5

Elementary Plotting Techniques

Plots, charts, and graphs are often used interchangeably.

Plotting data is one of the oldest forms of visualization. In fact, many of the standard plotting techniques were introduced in the late 18th century by William Playfair [Playfair 86, Playfair 01], a pioneer in information visualization. Even today, plotting is by far the most prevalant method for analyzing, correlating, condensing, and presenting scientific data. This is because, with a properly created plot, our visual system is easily able to distinguish patterns that may lead to insight about the underlying data. Conversely, with a bad plot, it is easy to confuse or even deceive the observer about the underlying data. Learning good plotting techniques should not be underestimated because of its importance in the scientific community for publishing and presenting results of hypotheses and experiments. Yet, the subject is often entirely left out of the curriculum for most college students in scientific disciplines!

It is important to note that the goals for plotting in a scientific setting are not the same as they are for those used in general media settings, such as newspapers and magazines. A more advanced knowledge base can be assumed about the scientific reader—less emphasis can be placed on extraneous or superfluous information and more emphasis can be placed on the data itself. The techniques described in this chapter are directed at the scientific community, though many of the principles apply in a more general setting.

There are two basic purposes for plots: data analysis and data communication. As readers and observers of publications and presentation, we are generally more familiar with the latter. However, the former may be of greater importance during the research phase where hypotheses are formed and tested. In either case, the process of creating a useful plot is more iterative than direct. The tack of performing experiments and gathering data can be time consuming, do not expect the analysis to be any different.

In a simplistic view, plotting is just reducing a large amount of information to a smaller form that is more easily understood. There is often a misconception that plotting is a way of presenting the data itself, taking the place of a table or list of the actual values. To the contrary, plotting should be used for displaying relationships within the data. Understanding the information that is being displayed

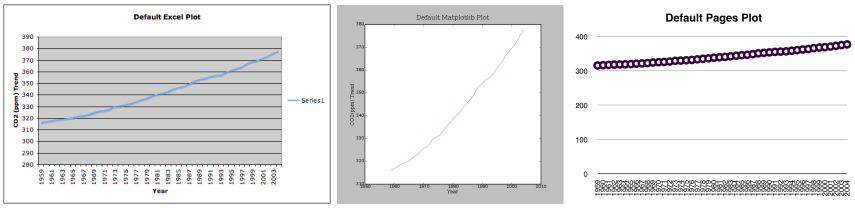


Figure 5.1: Default plot settings for several Excel, Matplotlib, and Pages.

often requires correlation and the detection of trends in otherwise independent samples. To this end, many of the principles and techniques described in this chapter target the reduction of the data to its simplest and cleanest form, such that the relationships inherent in the data are easily perceived.

In this chapter, we begin by describing some basic principles for creating and improving plots(§5.1). We then move on to discuss some of the basic plotting techniques that are commonly used within the scientific community and briefly touch on others that are not(§5.2).

5.1 Principles of Plotting

Because plotting is one of the most common forms of data visualization, there are many software packages available to assist in the creation process. Figure 5.1 shows three default plots generated using three such packages. The data set expresses the yearly average of carbon dioxide measurements at the Mauna Loa Observatory in Hawaii over a fourty six year period [Keeling and Whorf 05]. These plots demonstrate two important points. First, there is no obvious standard for what a plot should look like. This is easy to see by the differences in the axes and scale lines, the data rectangle inside the plot, and the actual representation of the data values. Second, creating a plot is an iterative process that can not be generally applied to all types of data. With all of these software packages, the properties of the plot require manipulation to result in a visually pleasing, and ultimately useful, plot.

So what should a plot look like? Because of the diversity of data and analysis goals, there are no magic formulas for creating a useful plot. However, some general principles have been advocated that can be applied to plots to improve their likelihood of being useful. In *Visualizing Data* [Cleveland 93] and *Elements of Graphing Data* [Cleveland 94] William S. Cleveland enumerates some of these principles in detail. In general, the principles fall into two categories: those that improve the *vision* and those that improve *understanding* of the plot. In this section, we simplify and summarize Clevelend's principles for plotting data, for a full treatise on the topic, we recommend reading his books.

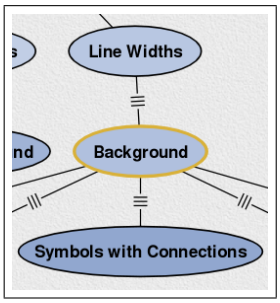
5.1.1 Improving the Vision

The first set of plotting principles deals with improving the vision of the plot. This could also be referred to as the readability of plot—the ability to visually disentangle all the information that is being presented.

Principle 1: Reduce clutter. The main focus of a plot should be on the data itself, any superflous elements of the plot that might obscure or distract the observer from the data needs to be removed. As an example, consider the default Excel plot in Figure 5.1. The low contrast background and dark horizontal grid lines draw attention away from the data. The Matplotlib plot in the middle is a little better because it leaves the area around the data white, but still uses an unnecessarily distracting gray frame around the data rectangle. In both of these cases, the plots fail to make the data stand out.

Principle 2: Use visually prominent data elements. The elements that are to represent the data need to be both distinct and prominent. Connecting lines should never obscure points and points should not obscure each other. If multiple samples overlap, a representation should be chosen for the elements that emphasizes the overlap, such as an alternate symbol for stacked points. If multiple data sets are represented in the same plot (superposed data), they must be visually separable. If this is not possible due to the data itself, the data can be separated into adjacent plots that share an axis (juxtaposed data). Of the three examples demonstrated in Figure 5.1, none show the data with visually prominent elements. The first two (Excel and Matplotlib) display a line that is not very visible due to the color and thickness. The third (Pages) has the opposite problem, the points symbols are so large they are difficult to distinguish visually.

Principle 3: Use proper scale lines and a data rectangle. The scale lines around the data rectangle are important for understanding the data values within the data rectangle. Two scale lines should be used on each axis (left and right, top and bottom) to frame to data rectangle completely. This serves two distinct purposes. First, it encloses the data points so that none of the information is overlooked. Second, it makes determining the data values at the extremes of the rectangle much easier. This is because our visual system is better at judging horizontal or vertical positions between a pair of tick marks than with only one. As discussed in the Principle 2, the data in the rectangle should remain prominent, this can be achieved by leaving a small margin between the data and the scale lines—the scale lines should never interfere with the data. Principle 1 should also be addressed with respect to the scale lines by using an appropriate number of tick marks and labels for each axis (3-10 is generally sufficient). By keeping these suggestions in mind, the scale lines can enhance the information being displayed without overshadowing it. Returning to the three plots in Figure 5.1, only the



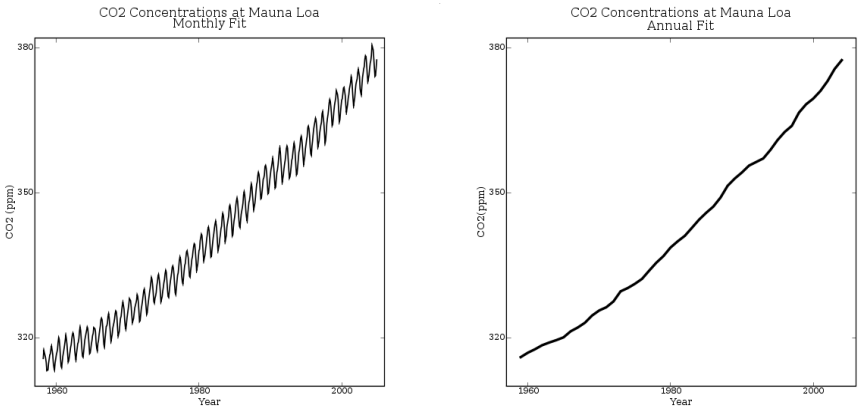


Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

middle (MatplotLib) follows this principle by using a proper scale line margin for the data and a manageable number of labels on each axis.

Principle 4: Be careful with reference lines, labels, notes, and keys. Reference lines are often used to show important values such as a threshold within the data. Labels and notes are similarly used to distinguish between different data points or draw conclusions. These types of elements should be used sparingly and in an unobtrusive way so as not to overshadow the data being represented. The data elements should be distinct enough from reference lines, labels, or notes, such that the correlations and trends can still be easily observed. The key for the data can also be distracting when displayed next to the data. When possible, this additional information should be moved to outside of the data rectangle to reduce the clutter.

These principles were applied to the Mauna Loa data using Matplotlib to produce the much improved plots shown in Figure 5.2. In particular, the margins were adjusted, the data lines were darkened, the gray frame was removed, and the labels and ticks on the axes were reduced.

5.1.2 Improving the Understanding

The next set of principles deals with improving the understanding of the plot. These principles ensure that the analysis of the plot is effectively communicated.

Principle 1: Provide explanations and draw conclusions. A graphical representation is often the means in which a hypothesis is confirmed or results are com-

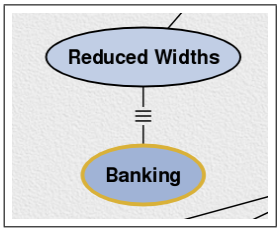
municated. Informative captions are often necessary to point out features in the data or to explain specific trends. Each element that is added to the plot should be properly explained to avoid confusion. In addition, since the plot and associated caption are highly visible, they should be properly proofread for correct content.

Principle 2: Use all available space. The empy space in the plot should be filled as much as possible horizontally and vertically. For skewed data that leaves excessive empty space, consider replacing the linear scale with a non-linear one (see Principle 4). It is often assumed that zero should always be included on the scale line, even if the data does not include zero in its range. The Pages plot on the right of Figure 5.1 uses this assumption. This should not be the case for scientific data, it should be assumed that the reader will look to the scale lines for clarification of the scale of the data.

Principle 3: Align juxtaposed plots. As mentioned previously, it is often desirable to extract data into separate plots to avoid clutter. This juxtaposition is also important when plotting higher dimensional data. These plots should be properly aligned along similar axes to facilitate comparison. Whenever possible, the scale lines should also be uniform across plots so that the reader is not deceived by the differences in the data. Figure 5.2 shows an example of two juxtaposed plots that are aligned along one axis and use the same scale. Because the default behavior for plotting software is to fit the data rectangle to the data, the scales usually require user intervention to make them uniform.

Principle 4: Use log scales when appropriate. Logarithmic scales are used to show multiplicity or factors in the data as well as to remove skewness that may leave much of the data clustered closely together. They can also be used in place of breaks in the scale for showing data that may have a few large values. Depending on the range of data, different bases may be used (*e.g.,* 2, 10, or e). When using a log scale, the axes should be properly labeled to draw attention to the scaling. In addition, it is often useful to show the log factor as well as the value for the tick marks by displaying each on a different axis (*e.g.,* the top scale contains values and the bottom scale contains the log factor).

Principle 5: Bank to 45°. The principle of banking to 45° was first introduced by Cleveland [Cleveland 93] as a means to automatically determine the aspect ratio of a plot. The slopes of the line segments that connect adjacent points in the plot is a visual indicator of the rate of change within the data. By optimizing the aspect ratio of these segments, the rate is more easily perceived. The obvious choice for optimizing in both horizontal and vertical directions is to use a slope of 1 (*i.e.,* 45°). To bank the data in a plot, the absolute values of the slopes for each line segment are averaged. This value is then used to adjust the aspect ratio until the average is 1. This method has recently been extended to multiple scales



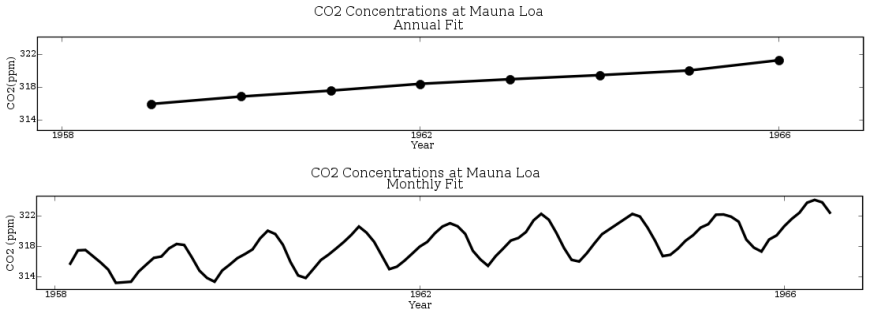


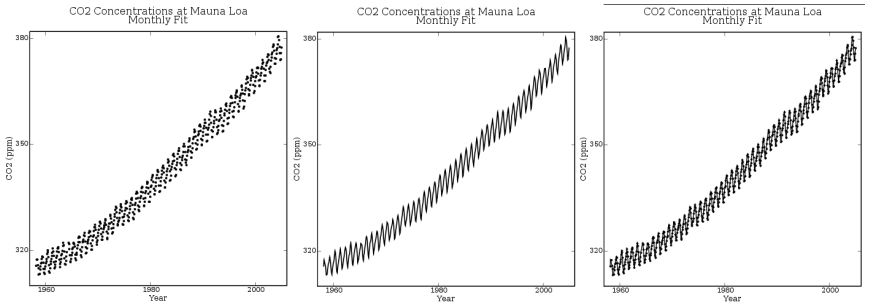
Figure 5.3: Plots of the Mauna Loa data set demonstrating the principle of banking to 45°. The annual data on top is improperly banked, making the change in rate difficult to see. The monthly data on the bottom is properly banked, making the change in rate much easier to see.

by banking not only the data, but the trends within the data as well [Heer and Agrawala 06]. Figure 5.3 shows an example of the principle of banking using a subset of the Mauna Loa data set. The yearly trend in the data is improperly banked, making it difficult to perceive any rate of change as is clearly visible in Figure 5.2. However, the monthly fit is now correctly banked, displaying features in the data that were not visible in Figure 5.2. The rate of increase is clearly not the same as the rate of decrease in the plot. Without banking, these features may go unnoticed.

These basic principles for improving the vision and understanding in the data are important for both analysis and communication of results. Often, they can be the difference in finding insight in the data. The techniques described here should be considered a rule of thumb, not a rigid standard. Different situations may require breaking one principle to conform to another—this is part of the challenge of the visualization process. But being aware of these principles can only help to improve the quality and speed in which this creative process can be accomplished.

5.2 Simple Plotting Techniques

We have covered the principles behind making clear and insightful plots. In this section, we address the problem of selecting a plotting technique based on the data to be visualized. As before, the information provided here should be considered as guidelines, since there is not a 'best plot' for all the data types that are acquired.



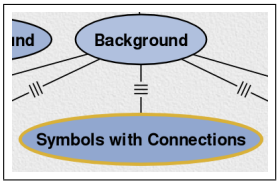


Figure 5.4: Plots of the Mauna Loa data set using only symbols (left), only connections (middle), and both symbols and connections (right).

5.2.1 Connected Symbol Plots

Connected symbol plots are probably the most common plotting technique. They are used for plotting a time series or other 1D data. For instance, a single measurement taken at different points in times or sequential samples from an experiment. There are three combinations of plotting elements that can be used depending on the properties of the data.

Symbols. For noisy data that shows high frequency characteristics, the symbols may be sufficient. This is especially the case when the trend of the data is more important for the viewer than the data points themselves. By only using the symbols, the plot may remain uncluttered.

Connections. For smooth data that shows low freqency characteristics, showing just the connnections is often the best choice. In these cases, the symbols will not provide any additional information to the viewer. Thus, they should be left out to improve the vision of the plot.

Connected Symbols. In many situations it is best to show both the symbols and their connections. The symbols demonstrate the actual concentrations of the data, while the path that the data takes can be better followed using the connections. For example, consider a data set is smooth in most places, but contains a large spike. With only connections, it may be unclear if the spike in the data is one outlier in the data or if it is an actual trend in the data. By using connected symbols, these features will be detectable.

Figure 5.4 shows the Manua Loa data using these three different combinations of elements. With this data, the symbols alone are difficult to visually assemble and the connected symbols do not provide any real additional information. In this case, the connections are sufficient to express the data.

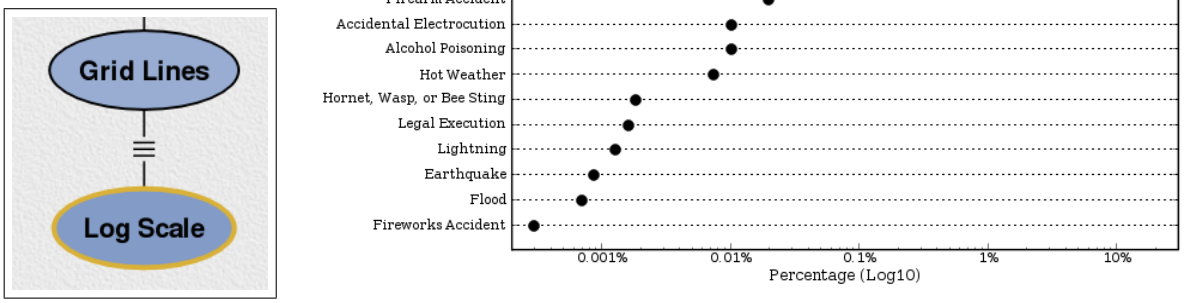


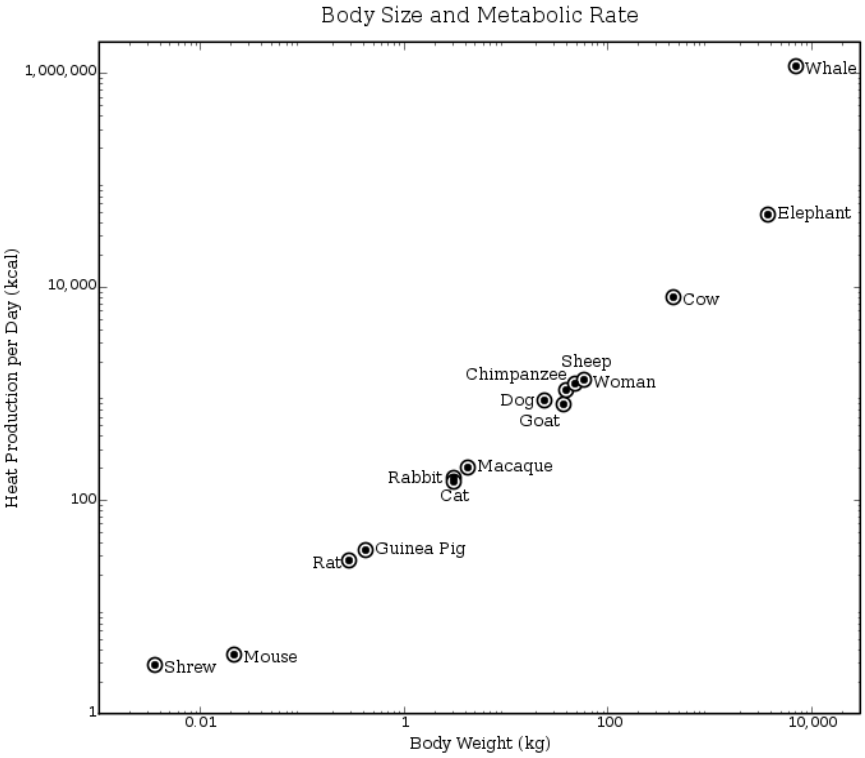
Figure 5.5: A dot plot showing the odds of dying. The log scaling prevents skewness and allows factors to easily be perceived.

5.2.2 Dot Plots

Another common plotting technique is the use of dot plots **??**. These are similar in nature to bar charts or pie charts, and should be used for quantitative labeled data. Figure 5.5 shows an example of a dot plot of data taken from the National Saftey Council statistics on deaths in the United States in 1993 [Council 93]. The data categorically represents the ways that people died in one year, represented as a percentage of all manners of dying.

The data in a dot plot can be arranged in several ways. The values should normally be sorted such that the largest value is at the top (as in Figure 5.5). The exception to this is when the labels of the data have an inherent order that must be preserved. As with other plotting techniques, a log scale should be used to reduce skewness in the data. Figure 5.5 uses a base 10 log scale that assists in the comparisons between many small percentages. For instance, the log scale in the plot shows that a person in the U.S. is almost 10 times more likely to die in firearm assault as they are from a firearm accident.

Unfortunately, in the real world data is not always univariate. To represent multi-dimensional data, a multiway dot plot can be used [Cleveland 84]. A multi-



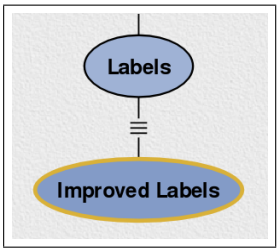
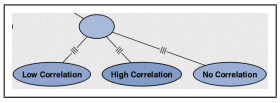


Figure 5.6: A scatter plot showing the biological principle of scaling for mammals. For each sample, the metabolic rate is plotted against the body mass to show a high correlation between the two variables. The points have also been labeled to provide additional information.

way dot plot is just several dot plots that share common labels and are juxtaposed such that they share an axis. As mentioned in § 5.1, ensuring that scale lines of the juxtaposed plots are uniform is essential for any useful comparisons.

5.2.3 Scatter Plots

Scatter plots are used to show how one variable is affected by another, or correlated, in 2D data. The two variables of data is generally mapped to the horizontal and vertical axes of a cartesian grid. The data elements are represented using symbols that should be both visible and distinct. Figure 5.6 shows an example of a scatter plot for data measured from mammals showing the effects of biological scaling [Kleiber 47]. On the horizontal axis, the log scale of body mass is represented and on the vertical axis, the log scale of metabolism is represented. Each point in the data represents one measurement for both these variables for



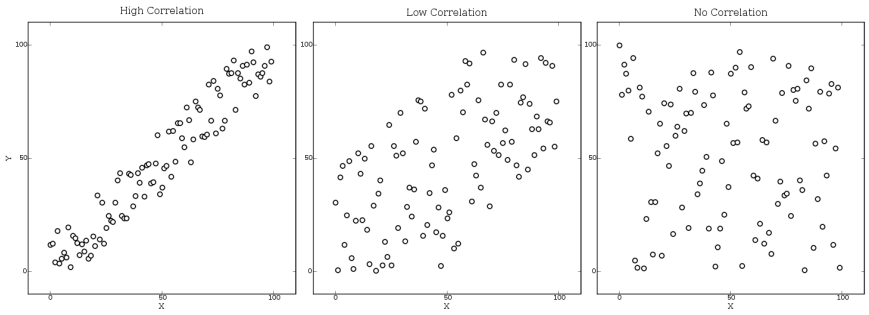


Figure 5.7: Scatter plots showing different levels (high, low, and no, respectively) of correlation for points generated with different magnitudes of randomness.

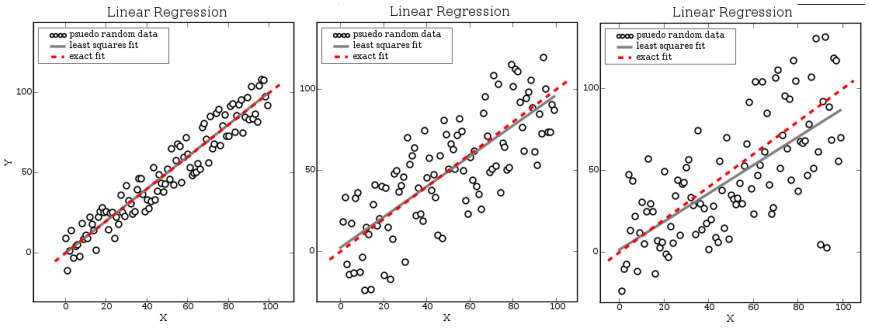
one mammal. Since the data has distinct labels, these have been added to the lable to provide additional information. Care was taken to make the symbols in the data stand out and keep the labels from obscuring the data and making the trend difficult to perceive.

One important aspect of scatter plots in that with enough samples, the correlation of the data can easily be computed. Figure 5.7 shows an example of the three types of correlation that emerge from dot plots: high, low, and no correlation. Though our visual system is generally good at perceiving the difference between these types of correlation, there are actual measures for it. One of the most common is the Pearson product-moment correlation coefficient [Pearson 96], which assumes linearity between the two variables. For a n measurements of x_i and y_i, where $i = 1, 2, \cdots, n$, the coefficient r can be computed as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.1}$$

where $\bar{x}$ and $\bar{y}$ are the means of the samples. Intuitively, the coefficient can be though of as a dot product between the two vectors that fit the two sample variables, or equivalently, the cosine of the angle between these vectors.

There have been several suggestions for interpreting the correlation based on the computed coefficient. For example, Cohen [Cohen 88] suggests that for possitive (increasing) correlation, values greater than 0.5 are highly correlation, while values less than 0.3 have no correlation. Obviously, these criteria are data dependent and should not be used as a guarantee of correlation, the definition of which is often be context dependent. In addition, the Pearson coefficient that has been described is highly sensitive to outliers, thus for certain data samples it would be a completely inaccurate measure. In such cases, more sophisticated methods such as rank correlation that do not assume linearity should be used instead [Spearman 04, Kendall 70, Snedacor and Cochran 89]



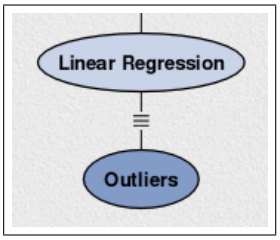


Figure 5.8: Linear regression using least squares fits a line to the data. The fit is good for high and low correlation (left and middle), but can result in problems in the case of outliers (right).

It is often desireable to express the correlation as a line that provides the best fit for the data. This fit is referred to as linear regression. The earliest form of linear regression is the method of least squares, which was introduced by both Legendre [Legendre 05] and Gauss [Gauss 09] as a means of predicting celestial paths. For the simple case of a regression line $y = a_0 + a_1 x$, the idea is to find the parameters a_0 and a_1 such that the summed squares of the vertical distances between the regression line and the function f data points are minimized:

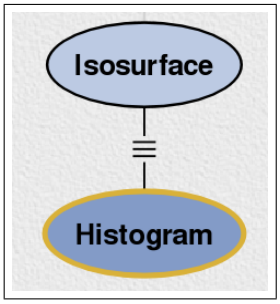
$$\sum_{i=0}^{n} (y_i - f(x_i))^2.$$

This minimization can be solved using a linear system:

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where ε_i are the errors between the samples and the linear regression after the fit. Figure 5.8 shows a few examples oflinear regression lines computed using this method. This summary and example use a simple regression line for the fit because we assumed data that follows a straight line, this is obviously not always the case. The regression fit can easily be extended to other functions (*e.g.*, $f(x) = a_o + a_1 x + a_2 x^2$) using the above matrix representation, but with additional columns in the x matrix and additional values in the a vector.

The method of least squares is sensitive to outliers in the data. For example, Figure 5.8 shows how two outliers can disrupt the fit. To this end, many methods have been introduced to weight points differently (weighted least squares)



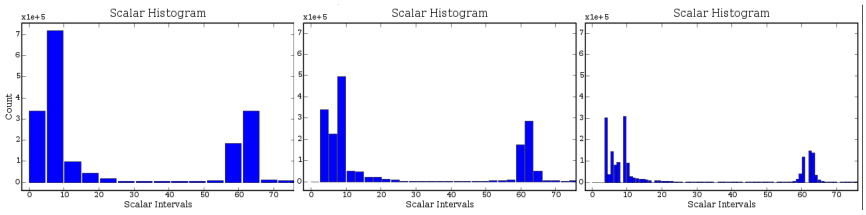


Figure 5.9: Several histograms of a CT data set of a head. The views for all have been zoomed into important regions to show how interval selection changes feature detection. The number of intervals selected are 50, 100, and 284 from left to right.

or ignoring points outside of a range of influence (moving least squares). These methods are more robust for data with large error and may be necessary, depending on the data.

As with dot plots, scatter plots can be used to represent data in higher dimensions. This is frequently done with a scatter plot matrix [Chambers et al. 83]. This assigns each dimension of the plot to a single row and column in the matrix. The variables are then plotted against each other as a standard scatter for each entry in the matrix. The diagonal of the matrix is generally left as a label for the entry that corresponds to the row and column. One side effect of this is that each plot is actually mirrored on the the upper and lower diagonals of the matrix. In general, however, the method is very good for correlating multidimensional data.

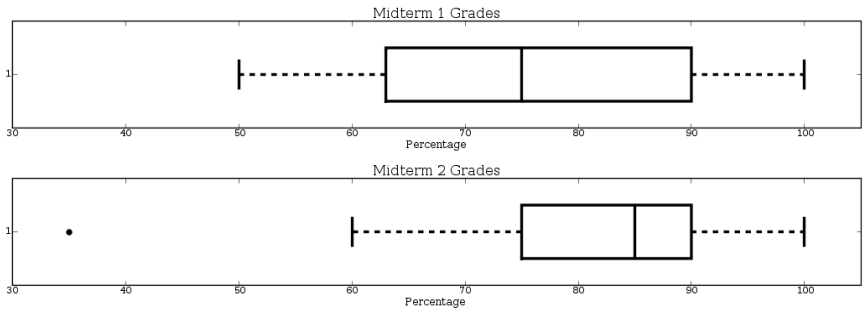
5.2.4 Histograms

Histograms are a special type of bar charts used for plotting distributions in data. They are typically used for a large number of data values because they reduce the information being displayed. The horizontal axis represents fixed intervals of the data and the vertical axis represents the number of values that lie within the intervals. Figure 5.9 shows examples of histograms of varying interval sizes generated from a CT scan of a head. Note that as the number of intervals increases, the larger distributions are broken up into smaller distributions and features that were hidden before become apparent. Because the choice of interval width determines the accuracy of the histogram, it should be chosen with care.

Methods have been developed to optimize the interval width by estimating the probability density function of the data. For n samples, the most common methods are to use the standard deviation σ [Scott 79]:

$$W = 3.49\sigma n^{-1/3}$$

or more robustly, the interquartile range (IRQ), which is the 75th percentile minus



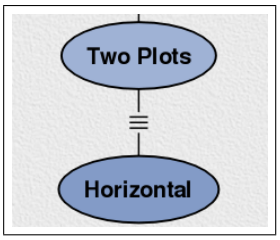


Figure 5.10: Box plots showing the statistical variation in the grades for two midterm exams.

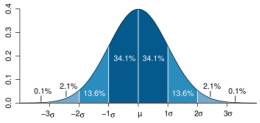
the 25th percentile [Freedman and Diaconis 81]:

$$W = 2(IQR)n^{-1/3}.$$

Using this definition on the CT data in Figure 5.9, the optimal bin width occurs at 284 intervals. This means that using more intervals will not likely show additional features of importance within the data. Keep in mind, however, that as with the computation of correlation, these methods are approximations that are only good for most data types. As such, they should not be considered a replacement for actual explorations of different histogram widths.

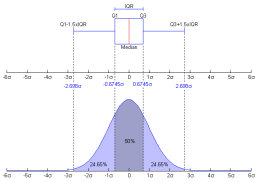
5.2.5 Box Plots

It is not uncommon for measurements to contain variation or uncertainty in the accuracy. To represent the statistical variation that is inherent in both empirical measurement as well as sparse sampling, several methods have been developed. One of the most prevelant is to represent the variation with error bars on a dot plot or a connected symbol plot. The dot or symbol represents the mean value, and lines with bars at the end (sometimes called *whiskers*) that extend from it represent the standard deviations. This effectively covers about 68% of a normal distribution. At times, to represent more accuracy, the bar is placed at 50% of the distribution and the line is extended beyond the bar to represent a higher interval, such as 95%. These values are referred to as the *confidence interval* for the statistical variation. Because the confidence intervals used may be data set dependent, it is important to document the choices of values used.



Redo this figure and add corresponding one and two tiered error bars.

Standard error bars have several disadvantes. First, they do not represent the shape of the distribution because it uses the mean value and is thus centered with equal length standard deviations. Second, they do not capture the outliers that may be of importance for the evaluation of a distribution. As a better method for representing statistical distributions, box plots were introduced **??**. Instead



Redo this figure.

of standard deviations, box plots represent the interquartile range (IQR), or in essence the spread, of the data. Figure 5.10 shows two examples of box plots that express the variation in the grades for two midterm exams. The box itself covers the IQR, going from the lower quartile (25th percentile) to the upper quartile (75th percentile). The median (50th percentile) is represented by a line or dot within the box. The whiskers of the plot extend to adjacent values of the plot, which are defined as the values closest to the lower quartile minus 1.5IQR and the upper quartile plus 1.5IQR. Any data values outside the adjacent values are considered outliers and drawn as point symbols.

Box plots are efficient for analyzing distributions. As shown in Figure 5.10, the shape of the distributions are readily apparent. The median grade improved significantly from the first to the second midterm exams, even though the number of students achieving a grade over 90% remained the same. By using quartiles, the outlier shown in midterm 2 does significantly effect the representation for the rest of the class as it would with a regular error bar that would be based on mean grades. Finally, because the size of the box on the right of the median is much smaller than below, the box plot shows that the distribution is skewed. This means that there was a larger variation in scores below the median (about 88%) than there was above.

As with histograms, box plots are a data reduction technique. They are very efficient for analyzing and comparing the statistical distribution in data, but may be a poor choice if the emphasis needs to be on the actual data values. Thus, depending on the application, other methods might be necessary along with box plots for a full analysis.

5.2.6 Others

There are other common plotting techniques that are used by the mass media such as pie charts, bar charts, and area plots. Cleveland [Cleveland 94] refers to these methods as *Pop Charts*. They are used much less frequently in the sciences because they typically result in poor pattern perception when compared with the techniques we have described above [Bertin 73,Tufte 83]. In particular, our visual system is better at recognizing changes in horizontal or vertical directions than comparing changes in area. An additional problem occurs when the data does not use a zero base for the scale—this type of data is not easily represented by any of these pop chart methods. In most cases, dot plots (see § 5.2.2) can be used instead for data that could be represented with a pop chart.