3…….N, we obtain log z with a normal distribution then the variable is said to be lognormally distributed.

4.2 Square root transformation

Taking logarithms will often normalize, or at least make symmetric, distributions that are strongly positively skewed, i.e. have g1 > 1.
Less pronounced positive skewness can be removed by taking square roots:

$$r = \sqrt{z}.$$

Equation 4.3

4.3 Angular transformation

This is sometimes used for proportions in the range 0 to 1, or 0 to 100 if expressed as percentages.
If p is the proportion then define

$$\phi = \sin^{-1} \sqrt{p}.$$

Equation 4.4

The desired transform is the angle whose sine is √p

4.4 Logit transformation

If, as above, p is a proportion (0 <p <1), then its logit is

$$l = \ln \frac{p}{1-p}$$

Equation 4.5

Note that the limits 0 and 1 are excluded; otherwise l would either go to -∞ or +∞.
If you have proportions that include 0 or 1 then you must make some little adjustment to use the logit transformation.