Chapter 4

Exploratory Data Analysis

A first look at the data.

As mentioned in Chapter 1, exploratory data analysis or "EDA" is a critical first step in analyzing the data from an experiment. Here are the main reasons we use EDA:

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis.

4.1 Typical data format and the types of EDA

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 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.

The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.

This chapter first discusses the non-graphical and graphical methods for looking at single variables, then moves on to looking at multiple variables at once, mostly to investigate the relationships between the variables.

4.2 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.

4.2.1 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 if we categorize subjects by College at Carnegie Mellon University as H&SS, MCS, SCS and "other", then there is a true population of all students enrolled in the 2007 Fall semester. If we take a random sample of 20 students for the purposes of performing a memory experiment, we could list the sample "measurements" as H&SS, H&SS, MCS, other, other, SCS, MCS, other, H&SS, MCS, SCS, SCS, other, MCS, MCS, H&SS, MCS, other, H&SS, SCS. Our EDA would look like this:

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%

Note that it is useful to have the total count (frequency) to verify that we have an observation for each subject that we recruited. (Losing data is a common mistake, and EDA is very helpful for finding mistakes.). Also, we should expect that the proportions add up to 1.00 (or 100%) if we are calculating them correctly (count/total). Once you get used to it, you won't need both proportion (relative frequency) and percent, because they will be interchangeable in your mind.

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

4.2.2 Characteristics of quantitative data

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. (See section 3.5.) Our observed data represent just one sample out of an infinite number of possible samples. *The characteristics of our randomly observed sample are not inherently interesting, except to the degree that they represent the population that it came from*.

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 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 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 (see 4.3.1). 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 (see section 3.5) 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 4.1 shows a histogram of a sample of size 200 from the infinite population characterized by distribution **C** of figure 3.1 from section 3.5. Remember that in that section we examined the parameters that characterize theoretical (population) distributions. Now we are interested in learning what we can (but not everything, because parameters are "secrets of nature") about these parameters from measurements on a (random) sample of subjects out of that population.

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). The definition of "outlier" for standard boxplots is described below (see 4.3.3). Another common definition of "outlier" consider any point more than a fixed number of standard deviations from the mean to be an "outlier", but these and other definitions are arbitrary and vary from situation to situation.

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. *But for categorical variables, none of these make any sense.*

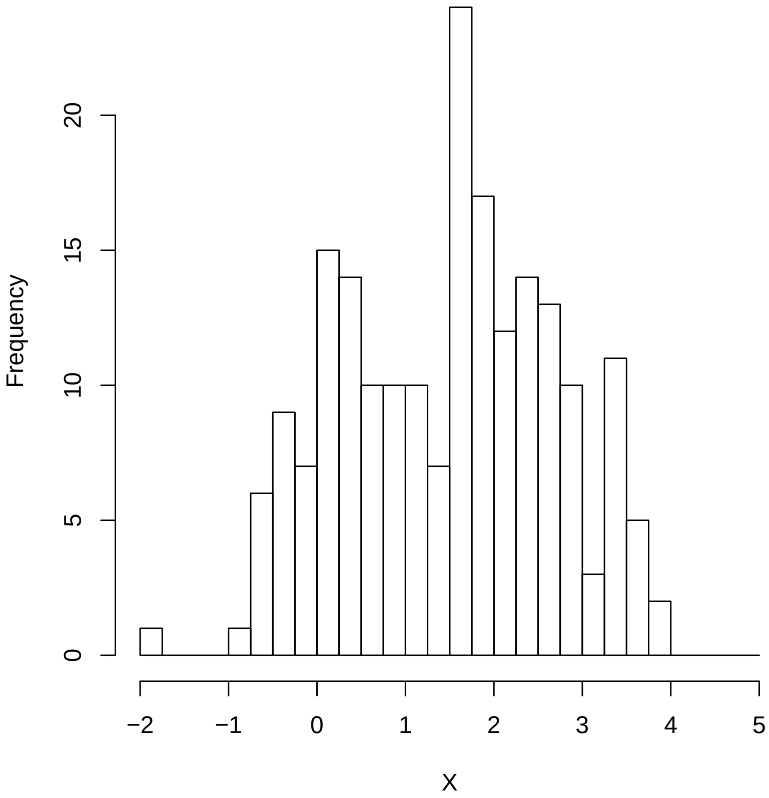

Figure 4.1: Histogram from distribution C.

4.2.3 Central tendency

The **central tendency** or “location” of a distribution has to do with typical or middle values. The common, useful measures of central tendency are the statistics called (arithmetic) mean, median, and sometimes mode. Occasionally other means such as geometric, harmonic, truncated, or Winsorized means are used as measures of centrality. While most authors use the term “average” as a synonym for arithmetic mean, some use average in a broader sense to also include geometric, harmonic, and other means.

Assuming that we have n data values labeled x_1 through x_n, the formula for calculating the sample (arithmetic) **mean** is

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

The arithmetic mean is simply the sum of all of the data values divided by the number of values. It can be thought of as how much each subject gets in a “fair” re-division of whatever the data are measuring. For instance, the mean amount of money that a group of people have is the amount each would get if all of the money were put in one “pot”, and then the money was redistributed to all people evenly. I hope you can see that this is the same as “summing then dividing by n”.

For any symmetrically shaped distribution (i.e., one with a symmetric histogram or pdf or pmf) the mean is the point around which the symmetry holds. For non-symmetric distributions, the mean is the “balance point”: if the histogram is cut out of some homogeneous stiff material such as cardboard, it will balance on a fulcrum placed at the mean.

For many descriptive quantities, there are both a sample and a population version. For a fixed finite population or for a theoretic infinite population described by a pmf or pdf, there is a single population mean which is a fixed, often unknown, value called the mean **parameter** (see section 3.5). On the other hand, the “sample mean” will vary from sample to sample as different samples are taken, and so is a random variable. The probability distribution of the sample mean is referred to as its **sampling distribution**. This term expresses the idea that any experiment could (at least theoretically, given enough resources) be repeated many times and various statistics such as the sample mean can be calculated each time. Often we can use probability theory to work out the exact distribution of the sample statistic, at least under certain assumptions.

The **median** is another measure of central tendency. The sample median is the middle value after all of the values are put in an ordered list. If there are an even number of values, take the average of the two middle values. (If there are ties at the middle, some special adjustments are made by the statistical software we will use. In unusual situations for discrete random variables, there may not be a unique median.)

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. For example, according to the US Census Bureau 2004 Economic Survey, the median income of US families, which represents the income above and below which half of families fall, was $43,318. This seems a better measure of central tendency than the mean of $60,828, which indicates how much each family would have if we all shared equally. And the difference between these two numbers is quite substantial. Nevertheless, both numbers are "correct", as long as you understand their meanings.

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, 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.

The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median may be preferred.

4.2.4 Spread

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 **variance** is a standard measure of spread. It is calculated for a list of numbers, e.g., the n observations of a particular measurement labeled x_1 through x_n, based on the n **sample deviations** (or just "deviations"). Then for any data value, x_i, the corresponding deviation is $(x_i - \bar{x})$, which is the signed (- for lower and + for higher) distance of the data value from the mean of all of the n data values. It is not hard to prove that the sum of all of the deviations of a sample is zero.

The variance of a population is defined as the mean squared deviation (see section 3.5.2). The sample formula for the variance of observed data conventionally has $n-1$ in the denominator instead of n to achieve the property of "unbiasedness", which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here, σ^2). The most commonly used symbol for sample variance is s^2, and the formula is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

which is essentially the average of the squared deviations, except for dividing by $n-1$ instead of n. This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets. (In most cases, squaring is better than taking the absolute value because it puts special emphasis on highly deviant values.) As usual, a sample statistic like s^2 is best thought of as a characteristic of a particular sample (thus varying from sample to sample) which is used as an estimate of the single, fixed, true corresponding parameter value from the population, namely σ^2.

Another (equivalent) way to write the variance formula, which is particularly useful for thinking about ANOVA is

$$s^2 = \frac{\text{SS}}{\text{df}}$$

where SS is "sum of squared deviations", often loosely called "sum of squares", and df is "degrees of freedom" (see section 4.6).

Because of the square, variances are always non-negative, and they have the somewhat unusual property of having squared units compared to the original data. So if the random variable of interest is a temperature in degrees, the variance has units “degrees squared”, and if the variable is area in square kilometers, the variance is in units of “kilometers to the fourth power”.

Variances have the very important property that they are additive for any number of different independent sources of variation. For example, the variance of a measurement which has subject-to-subject variability, environmental variability, and quality-of-measurement variability is equal to the sum of the three variances. This property is not shared by the “standard deviation”.

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 a theoretical Gaussian distribution, we learned in the previous chapter that mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7% of the probability respectively, and this should be approximately true for real data from a Normal distribution.

The variance and standard deviation are two useful measures of spread. The variance is the mean of the squares of the individual deviations. The standard deviation is the square root of the variance. For Normally distributed data, approximately 95% of the values lie within 2 sd of the mean.

A third measure of spread is the **interquartile range**. To define IQR, we first need to define the concepts of **quartiles**. 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. The range of a sample is the distance from the minimum value to the maximum value: range = maximum - minimum. 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.

The interquartile range (IQR) is a robust measure of spread.

4.2.5 Skewness and kurtosis

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. Sample estimates of skewness and kurtosis are taken as estimates of the corresponding population parameters (see section 3.5.3). If the sample skewness and kurtosis are calculated along with their standard errors, we can roughly make conclusions according to the following table where e is an estimate of skewness and u is an estimate of kurtosis, and $SE(e)$ and $SE(u)$ are the corresponding standard errors.

Skewness (e) or kurtosis (u)	**Conclusion**
$-2\text{SE(e)} < e < 2\text{SE(e)}$	not skewed
$e \leq -2\text{SE(e)}$	negative skew
$e \geq 2\text{SE(e)}$	positive skew
$-2\text{SE(u)} < u < 2\text{SE(u)}$	not kurtotic
$u \leq -2\text{SE(u)}$	negative kurtosis
$u \geq 2\text{SE(u)}$	positive kurtosis

For a positive skew, values far above the mode are more common than values far below, and the reverse is true for a negative skew. When a sample (or distribution) has positive kurtosis, then compared to a Gaussian distribution with the same variance or standard deviation, values far from the mean (or median or mode) are more likely, and the shape of the histogram is peaked in the middle, but with fatter tails. For a negative kurtosis, the peak is sometimes described has having “broader shoulders” than a Gaussian shape, and the tails are thinner, so that extreme values are less likely.

Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peakedness compared to a Gaussian distribution.

4.3 Univariate graphical EDA

If we are focusing on data from 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; therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also required.

4.3.1 Histograms

The only one of these techniques that makes sense for categorical data is the histogram (basically just a barplot of the tabulation of the data). A pie chart is equivalent, but not often used. The concepts of central tendency, spread and skew have no meaning for nominal categorical data. For ordinal categorical data, it sometimes makes sense to treat the data as quantitative for EDA purposes; you need to use your judgment here.

The most basic graph is the **histogram**, which is a barplot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Typically the bars run vertically with the count (or proportion) axis running vertically. To manually construct a histogram, define the range of data for each bar (called a **bin**), count how many cases fall in each bin, and draw the bars high enough to indicate the count. For the simple data set found in EDA1.dat the histogram is shown in figure 4.2. Besides getting the general impression of the shape of the distribution, you can read off facts like "there are two cases with data values between 1 and 2" and "there are 9 cases with data values between 2 and 3". Generally values that fall exactly on the boundary between two bins are put in the lower bin, but this rule is not always followed.

Generally you will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution. Of course you need to see the histogram to know the shape of the distribution, so this may be an iterative process. It is often worthwhile to try a few different bin sizes/numbers because, especially with small samples, there may sometimes be a different shape to the histogram when the bin size changes. But usually the difference is small. Figure 4.3 shows three histograms of the same sample from a bimodal population using three different bin widths (5, 2 and 1). If you want to try on your own, the data are in EDA2.dat. The top panel appears to show a unimodal distribution. The middle panel correctly shows the bimodality. The bottom panel incorrectly suggests many modes. There is some art to choosing bin widths, and although often the automatic choices of a program like SPSS are pretty good, they are certainly not always adequate.

It is very instructive to look at multiple samples from the same population to get a feel for the variation that will be found in histograms. Figure 4.4 shows histograms from multiple samples of size 50 from the same population as figure 4.3, while 4.5 shows samples of size 100. Notice that the variability is quite high, especially for the smaller sample size, and that an incorrect impression (particularly of unimodality) is quite possible, just by the bad luck of taking a particular sample.

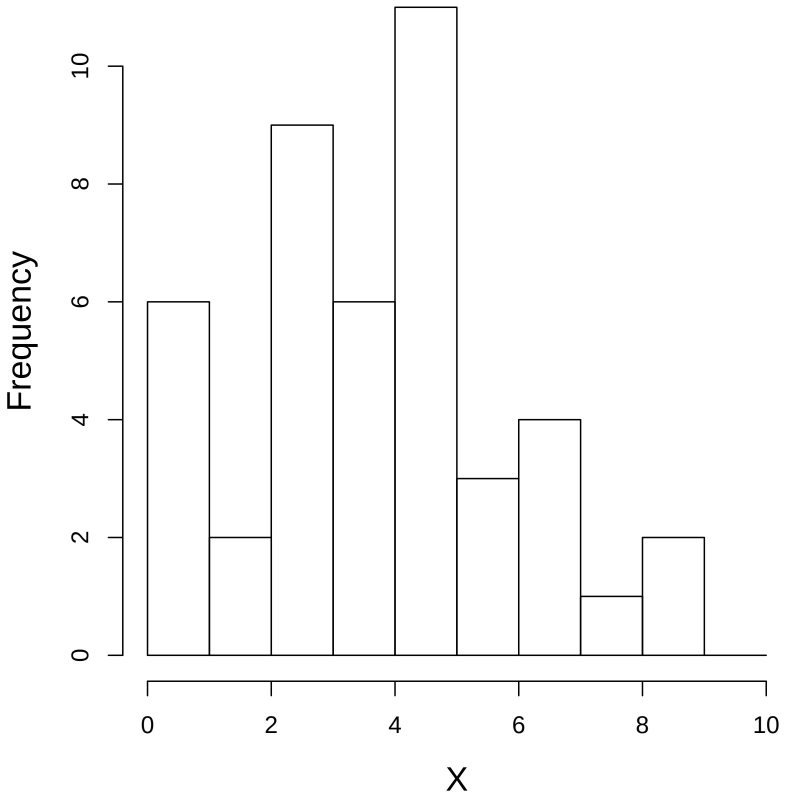

Figure 4.2: Histogram of EDA1.dat.

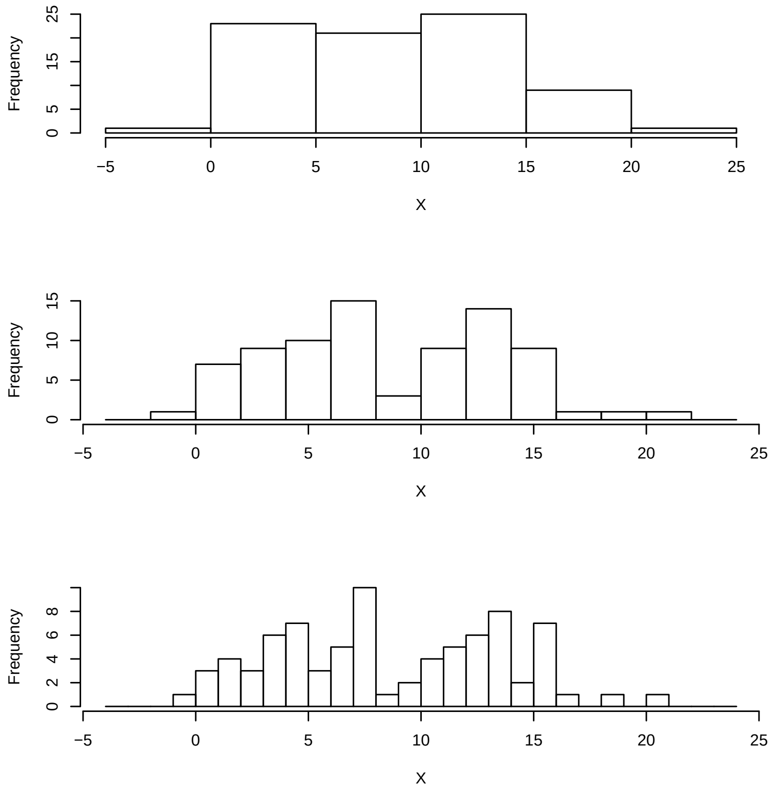

Figure 4.3: Histograms of EDA2.dat with different bin widths.

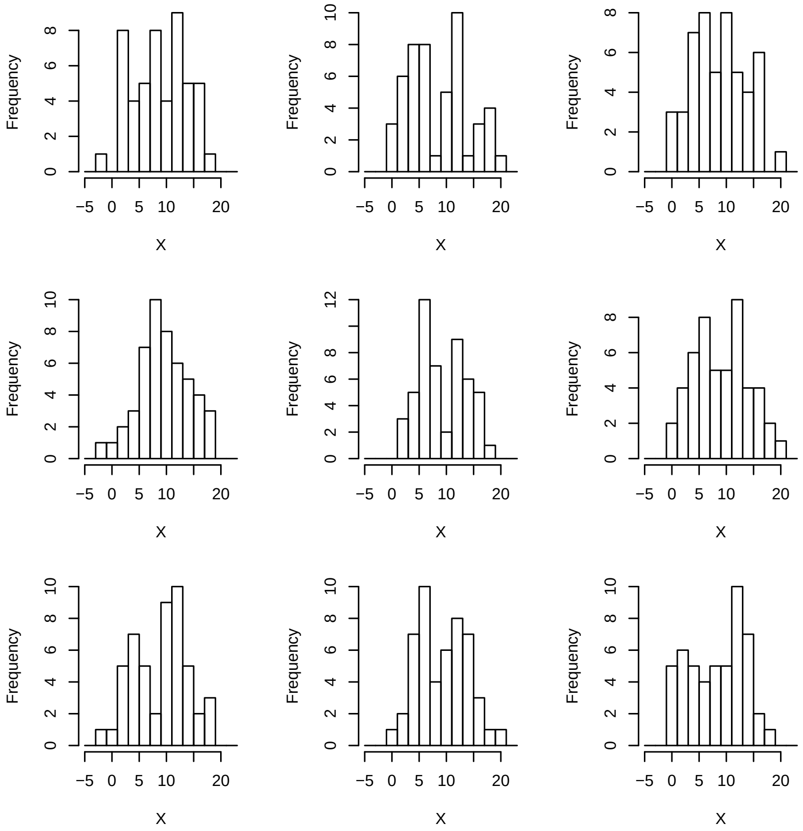

Figure 4.4: Histograms of multiple samples of size 50.

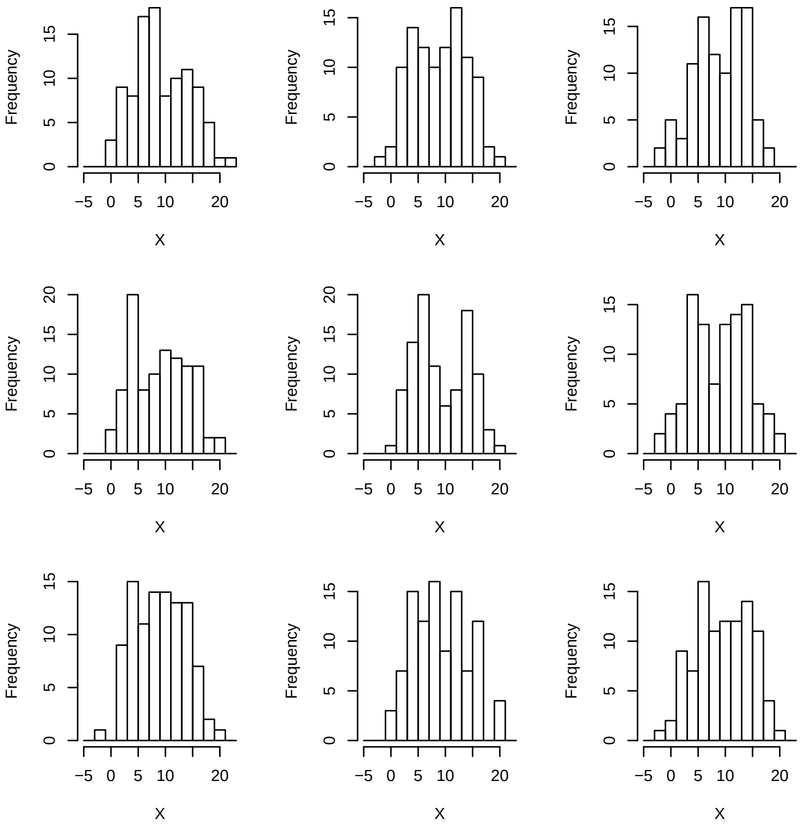

Figure 4.5: Histograms of multiple samples of size 100.

With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

4.3.2 Stem-and-leaf plots

A simple substitute for a histogram is a **stem and leaf plot**. A stem and leaf plot is sometimes easier to make by hand than a histogram, and it tends not to hide any information. Nevertheless, a histogram is generally considered better for appreciating the shape of a sample distribution than is the stem and leaf plot. Here is a stem and leaf plot for the data of figure 4.2:

```
The decimal place is at the "|".
1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

Because this particular stem and leaf plot has the decimal place at the stem, each of the 0's in the first line represent 1.0, and each zero in the second line represents 2.0, etc. So we can see that there are six 1's, two 2's etc. in our data.

A stem and leaf plot shows all data values and the shape of the distribution.

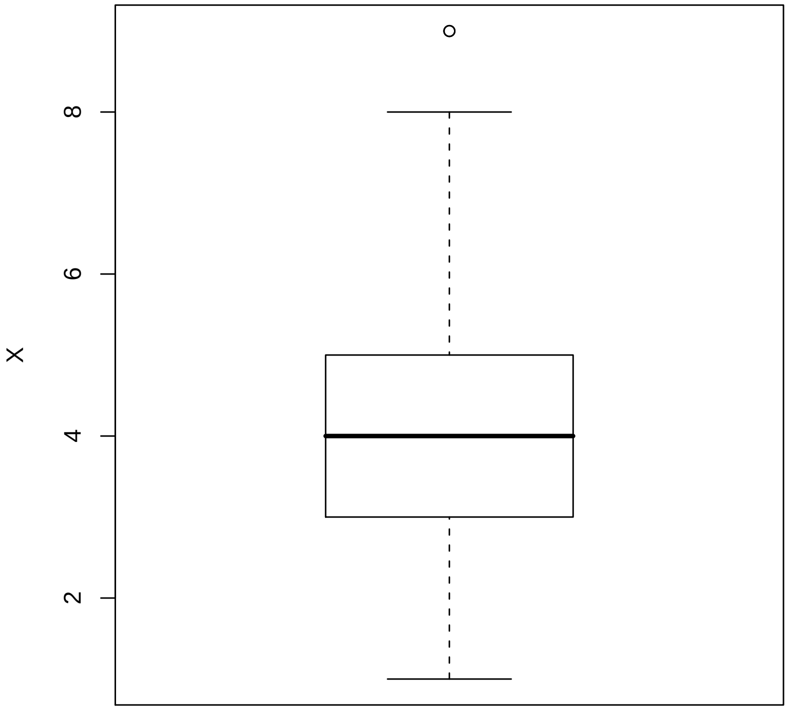

Figure 4.6: A boxplot of the data from EDA1.dat.

4.3.3 Boxplots

Another very useful univariate graphical technique is the **boxplot**. The boxplot will be described here in its vertical format, which is the most common, but a horizontal format also is possible. An example of a boxplot is shown in figure 4.6, which again represents the data in EDA1.dat.

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. One of the best uses of boxplots is in the form of side-by-side boxplots (see multivariate graphical analysis below).

Figure 4.7 is an annotated version of figure 4.6. Here you can see that the boxplot consists of a rectangular box bounded above and below by “hinges” that represent the quartiles Q3 and Q1 respectively, and with a horizontal “median”

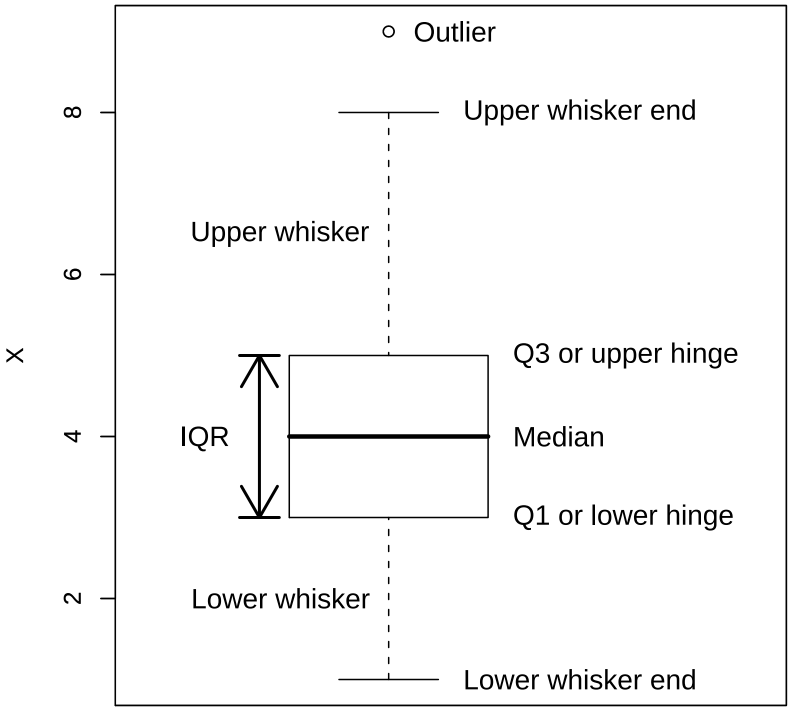

Figure 4.7: Annotated boxplot.

line through it. You can also see the upper and lower "whiskers", and a point marking an "outlier". The vertical axis is in the units of the quantitative variable.

Let's assume that the subjects for this experiment are hens and the data represent the number of eggs that each hen laid during the experiment. We can read certain information directly off of the graph. The median (**not mean!**) is 4 eggs, so no more than half of the hens laid more than 4 eggs and no more than half of the hens laid less than 4 eggs. (This is based on the technical definition of median; we would usually claim that half of the hens lay more or half less than 4, knowing that this may be only approximately correct.) We can also state that one quarter of the hens lay less than 3 eggs and one quarter lay more than 5 eggs (again, this may not be exactly correct, particularly for small samples or a small number of different possible values). This leaves half of the hens, called the "central half", to lay between 3 and 5 eggs, so the interquartile range (IQR) is Q3-Q1=5-3=2.

The interpretation of the whiskers and outliers is just a bit more complicated. Any data value more than 1.5 IQRs beyond its corresponding hinge in either direction is considered an "outlier" and is individually plotted. Sometimes values beyond 3.0 IQRs are considered "extreme outliers" and are plotted with a different symbol. In this boxplot, a single outlier is plotted corresponding to 9 eggs laid, although we know from figure 4.2 that there are actually two hens that laid 9 eggs. This demonstrates a general problem with plotting whole number data, namely that multiple points may be superimposed, giving a wrong impression. (Jittering, circle plots, and starplots are examples of ways to correct this problem.) This is one reason why, e.g., combining a tabulation and/or a histogram with a boxplot is better than either alone.

Each whisker is drawn out to the most extreme data point that is less than 1.5 IQRs beyond the corresponding hinge. Therefore, the whisker ends correspond to the minimum and maximum values of the data *excluding* the "outliers".

Important: The term "outlier" is not well defined in statistics, and the definition varies depending on the purpose and situation. The "outliers" identified by a boxplot, which could be called "boxplot outliers" are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This *does not* by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. It is also important to realize that the number of boxplot outliers depends strongly on the size of the sample. In fact, for data that is perfectly Normally distributed, we expect 0.70 percent (or about 1 in 150 cases) to be "boxplot outliers", with approximately half in either direction.

The boxplot information described above could be appreciated almost as easily if given in non-graphical format. The boxplot is useful because, with practice, all of the above and more can be appreciated at a quick glance. The additional things you should notice on the plot are the symmetry of the distribution and possible evidence of "fat tails". Symmetry is appreciated by noticing if the median is in the center of the box and if the whiskers are the same length as each other. For this purpose, as usual, the smaller the dataset the more variability you will see from sample to sample, particularly for the whiskers. In a skewed distribution we expect to see the median pushed in the direction of the shorter whisker. If the longer whisker is the top one, then the distribution is positively skewed (or skewed to the right, because higher values are on the right in a histogram). If the lower whisker is longer, the distribution is negatively skewed (or left skewed.) In cases where the median is closer to the longer whisker it is hard to draw a conclusion.

The term **fat tails** is used to describe the situation where a histogram has a lot of values far from the mean relative to a Gaussian distribution. This corresponds to positive kurtosis. In a boxplot, many outliers (more than the 1/150 expected for a Normal distribution) suggests fat tails (positive kurtosis), or possibly many data entry errors. Also, short whiskers suggest negative kurtosis, at least if the sample size is large.

Boxplots are excellent EDA plots because they rely on robust statistics like median and IQR rather than more sensitive ones such as mean and standard deviation. With boxplots it is easy to compare distributions (usually for one variable at different levels of another; see multivariate graphical EDA, below) with a high degree of reliability because of the use of these robust statistics.

It is worth noting that some (few) programs produce boxplots that do not conform to the definitions given here.

Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

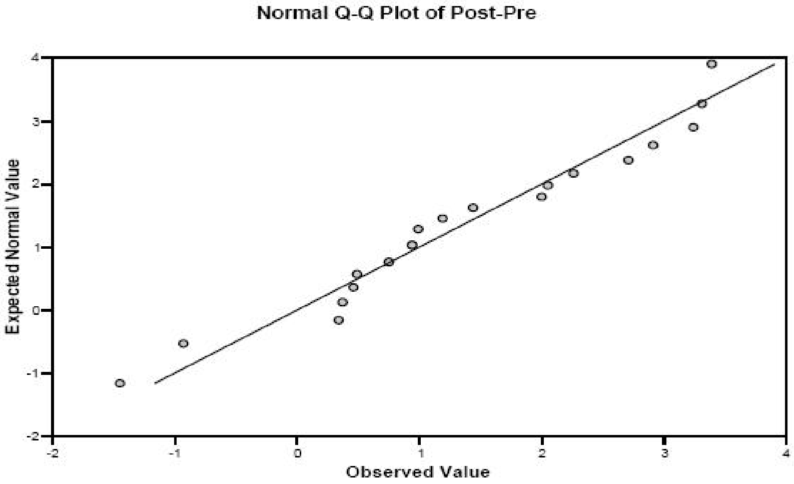

Figure 4.8: A quantile-normal plot.

4.3.4 Quantile-normal plots

The final univariate graphical EDA technique is the most complicated. 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 4.8 shows 20 data points that are approximately normally distributed. **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. For many computer programs, the word "quantile" is also in the axis labels.

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. As we will see in the ANOVA and regression chapters, the most important situation where we use a QN plot is not for EDA, but for examining something called "residuals" (see section 9.4). For basic interpretation of the QN plot you just need to be able to distinguish the two situations of “OK” (points fall randomly around the line) versus “non-normality” (points follow a strong curved pattern rather than following the line).

If you are still curious, here is a description of how the QN plot is created. Understanding this will help to understand the interpretation, but is not required in this course. Note that some programs swap the x and y axes from the way described here, but the interpretation is similar for all versions of QN plots. Consider the 20 values observed in this study. They happen to have an observed mean of 1.37 and a standard deviation of 1.36. Ideally, 20 random values drawn from a distribution that has a true mean of 1.37 and sd of 1.36 have a perfect bell-shaped distribution and will be spaced so that there is equal area (probability) in the area around each value in the bell curve.

In figure 4.9 the dotted lines divide the bell curve up into 20 equally probable zones, and the 20 points are at the probability mid-points of each zone. These 20 points, which are more tightly packed near the middle than in the ends, are used as the “Expected Normal Values” in the QN plot of our actual data.

In summary, the sorted actual data values are plotted against “Expected Normal Values”, and some kind of diagonal line is added to help direct the eye towards a perfect straight line on the quantile-normal plot that represents a perfect bell shape for the observed data.

The interpretation of the QN plot is given here. If the axes are reversed in the computer package you are using, you will need to correspondingly change your interpretation. If all of the points fall on or nearly on the diagonal line (with a random pattern), this tells us that a histogram of the variable will show a bell shaped (Normal or Gaussian) distribution.

Figure 4.10 shows all of the points basically on the reference line, but there are several vertical bands of points. Because the x-axis is “observed values”, these bands indicate ties, i.e., multiple points with the same values. And all of the observed values are at whole numbers. So either the data are rounded or we are looking at a discrete quantitative (counting) variable. Either way, the data appear

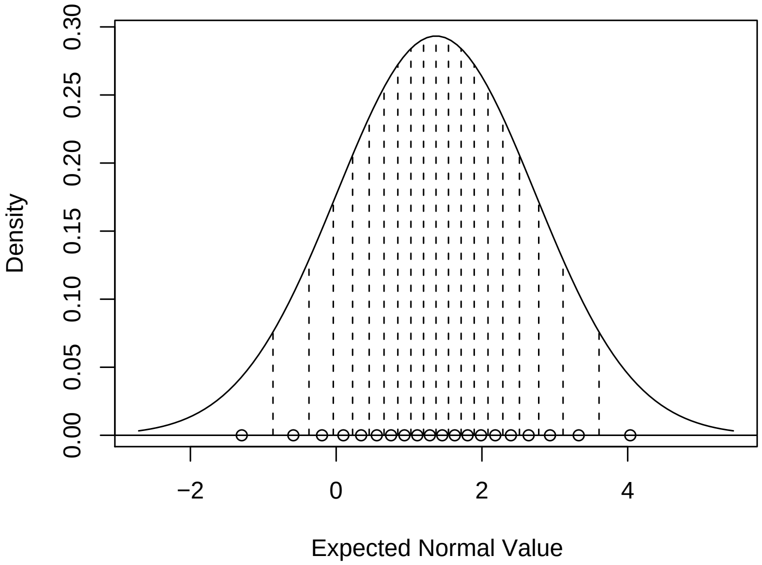

Figure 4.9: A way to think about QN plots.

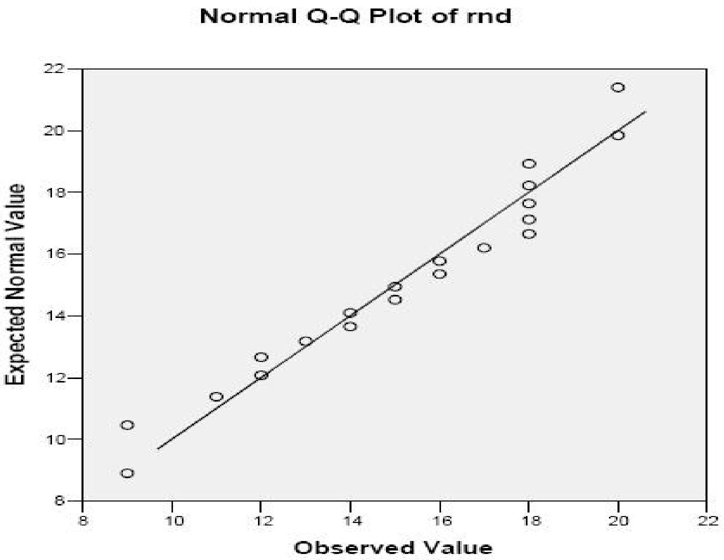

Figure 4.10: Quantile-normal plot with ties.

to be nearly normally distributed.

In figure 4.11 note that we have many points in a row that are on the same side of the line (rather than just bouncing around to either side), and that suggests that there is a real (non-random) deviation from Normality. The best way to think about these QN plots is to look at the low and high ranges of the Expected Normal Values. In each area, see how the observed values deviate from what is expected, i.e., in which "x" (Observed Value) direction the points appear to have moved relative to the "perfect normal" line. Here we observe values that are too high in both the low and high ranges. So compared to a perfect bell shape, this distribution is pulled asymmetrically towards higher values, which indicates positive skew.

Also note that if you just *shift* a distribution to the right (without disturbing its symmetry) rather than skewing it, it will maintain its perfect bell shape, and the points remain on the diagonal reference line of the quantile-normal curve.

Of course, we can also have a distribution that is skewed to the left, in which case the high and low range points are shifted (in the Observed Value direction) towards lower than expected values.

In figure 4.12 the high end points are shifted too high and the low end points are shifted too low. These data show a positive kurtosis (fat tails). The opposite pattern is a negative kurtosis in which the tails are too "thin" to be bell shaped.

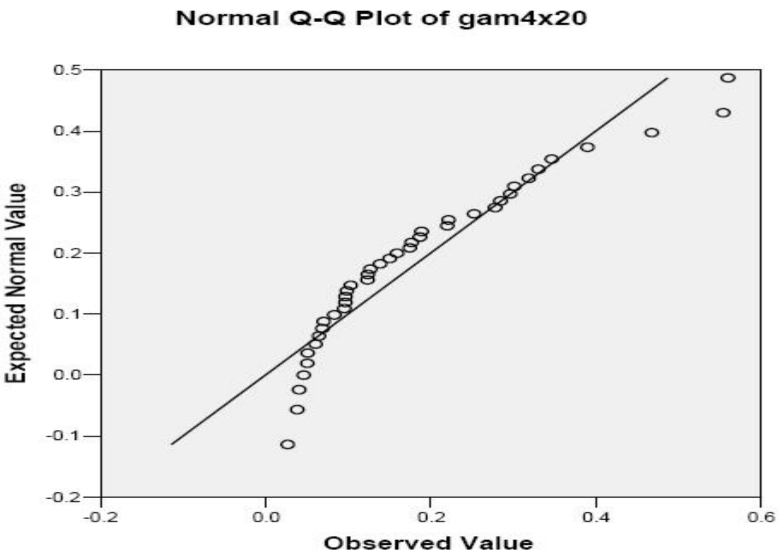

Figure 4.11: Quantile-normal plot showing right skew.

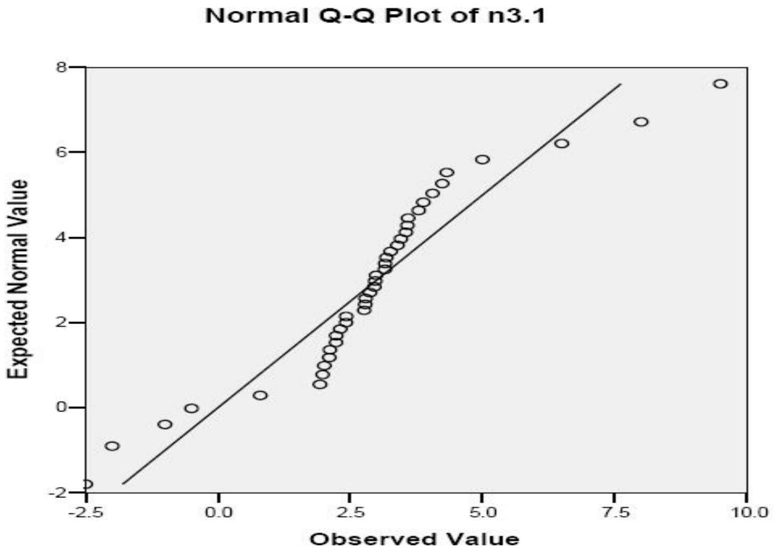

Figure 4.12: Quantile-normal plot showing fat tails.

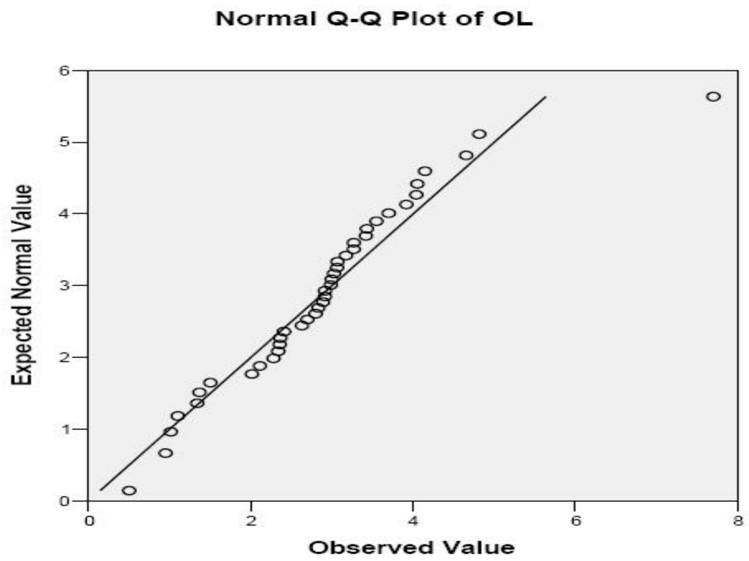

Figure 4.13: Quantile-normal plot showing a high outlier.

In figure 4.13 there is a single point that is off the reference line, i.e. shifted to the right of where it should be. (Remember that the pattern of locations on the Expected Normal Value axis is fixed for any sample size, and only the position on the Observed axis varies depending on the observed data.) This pattern shows nearly Gaussian data with one "high outlier".

Finally, figure 4.14 looks a bit similar to the "skew left" pattern, but the most extreme points tend to return to the reference line. This pattern is seen in bi-modal data, e.g. this is what we would see if we would mix strength measurements from controls and muscular dystrophy patients.

Quantile-Normal plots allow detection of non-normality and diagnosis of skewness and kurtosis.

4.4 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.

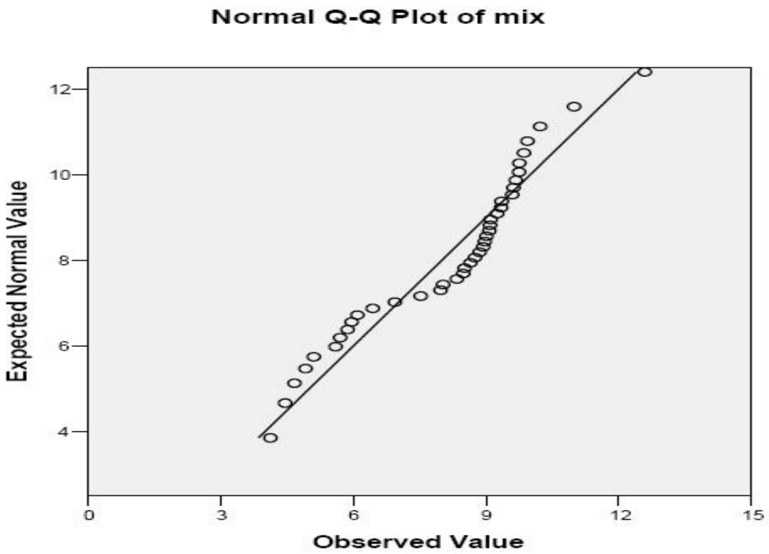

Figure 4.14: Quantile-normal plot showing bimodality.

4.4.1 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, or 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 add to 100% over all cells) are also useful.

Here is an example of a cross-tabulation. Consider the data in table 4.1. For each subject we observe sex and age as categorical variables.

Table 4.2 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 $2/11 \times 100 = 18.2\%$, the row percentage is $2/5 \times 100 = 40.0\%$, and the column percentage is $2/7 \times 100 = 28.6\%$.

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.

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 4.1: Sample Data for Cross-tabulation

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

Table 4.2: Cross-tabulation of Sample Data

For example, we could make separate age by gender tables for each education level.

Cross-tabulation is the basic bivariate non-graphical EDA technique.

4.4.2 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, and that material is currently beyond the scope of this book.

4.4.3 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.

Especially for a categorical explanatory variable and a quantitative outcome variable, it is useful to produce a variety of univariate statistics for the quantitative variable at each level of the categorical variable.

4.4.4 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 from section 3.5. 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.

Sample covariance is calculated by computing (signed) deviations of each measurement from the average of all measurements for that variable. Then the deviations for the two measurements are multiplied together separately for each subject. Finally these values are averaged (actually summed and divided by n-1, to keep the statistic unbiased). Note that the units on sample covariance are the products of the units of the two variables.

Positive covariance values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa. Negative covariances suggest that when one variable is above its mean, the other is below its mean. And covariances near zero suggest that the two variables vary independently of each other.

Technically, independence implies zero correlation, but the reverse is not necessarily true.

Covariances tend to be hard to interpret, so we often use correlation instead. The correlation has the nice property that it is always between -1 and +1, with -1 being a “perfect” negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated. The symbol r or $r_{x,y}$ is often used for sample correlations.

The general formula for sample covariance is

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

It is worth noting that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

If you want to see a “manual example” of calculation of sample covariance and correlation consider an example using the data in table 4.3. For each subject we observe age and a strength measure.

Table 4.4 shows the calculation of covariance. The mean age is 50 and the mean strength is 19, so we calculate the deviation for age as age-50 and deviation for strength and strength-19. Then we find the product of the deviations and add them up. This total is 1106, and since n=11, the covariance of x and y is -1106/10=-110.6. The fact that the covariance is negative indicates that as age goes up strength tends to go down (and vice versa).

The formula for the sample correlation is

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y.

In this example, $s_x = 18.96$, $s_y = 6.39$, so $r = \frac{-110.6}{18.96 \cdot 6.39} = -0.913$. This is a strong negative correlation.

Subject ID	Age	Strength
GW	38	20
JA	62	15
TJ	22	30
JMA	38	21
JMO	45	18
JQA	69	12
AJ	75	14
MVB	38	28
WHH	80	9
JT	32	22
JKP	51	20

Table 4.3: Covariance Sample Data

4.4.5 Covariance and correlation matrices

When we have many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix. Note that the covariance of X with X is the variance of X and the correlation of X with X is 1.0. For example the covariance matrix of table 4.5 tells us that the variances of X, Y, and Z are 5, 7, and 4 respectively, the covariance of X and Y is 1.77, the covariance of X and Z is -2.24, and the covariance of Y and Z is 3.17.

Similarly the correlation matrix in figure 4.6 tells us that the correlation of X and Y is 0.3, the correlation of X and Z is -0.5. and the correlation of Y and Z is 0.6.

Subject ID	Age	Strength	Age-50	Str-19	product
GW	38	20	-12	+1	-12
JA	62	15	+12	-4	-48
TJ	22	30	-28	+11	-308
JMA	38	21	-12	+2	-24
JMO	45	18	-5	-1	+5
JQA	69	12	+19	-7	-133
AJ	75	14	+25	-5	-125
MVB	38	28	-12	+9	-108
WHH	80	9	+30	-10	-300
JT	32	22	-18	+3	-54
JKP	51	20	+1	+1	+1
Total			0	0	-1106

Table 4.4: Covariance Calculation

	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

Table 4.5: A Covariance Matrix

The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

4.5 Multivariate graphical EDA

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

	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

Table 4.6: A Correlation Matrix

4.5.1 Univariate graphs by category

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, 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**, as in figure 4.15. Here we see the data from EDA3.dat, which consists of strength data for each of three age groups. You can see the downward trend in the median as the ages increase. The spreads (IQRs) are similar for the three groups. And all three groups are roughly symmetrical with one high strength outlier in the youngest age group.

Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.

4.5.2 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, it is a very, very strong convention to put the outcome on the y (vertical) axis.*

One or two additional categorical variables can be accommodated on the scatterplot by encoding the additional information in the symbol type and/or color.

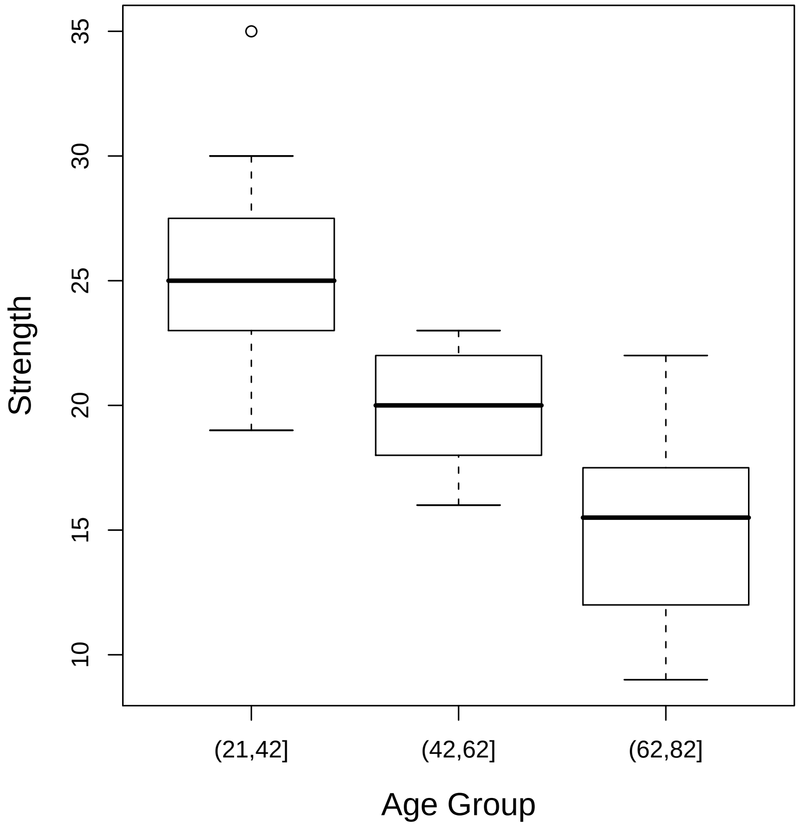

Figure 4.15: Side-by-side Boxplot of EDA3.dat.

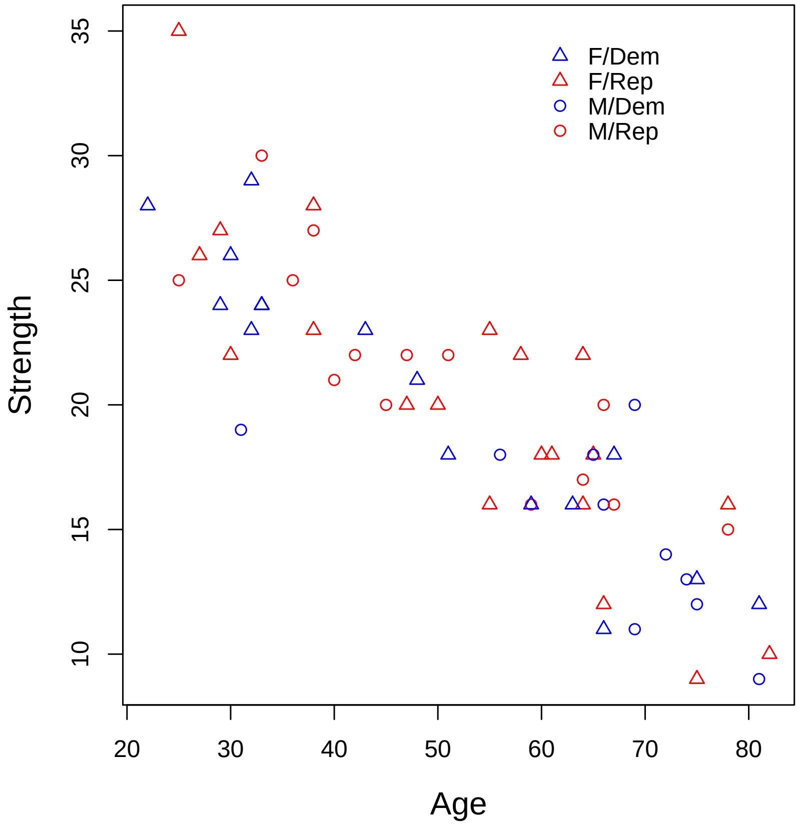

Figure 4.16: scatterplot with two additional variables.

An example is shown in figure 4.16. Age vs. strength is shown, and different colors 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!

4.6 A note on degrees of freedom

Degrees of freedom are numbers that characterize specific distributions in a family of distributions. Often we find that a certain family of distributions is needed in a some general situation, and then we need to calculate the degrees of freedom to know which specific distribution within the family is appropriate.

The most common situation is when we have a particular statistic and want to know its sampling distribution. If the sampling distribution falls in the "t" family as when performing a t-test, or in the "F" family when performing an ANOVA, or in several other families, we need to find the number of degrees of freedom to figure out which particular member of the family actually represents the desired sampling distribution. One way to think about degrees of freedom for a statistic is that they represent the number of independent pieces of information that go into the calculation of the statistic,

Consider 5 numbers with a mean of 10. To calculate the variance of these numbers we need to sum the squared deviations (from the mean). It really doesn't matter whether the mean is 10 or any other number: as long as all five deviations are the same, the variance will be the same. This make sense because variance is a pure measure of spread, not affected by central tendency. But by mathematically rearranging the definition of mean, it is not too hard to show that the sum of the deviations (not squared) is always zero. Therefore, the first four deviations can (freely) be any numbers, but then the last one is forced to be the number that makes the deviations add to zero, and we are not free to choose it. It is in this sense that five numbers used for calculating a variance or standard deviation have only four degrees of freedom (or independent useful pieces of information). In general, a variance or standard deviation calculated from n data values and one mean has $n - 1$ df.

Another example is the "pooled" variance from k independent groups. If the sizes of the groups are n_1 through n_k, then each of the k individual variance estimates is based on deviations from a different mean, and each has one less degree of freedom than its sample size, e.g., $n_i - 1$ for group i. We also say that each numerator of a variance estimate, e.g., SS_i, has $n_i - 1$ df. The pooled estimate of variance is

$$s^2_{\text{pooled}} = \frac{\text{SS}_1 + \cdots + \text{SS}_k}{\text{df}_1 + \cdots + \text{df}_k}$$

and we say that both the numerator SS and the entire pooled variance has $\text{df}_1 + \cdots +$

df_k degrees of freedom, which suggests how many independent pieces of information are available for the calculation.