Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 1: Overview of Statistics and Exploratory Data Analysis

(I.) Overview of Statistics

Originally, the word "Statistics" referred to the political science concerned with the facts of a state or a community. It was only at the end of the 19th century that the term acquired its present meaning and began to denote the science of collecting, organising, and interpreting data, usually with the goal of inferring certain properties of a population from the properties observed in a representative sample. The information collected about the representative sample often takes the form of measurements of one or more "variables" for the "individuals" or "units" in the sample, where a "variable" is loosely defined as a characteristic whose values change across members of a population. The resulting collection of measurements is termed the "data set" or the "data." In this case, a "data set" refers to the measurements of p (≥ 1) variables of interest for each of n (≥ 1) units. As an example, we might consider a data set consisting of the age, socio-economic background, educational level, years of experience, performance rating, and pay of 100 individuals randomly selected from a certain division of a company. This type of data, i.e., a collection of the values of certain variables for more than one unit, where each variable is measured at only one point in time, is referred to as "cross-sectional data." The two other types of data include "time series data," which results from measuring a variable for one unit at several points in time, and "panel data" or "longitudinal data," which is a cross between cross-sectional and times series data and consists of measurements of variables, possibly observed at more than one point in time, for n (≥ 1) units. An example of time series data could be a data set containing the revenue of Coca-Cola from 1982-1987, and an example of panel data might be a data set containing the blood pressure levels for 50 individuals given a placebo and 50 individuals taking blood pressure reducing medicine, measured at 6 month intervals over a five year period. In this course, however, we will focus on cross-sectional data.

Before turning to a discussion of the reasons for collecting and analysing cross-sectional data for a representative sample, we should note the important division between a "population" and a "sample." The term "population" means the entire collection of units or measurements of those units about whom or which information is available, or, in a

slightly more formal statement of the latter case, the set of values for one or more variables taken by all units. The term "sample" denotes the subset of the population selected for study or the values of the variable(s) recorded for those selected units. Another essential dichotomy is that of "parameters," which are numerical characteristics of a population, and "statistics," which are numerical characteristics of a sample. Further, an "estimator" is a statistic that is specifically designed to measure a particular parameter of a population.

Usually, we are interested in learning about certain "attributes" or properties of a population, such as its parameters or, alternatively, its "structure" or "distribution." However, in most cases, we cannot observe a population's attributes because doing so would require analysing the whole population, which is generally not possible for a number of reasons (e.g., it may take a lot of time and money, it may cause the destruction of the population, etc.). So, instead, we select a sample from the population, collect cross-sectional data for it,[1] and then, from the resulting data, calculate certain statistics and/or produce certain graphs that correspond to the population attributes of interest; hopefully, these statistics and graphs adequately resemble the population properties to which they correspond. As an example, if we were interested in knowing the mean height of a population of students, we could select a sample from that population and then obtain its sample mean; we would hope that the sample mean would provide a reasonable estimate (i.e., an approximation) of the (unknown) parameter value. This science of deducing the properties of an underlying population from a data set is called "statistical inference." In this course, we will focus on using cross-sectional data sets to draw inference about their underlying population; the methods used to draw inference about the populations underlying times series and panel data sets, such as spectral analysis and mixed effects models, respectively, will not be addressed in this course, but represent important areas of statistics.

Throughout this course and in the future, it will be important to bear in mind that, in most cases, the ultimate goal of data analysis is to draw inference about certain attributes of an underlying population using the information contained in the data sampled from that population.

(II.) Sample Design, the Phases of Statistical Analysis, and the Structure of This Course

Once a researcher has defined the population in which he/she is interested, a variety of methods for sampling from that population exist; these sampling methods and the resulting data set will be discussed in Sections IV and V below after a short digression on the role of computers in statistical analysis. Once a sampling method has been decided upon and data collection, which will not be covered in this course, has been carried out,

[1] Frequently, sample selection and data collection will be performed by someone other than the data analyst.

the statistical analysis of the resulting data set should proceed in four stages, as described by Cox and Snell (1981: 6). Before presenting these four stages, we should note that two stages of analysis precede statistical inference. In fact, it is almost uniformly recommended that, before embarking on any formal statistical analysis, one should calculate "descriptive statistics" for the data, which are numerical methods for organising and summarising data. Only Phase 3 below corresponds directly to statistical inference, which entails calculating "inferential statistics," i.e., methods that are used to generalise conclusions obtained from a sample to the population from which the sample was selected. As a note, this descriptive vs. inferential division of statistics is often employed, but it is important to be aware that a given statistic can be used either to summarise the data or to draw inference about the underlying population (e.g., the sample mean).

(1.) Initial data manipulation: This involves putting the data in the right format to carry out checks of data quality before any initial analysis of the data commences. This step is an important one and should include: reviewing the method(s) of data collection in order to look for possible sources of bias that might invalidate conclusions drawn from the data, checking for discrepant observations (some of which might be the result of measurement or typing errors), and searching for missing observations.

(2.) Preliminary analysis or "Exploratory Data Analysis" (EDA): Once the relevant checks have been performed, simple analysis of the data should be done in order to clarify their general form, to again check for discrepant or "outlying" observations, to suggest possible directions for more complicated analysis, and to investigate assumptions required by the subsequent definitive analysis. Usually, this step involves producing graphical, tabular, and numerical summaries of the data.

(3.) Definitive analysis: This phase entails using formal and possibly also informal techniques of statistical inference in order to draw conclusions about certain attributes of interest for the underlying population.

(4.) Presentation of conclusions: This step involves presenting the graphical and numerical results from 2) and 3) in an accurate, concise, and lucid form. Usually, these results are then interpreted in light of the subject matter at hand.

In this course, we will skip over Phase 1 and assume that the data are in satisfactory condition for the subsequent stages. Phase 2 involves using a variety of descriptive statistics, tabular methods, and/or graphical methods, many of which will be presented momentarily in Section VI. As for Phase 3, it entails using various techniques to draw inference about the population attributes of interest, as was stated above; these statistical inference methods will be presented in Lectures 3-8. However, since these methods are derived using probability theory and rely on probabilistic assumptions, Lecture 2 will be devoted to a brief overview of probability that is necessary to fully comprehend the methods of statistical inference presented in Lectures 3-8. Lastly, we turn to Phase 4, which will not be addressed in this course. However, we will note here two points relating to this phase: first, a simple presentation of results is often best, and secondly, it is

important to remember that statistical significance does not necessarily imply scientific significance.

(III.) The Role of Computers in Statistical Analysis

Computers can be used in each of the four phases of statistical analysis enumerated above, as well as for data collection and entry. More specifically, computers can be used to perform the following tasks:

- *Data collection*: preferably done using pre-designed software that works for the data of interest.
- *Data entry*: can be done using a spreadsheet or a statistical package.
- *Data checking (Phase 1)*: might include checking for double entries and logical errors.
- *Data screening (Phases 1 and 2)*: checking for outliers and missing values and getting a feel for the data.
- *Definitive analysis (Phase 3)*: the specific analyses used will depend on the design of the study, the type of data and variables collected, and the questions of interest.
- *Presentation of results (Phase 4)*: again, the analyst should remember that a simple presentation of results is often best and, thus, should resist the temptation to use too many colours or unnecessarily complex graphics.

Using a computer to perform statistical analysis results in numerous advantages, including: increased accuracy, speed, and versatility; the ability to handle greater amounts of data with only marginally more computation time; the facility with which graphics can be produced; the ease with which slight modifications can be made to a given analysis; greater flexibility and the ability to create new variables; and the capacity to transfer data either electronically or on disc.

There are, of course, also certain disadvantages, such as: there might be errors in the software used (sometimes even in the best written programmes); the versatility offered by the software makes it easy to use an inappropriate statistical procedure; similarly, statistical software often allows the researcher to perform complex analyses when he/she doesn't understand them or when simple analyses would be sufficient and possibly more appropriate; and, lastly, because a computer allows many analyses to be made quickly and easily, it can facilitate "data dredging," which refers to searching for significant relationships by performing a large number of analyses. Here, it is worth quoting Cox and Snell (1981:24) at length on the problems that may ensue when a computer is used to perform statistical analysis:

> In the central analysis, the wide availability of computer packages has made quite complicated methods of analysis painlessly available even to those with little statistical training, and the dangers of this have often been discussed. There is another side, however; it is now also painless for the statistician to generate large numbers of plots of residuals, tests of normality, analyses on numerous scales, etc., and thus to get very large amounts of computer output even for very simple data. Without some ruthlessness in appreciating that the great majority of this must be discarded, the objectives of achieving and reporting simple conclusions are threatened.

As a result, caution must be exercised when using computers to perform statistical analysis. Before using a package, one should first understand what it is doing; the best package is that which one understands best. In addition, it must be borne in mind that the computer has no magical powers; it merely follows orders, and it is only a tool to help the researcher perform the planned analyses. Further, far from being a helpful tool, some software packages may prevent a researcher from drawing valid conclusions from his/her data because the package's functions contain errors. It should not necessarily be assumed that the results from a software package are valid, and results should be checked, at the very least for reasonability. Lastly, researchers should guard against the temptation to produce large amounts of computer output not required by the planned analysis.

The above caveats aside, using a computer to perform statistical analysis is generally advisable.

(IV.) Sample Selection

As was stated above, in this course we will be focusing on data of the cross-sectional variety only, which consists of the values of p (≥ 1) variables of interest for each of n (≥ 1) units in a representative sample of the population of interest. This type of data is especially common in applications such as demographics, medicine, psychology, sociology, and zoology, and it is relatively straightforward to draw inference about its underlying population compared to doing so for times series and panel data.

For one such data set, the questions that the data analyst currently hopes to answer using it may not coincide with the initial objectives driving the data set's collection. This is possibly true because the goals of the data collector have changed or because the person who designed and implemented the method of data collection may not be the same as the data analyst. This said, the data may have been collected with either of two types of goals in mind:

- descriptive inference: when the main objective is to describe, in some way, a large group, using information from a sample from that group.
- analytical inference: when the main objective is to study the properties of and relationships between variables, using a small sample, in hopes that the results from that sample can be generalised to a larger population.

An example of the first type of goal might be a desire to determine the racial composition of a major American city, and an example of the latter goal might be a study of the effects of taking vitamins on longevity. Additional examples of both types of goals in different application areas include:

Science and Industry:	Clinical trials, experimental design, quality control
Government:	Population and Economic trends, official statistics
Market Research:	TV ratings, readership of newspapers, impact of advertisement campaigns, opinion polls

In addition to stating the goals of data collection and the questions that he/she hopes the study will address, the data collector should also clearly enumerate the variables that will need to be measured in order to answer these questions and also clearly define the population of interest. This population may be finite or infinite. In the latter case, it is obvious that only a sample from the overall population can be studied. However, in the former case, it might have been possible to measure the variables of interest for all units in the population, which would be termed a "census." However, following Cochran (1980), we can establish the following advantages of collecting data for only a sample from the population rather than conducting a census:

- *Reduced cost:* Lower costs may result in instances where it is difficult to locate units or to measure the variables of interest.
- *Greater speed:* The collection, screening, and analysis of the data are performed more quickly for a sample than for a census.
- *Greater depth:* The gains in speed and cost that are achieved by sampling can be applied to increasing the breadth and depth of the study (i.e., a greater number of variables can be measured).
- *Greater accuracy:* Sometimes obtaining truly accurate measurements requires highly trained personnel and/or specialised equipment, which, of course, makes it inadvisable to examine the whole population because doing so might result in less accurate measurements.
- *Limited destruction:* In some cases, a unit must be destroyed in order for the variables of interest for be measured, which obviously makes it undesirable to do so for the entire population

In this course, we will address only data sets that represent samples from the population of interest rather than those which contain all units in a population. In addition, for probabilistic reasons, we will assume throughout this course that the underlying population of interest is infinite, unless stated otherwise. Here, we should note the distinction between the "target population," which is the one about which we desire to draw inferences, and the "population to be sampled," which consists of those units for whom it is possible to measure the relevant variables. The population to be sampled is

often much more restricted than the target population, and the sample must, of course, be drawn from this more restricted population.

Once the data collector has clearly defined the above populations, it is necessary to decide what method of sampling will be used to select a subset of the population to be sampled. Different methods of sampling include simple random sampling, stratified sampling, and cluster sampling, among others; in addition, sampling can occur in one or more phases. The theory behind and examples of these sampling schemes are provided in the classic texts of Cochran (1980) and Hansen *et al.* (1953).

A "simple random sample" of size n is a subset containing n units from the population of interest. These units are chosen in such a way that every possible subset of size n has the same probability of being selected as any other; in other words, every unit in the relevant population has the same probability of being selected for the sample.

However, sometimes the population to be sampled is divided into non-overlapping subpopulations called strata, across which it is suspected that the answers to the questions of interest may differ. Although we desire to sample from the entire population (i.e., the union of these strata), we might do so by collecting a "stratified" rather than a simple random sample because using the former may improve the precision of our estimate of the parameter of interest. In this stratified sampling method, the total desired sample size is divided between the strata in a manner that reflects the properties of the variables of interest within each stratum and the costs of sampling within each stratum. Once the number to be sampled from each stratum has been fixed, the appropriate number of units is selected, randomly, from each stratum. An example of a stratified sample might occur if a researcher desired to estimate the average height in the U.S. adult population and, after deciding on a total sample size of 1000, randomly sampled 500 females and 500 males from the U.S. population because he/she suspected that the mean height might differ drastically for these two groups.

In cluster sampling, the population is again divided into non-overlapping groups or "clusters"; however, in this case, the groups are not assumed to differ systematically but, instead, each is assumed to be representative of the entire population. If this case, a "cluster sample," in which only several of these clusters are first selected and then units are randomly sampled within each cluster, might be collected because doing so could reduce the time, money, and effort required to collect data. An example of cluster sampling might occur if a researcher hoped to estimate the mean parental income of Oxford students. If he/she assumed that all colleges were more or less the same in this regard, he/she could randomly select several colleges and then measure parental income for randomly selected students within these colleges only.

To compare stratified sampling to cluster sampling, in order to employ the latter method, we should suspect that the answers to the questions of interest vary minimally across groups (clusters) and maximally within groups, whereas, for stratified sampling, we should suspect that the answers vary maximally across groups (strata) and minimally

within them. Further, in stratified sampling, all the strata are sampled from, but in cluster sampling, not all the clusters are sampled from.

Here, we should note that there may be several layers of clusters or strata; for instance, the population of interest could be divided into school districts, each of which could then be subdivided into its component schools. If this were the case, then "multi-stage" sampling, in which groups, then subgroups, and then units within subgroups are selected, could be employed. Obviously, this method can be generalised to include more than two levels of grouping.

Lastly, sampling does not necessarily occur in only one phase. As an example of "multi-phase sampling," an initial random sample from the overall population might be used to estimate certain properties of the variables of interest in each of the strata in that population. Then, the resulting estimates could be employed in order to determine how the total number of units that will be sampled in the second phase should be divided amongst the strata for more in-depth sampling in that phase.

Before proceeding, we should note that all of these sampling methods employ random selection in one way or another. The importance of choosing units randomly is twofold: it helps to avoid biases, and it is an explicit assumption of many statistical methods. In fact, many statistical methods that we will examine in this course assume that the data at hand comes from a simple random sample of the population.

(V.) The Resulting Data Set

Types of Variables

Once the goal of the study has been clearly established, the target population and population to be sampled have been identified, a sampling scheme has been settled upon and units selected in accordance with that scheme, and the variables of interest have been measured for those units, a data set will result. This data set contains measurements of p variables for n units. Each of these p variables will belong to one of the following six variable types. A variable's type is determined by the set of values that it can potentially take, and each of the possible types belongs to either the categorical variety or to the numerical variety.

(a.) Types of Categorical Variables:

(1.) Nominal: a variable that can take on only a finite set of values (i.e., the units fall into a finite set of "categories"), where these categories or "levels" have no intrinsic ordering (e.g., race).

(2.) Ordinal: a variable that can take on only a finite set of values, where the categories in this set do have an intrinsic ordering, but not on a well-defined scale (e.g., quality of service: poor, decent, good, excellent).

(3.) Interval: a variable that can take on only a finite set of values, where the categories in this set not only have an intrinsic ordering, but also have numerical scores or labels attached (e.g., quality of service: 1-5). These labels are often treated as category averages, means, or medians, and the differences between them can be used as a measure of the separation between two categories. This type of variable can result from coarsely observing a numerical variable, i.e., when the possible range of values for a numerical variable is divided into a number of bins and only the bin location is observed for each unit.

Before enumerating the types of numerical variables, we should note that the categories of the three variable types above are, to some extent, arbitrary and do not have a strict numerical interpretation. As another point, a categorical variable with only two categories can be referred to as "binary" or "dichotomous," whereas a categorical variable with more than two levels can be referred to as "polytomous." However, a polytomous variable can always be reduced to a binary variable by merging categories.

(b.) Types of Numerical Variables:

(4.) Discrete: a variable that takes on integer or counting number values and is, in effect, counting the numbers of occurrences of some phenomenon (e.g., number of siblings, number of lectures attended in one term).

(5.) Non-ratio continuous: a variable that takes on values along an effectively continuous, but relatively ill-defined, scale. Often, for variables of this type, a value of 0 should not be considered as the lack of the characteristic, and the variable's scale will not be linear. For example, for this type of variable, the difference between, say, 5 and 10, may not the same as the difference between 80 and 85. This would also mean that a value of, say, 40 would not correspond to twice the value of 20 or half the value of 80. For instance, consider the variable 'temperature': in this example, 0 C does not imply an absence of heat, and a day that is 40 C is not necessarily twice as hot as a day that is 20 C.

(6.) Ratio: a variable that takes on values along an effectively continuous and well-defined scale. For variables of this type, a difference of one unit has the same interpretation at any part of the scale, and a value of 0 truly denotes the absence of the characteristic (e.g., height, weight).

Here, we should note that types (5) and (6) are often referred to as "continuous" variables; often, no differentiation is made between these two types of variables. Due to the limitations of measuring devices, variables of the continuous variety cannot actually take on any real-numbered value in practice, and they are thus only effectively continuous and

not continuous in a strict mathematical sense. Lastly, we should point out that numerical variables can be reduced to interval categorical variables by coarsening their values (i.e., by grouping their values into bins).

No matter which type of variable(s) we are considering, when we are interested in the properties of only one of the data set's p variables, the resulting analysis is called "univariate"; alternatively, an analysis that investigates the relationship or association between two variables is termed "bivariate," or, similarly, "multivariate" for two variables or more.

Distributions

For variables belonging to any of these six types, their measured values will always vary across members of a population or across members of a sample from that population. For a given variable, some values will appear more often than others in the population and also probably in a randomly selected sample from this population. The pattern of occurrence of the various values of a variable is called its "distribution." Primarily, a distribution describes the possible values that a variable can take and the relative frequency with which each of these different values occur; this description might occur in a variety of forms, such as a mathematical function, a table, or a graph. The distribution of values for all units in the relevant population will be termed the "population distribution," which will be discussed in Lecture 2; population distributions cannot usually be observed because the entire underlying population is not known or has not been observed or measured. The distribution of values for the units in a sample selected from the relevant underlying population will be termed the "empirical distribution," which can, of course, be observed. In many cases, we will assume that the empirical distribution for a sample is a good representative of the underlying population distribution. As an example of an empirical distribution for a categorical variable, consider the data obtained by R. Wolf, who, in 1882, tossed a die 20,000 times and recorded the number of times each of the six different faces showed. The resulting empirical distribution was:

Face	1	2	3	4	5	6
Frequency	3,407	3,631	3,176	2,916	3,448	3,422

[N.B. In this example, the variable of interest is a categorical variable of the nominal type. This is the case because the values simply denote which face fell, and the numbers attached to each face are arbitrary labels without an order structure rather than quantities (unless we were specifically interested in the *number* of points on the showing face rather than the showing face itself).] As for an example of the empirical distribution of a continuous variable, it is not possible to present one in the tabular form used above without first placing the variable's observed values into bins (i.e., intervals); this is the case because, for continuous variables, most numerical values are observed only once in the sample. Once the continuous variable's possible values are divided into bins, the number of observations in each bin can be counted, and a frequency table similar to the

one above can be constructed. However, although the empirical distribution of a given continuous variable can be presented in tabular form, a graphical portrayal will often be more easily interpretable.

The above discussion refers to "univariate" distributions since only one variable at a time is considered. However, if we consider more than one variable at a time, we can talk about "multivariate distributions" or the "joint distribution" of the variables. A joint distribution for q variables describes all possible q-tuplets of values taken by the variables (e.g., for q=3, [length, width, height]) and the relative frequency with which each of these possible q-tuplets occurs. Again, the distribution might be summarised in functional, tabular, or, for q=2, graphical form. The "population joint distribution" (Lecture 2), which usually cannot be observed, refers to the distribution of the values of the q variables for all units in the underlying population, and the "empirical joint distribution," which can be observed, refers to the distribution of the values of the q variables for the sample units. As an example of an empirical joint distribution for two categorical variables (i.e., an empirical bivariate distribution), suppose that R. Wolf had rolled a red die and a green die (at the same time) 20,000 times and recorded the resulting pair of faces for each roll. Then, the empirical distribution could be summarised by the following "two-way table," which crosses the possible categories for the two categorical variables and then records the number of observations that fall into each of the 36 classes that result from these crossings. The numbers of observations in these classes will be termed the (absolute) "joint frequencies."

Joint Freq.	**Green Face**					
Red Face	1	2	3	4	5	6
1	632	531	499	543	623	573
2	498	598	501	610	576	654
3	664	629	612	432	527	453
4	582	743	653	476	694	467
5	489	624	537	492	476	515
6	535	467	598	578	435	484

As was true for the empirical distribution of one continuous variable, the empirical bivariate distribution for two continuous variables can be presented in the above tabular form if the possible continuum for each of the variables is divided into intervals or categories, these categories are crossed to produce classes, and then the number of observations falling into each class is counted.

(VI.) Exploratory Data Analysis

We have now discussed possible reasons for collecting data, the population from which the data is collected, the possible schemes for sampling from that population, and various characteristics of the resulting data set. Next, since we will assume that glaring mistakes

in the data set have been investigated and corrected and that missing values have been noted[2], we can skip over Phase 1 of data analysis and proceed to Phase 2, which is often called Exploratory Data Analysis or EDA.

There is a myriad of reasons for performing EDA before addressing the ultimate questions of interest using methods of statistical inference. To begin, EDA techniques can reduce the information contained in a data set to a few indicators that describe or summarise its main characteristics and therefore give the analyst and his audience a better overall picture of the data. Although some particular features of the data may be lost by summarising the data, doing so could reveal certain trends or patterns in the data, which might be relevant to the questions of interest. Second, certain EDA techniques will highlight departures from these trends/patterns in the data set; these departures are known as "outliers" and can be thought of as values that cause surprise in relation to the majority of the sample. Although outliers often result from measurement or recording errors, they can also simply correspond to anomalous units. Lastly, EDA can provide the opportunity to informally investigate the assumptions that will be required for the statistical inference phase (Phase 3) of data analysis. For example, some methods of statistical inference require the assumption that the population underlying the data for one variable has a normal distribution; by making a histogram of the observations for that variable, the analyst can get an idea of whether or not this assumption seems reasonable.

EDA methods can be numerical (i.e., descriptive statistics), graphical, or tabular. The specific method of EDA that is appropriate depends on whether the properties of only one variable (univariate analysis) or the relationships between multiple variables (multivariate analysis) are being investigated. The specific EDA method that is appropriate also depends on the type of variable(s) being considered. In general, there are more possibilities for describing numerical variables because we can perform arithmetical operations on their values and also because we can appropriate methods designed for categorical variable if we divide the continuous variables' possible ranges of values into interval classes.

(VI.a.) Tabular EDA Methods

For a nominal or ordinal categorical variable, a frequency table or "one-way table," such as the one given for the 20,000 die rolls, is more or less the only possible way of describing the data. For each category, this table can show either its "absolute frequency" (the number of occurrences of the category) or its "relative frequency" (the number of occurrences of the category divided by the total number of occurrences of all categories). If we are interested in the relationship or association between two or more ordinal and/or nominal variables, two-way and multi-way frequency tables, respectively, (often referred to as "contingency tables") should suffice to informally show trends in the data and are again essentially the only way of describing the data. As was the case for one-way tables, these tables can contain either absolute or relative frequencies. As an example of how a

[2] However, in the remainder of this course, we will assume that the data set does not contain any missing values.

two-way table can reveal the association between two variables, consider the following absolute frequency table relating number of marriages and education for a sample of 1,436 married women listed in Who's Who in 1949.

Education	Married Once	Married more than once	Total
College	550	61	611
No College	681	144	825
Total	1,231	205	1,436

First, note that, of the women who went to college, 10% had been married more than once; for those without college education, this figure is 17%. Also, for women married more than once, 30% had a college education; for women married only once, that figure is 45%. The tables shows an association between the two categorical variables in this data set; specifically it shows that, for these data, having a college education increases the chances of being married only once.

As was discussed in Section V above, one-way and multi-way frequency tables can also be used to describe continuous variables if the possible values of these variables are divided into interval classes or bins and the number (or frequency) of observations in each class is then counted.

(VI.b.) Numerical EDA Methods

(VI.b.i.) Univariate Methods

Sample Quantiles and Deciles

"Sample quantiles"[3] can be used to describe a variable of either the categorical or the continuous variety. The "α-th sample quantile," denoted η(α), is the smallest value such that $(100 \times \alpha)\%$ of the observations of that variable take values which are less or equal than η(α). For instance, five percent of the observed values for a given variable are smaller than its 5th sample quantile.

One particular set of sample quantiles that is used frequently is the "sample deciles," which are the 9 values that divide the variable's observed range into 10 intervals, each of which has the same number of observations (i.e., 10% of the total number). As an example, sample deciles are often used to compare two different income distributions. Another important set of quantiles is the "sample quartiles," a set of values that divides the observed range of a variable into four intervals, each containing 25% of the observations. The sample quartiles are denoted by Q_1, Q_2 and Q_3 and are referred to as the "lower quartile," "the sample median," and the "upper quartile," respectively. Frequently, the sample quartiles are combined with the sample minimum and the sample

[3] Although in this course we will use the terms "sample quantiles," "sample mean," etc., in order to distinguish them from their population counterparts, in many cases, the adjective "sample" is omitted when referring to these quantities.

maximum (i.e., respectively, the smallest and largest of the n observed values in the data set for the variable) in order to produce a "five number summary." In addition, the quantiles are often used to calculate a variable's "sample inter-quartile range (IQR)," which is defined as the quantity Q_3-Q_1. As an example, consider the following data, which refer to the time intervals between successive coal mine disasters (i.e. accidents involving ten or more men killed) in Britain between 1851 and 1962. The time interval covers 40,550 days, during which period there were 190 disasters. The five number summary for this 'inter-disaster interval' variable is:

Min	Q1	Median	Q3	Max
0	37.75	113.5	270	2,366

Looking at this five number summary, we might have some suspicions about the minimum (0); however, it happens that there were two accidents in the same day once. Note that even though the data are recorded integers, the first two quartiles are non-integer values. This is because most statistical software packages use a convention of interpolating between observed values if the quartile does not correspond exactly to an observed value (e.g., the sample median is the average of the two middle values when n is even). For instance, if we have the data {1, 5, 8, 15}, the quartiles would be (4, 6.5 and 9.75). Lastly, we can calculate the IQR for this variable, which has a value of 232.25 = 270-37.75.

Location

This property is concerned with finding, for a given variable, the position of the value in the data set that best characterises it. The sample "median," "mean," and "mode" can all be used to describe this position. Here, we should note that although the mode has meaning for all six types of variables, calculation of the median only makes sense for variable types other than nominal, and calculation of the mean is only possible for variable types other than nominal and ordinal. In addition, we should point out that these measures of location are not particularly informative when the empirical distribution for a given variable isn't unimodal (i.e., when it has more than one 'hump').

The sample median can be calculated by ranking the n observed values of the variable of interest from smallest to largest; then, the median is the middle value in this ordered list if n is odd or the average of the two middle values in this list if n is even. The sample median indicates the 'centre' of the empirical distribution of a given variable in the sense that half of the values are smaller than or equal to the sample median and half of them are larger than or equal to it. Another measure of location is the sample mode, which is simply the value of the variable that appears with the highest frequency in the data set. The sample mode is not necessarily unique because two different values may occur with the same highest frequency; for this reason and others, the sample mode is not always a good indication of location. Lastly, the sample mean is probably the most widely used location measure. If we are interested in a variable called X, then the sample mean is denoted $\overline{X}$. $\overline{X}$ is calculated by summing the n observed values of X and then dividing that sum by n. Like the sample median, the sample mean also indicates the centre of the

distribution, but here in the sense of a centre of gravity: if the observations for a variable were put in numerical order along a bar that was then placed on a pivot, the sample mean would be the point along the bar where placement of the pivot allowed the bar to balance perfectly horizontally. Alternatively, the sample mean can be thought of as the average of the observed values of the given variable.

As for a comparison of the sample median and the sample mean, if the empirical distribution of the variable is symmetric with respect to the mean, then the median and the mean have the same value. For instance, let us consider a data set containing measurements of length of the forearm (in inches) for 140 adult males. For this data, the median and the mean are 18.8 and 18.802, reflecting the symmetric nature of the distribution. However, the sample mean and median are not coincident when the empirical distribution of the variable is asymmetrical. This is especially true when outliers are present for a variable, since the sample mean is greatly affected by outliers whereas the sample median is not. As an example of this, consider the coal mine data introduced above. For the 'inter-disaster interval' variable, the sample median and the mean are 113.5 and 213.4, respectively. This discrepancy occurs because the maximum value of 2,366 is an outlier in comparison to the rest of the coal mine data and therefore pulls the sample mean away from the sample median. This value does not affect the sample median because the median merely selects the middle observed value and thus extra-small and extra-large values to do enter into its calculation. For this reason, we would say that the median exhibits "robustness against outliers," which is often a desirable statistical property, especially in the case where outliers represent measurement or recording errors.

Spread (or dispersion)

The "spread" or "dispersion" of a variable measures the degree to which the observed values for that variable are concentrated around a location measure; a 'smaller' spread indicates that the observed values are more tightly clustered around the 'centre' of the variable's empirical distribution. Measures of spread include the "sample range," the IQR, the "sample variance," the "sample standard deviation," and the "sample coefficient of variation." As was true for the sample mean, it is not possible to calculate these quantities for nominal or ordinal variables. In addition, the sample coefficient of variation is valid only for the discrete and ratio types of variables.

The simplest measure of spread is the sample range, which is defined as the difference between the sample maximum and the sample minimum for a variable. However, since the sample range depends on only two observations, it is highly sensitive to outliers and, as a result, may not be a reliable indicator of spread. Another measure of spread is the previously defined IQR. Yet another measure of spread is the sample variance, which is often denoted by s^2 or by $\hat{\sigma}^2$. The sample variance is loosely defined as the average of

the squared differences between the observations and the sample mean (i.e., the average of the squared deviations of the observations from the sample mean). Specifically, [4]

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 .$$

[Before proceeding, we should point out that the sample variance is not always calculated by taking the traditional average of the squared deviations of the observed values from the mean: frequently, in practice, the sum of these deviations is divided by *n-1* rather than by *n* as would be done for a traditional average. Theoretical reasons for doing so exist, and they will be explained in Lecture 4.] As a result of the way in which the sample variance is defined, it is always non-negative and will be 0 only if all of the observed values for a variable are identical (i.e., there is no variation). The sample variance is expressed in squared units, which can make it a difficult quantity to intuitively interpret. For this reason, it is common to use the (positive) square root of the sample variance as a measure of the variable's spread; this value is called the sample standard deviation and is frequently denoted by s. Here, we should note that the sample standard deviation is <u>not</u> the average of the differences between each observation and the mean. Obviously, the sample range, variance, and standard deviation depend on the units in which a variable is measured. As a result, a variable that takes on values in the 1,000s and whose observed values are tightly clustered about its mean could have a larger sample standard deviation than a variable that takes on only positive values less than 50 and whose values are very spread out relative to the mean, even though one might want to say that the latter variable has greater spread than the former. However, if we desire to compare two or more dispersions, we can circumvent this difficulty by employing the sample coefficient of variation (CV), which uses the absolute sizes of the means to adjust for the above phenomenon. The sample coefficient of variation is calculated using the formula

$$CV = s \Big/ \left|\overline{X}\right| .$$

This coefficient has no units, which allows us to use it for comparing dispersions of variables measured in different units. As was stated above, note that the sample coefficient of variation is only valid for discrete and ratio variables; obviously, it is not possible to calculate this quantity for ordinal and nominal variables, and, for non-ratio continuous variables, using the mean as an adjustment factor has little validity because of the arbitrary position of zero on the variable's scale. Also, we should note that if the sample mean for a variable happens to be very near 0, then the value of the coefficient of variation might be artificially inflated and therefore suggest a greater degree of dispersion that is actually present.

As an example of a comparison between the sample standard deviation and coefficient of variation, suppose that we would like to compare the variability of house prices with the variability of car prices and that we have the following values:

[4] For those unfamiliar with sigma (Σ) notation (also known as summation notation), consider the following basic example: $\sum_{i=1}^{n} X_i$ means "sum all the X_is (the quantity to the right of the summation sign) from *i=1* (*1* is the lower index of the sign) to *i=n* (*n* is the upper index of the sign)." In other words, $\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_{n-1} + X_n$. Similarly, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_{n-1} - \overline{X})^2 + (X_n - \overline{X})^2$.

	Houses	Cars
mean	£120,000	£12,000
standard deviation	£2,000	£2,000

Although the standard deviation is £2,000 for both houses and cars, the dispersion of prices around the mean is clearly more spread out for cars than for houses. This fact is revealed by the sample coefficients of variations, which are 1/60 and 1/6 for houses and cars, respectively, reflecting a greater variability for the latter. This would not be apparent if we had compared the standard deviations only.

Of the measures of dispersion that we have reviewed, only the IQR is relatively robust to outliers. Because they involve squaring the deviations of observations from the sample mean, the sample variance, standard deviation, and coefficient of variation are all extremely sensitive to outliers.

Skewness

"Skewness" refers to deviations from symmetry with respect to a location measure. The quantity, often referred to as b_1, that is commonly used as a measure of asymmetry is calculated by dividing the average of the cubes of the differences between the observations and their sample mean by the cube of the sample standard deviation. Specifically,

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}.$$

The resulting quantity, b_1, is unit-free. If the variable's empirical distribution is symmetric around its sample mean, then b_1 has a value of 0. Positive values of b_1 indicate that the variable is "right-skewed" (i.e., there is a longer or fatter tail for values larger than the mean); a negative value of b_1 provides evidence of a longer or fatter tail for values smaller than the mean (i.e., the variable is "left-skewed").

Kurtosis

"Kurtosis" denotes the degree of 'peakedness' of the distribution, often as compared to a Normal (Gaussian) distribution. The "coefficient of kurtosis," usually referred as b_2, is calculated by dividing the average of the fourth power of the differences between the observations and their sample mean by the fourth power of the sample standard deviation. Specifically,

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}.$$

As a result of this definition, b_2 is always non-negative and is unit-free. This coefficient takes the value of 3 for the normal distribution, which is described as "mesokurtic." A value of this coefficient that is smaller than 3 indicates a distribution that is "platykurtic", i.e., a distribution that is fatter- and/or shorter-tailed than the normal distribution or,

identically, less peaked than the normal distribution. For the flattest (least-peaked) of all distributions, the uniform distribution (i.e., a distribution in which all possible values occur with the same frequency), b_2 takes a value of 1.8. On the other hand, if b_2 has a value larger then three, then this indicates a distribution that is "leptokurtic" (i.e., a distribution that has longer and/or slimmer tails than the normal distribution and is thus more peaked). Before proceeding, we should point out that some packages standardise the coefficient of kurtosis by subtracting it from 3, so that it takes positive values for flatter (platykurtic) distributions, 0 for the normal (mesokurtic), and negative values for peaked (leptokurtic) distributions.

As an example of b_1 and b_2, we turn to the forearm lengths data, for which the values of b_1 and b_2 are (-0.108, 2.53); these values are quite near 0 and 3, indicating a symmetric, mesokurtic distribution. In contrast, the coal data yield values of (3.539, 18.99), which correspond to a positive-skewed, leptokurtic distribution.

As a last point of interest, we should note that both b_1 and b_2 are highly sensitive to outliers.

(VI.b.ii.) Multivariate Methods

Measures of Association Between Two Variables

If we are interested in assessing the degree of association between two numerical variables, "Pearson's correlation coefficient," "Spearman's rank correlation coefficient," and/or "Kendall's rank correlation coefficient" can give us a good idea of the strength of the relationship between two such variables. However, these measures of association are not appropriate for categorical variables; if we desire to investigate the relationship between two or more categorical variables, a multi-way frequency table can be very revealing, as was discussed in Section VI.b above.

For numerical variables X and Y, Pearson's correlation coefficient provides a measure of <u>linear</u> association. To calculate this coefficient, one must first calculate the "sample covariance" for these two variables, which is defined as

$$\text{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right),$$[5]

where X_i and Y_i are the observations of variables X and Y for unit i, and $\overline{X}$ and $\overline{Y}$ are the sample means for those variables. If most of the X observations that are above the mean for X are paired with Y observations that are above (below) the mean for Y, and if the observations below the mean for X correspond to observations below (above) the mean for Y, then most of the terms in the sum defining the covariance will be positive

[5] For those unfamiliar with sigma (summation) notation,
$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \left(X_1 - \overline{X}\right)\left(Y_1 - \overline{Y}\right) + \left(X_2 - \overline{X}\right)\left(Y_2 - \overline{Y}\right) + \cdots + \left(X_{n-1} - \overline{X}\right)\left(Y_{n-1} - \overline{Y}\right) + \left(X_n - \overline{X}\right)\left(Y_n - \overline{Y}\right)$.

(negative), resulting in a positive (negative) covariance. Obviously, just as the sample variance and standard deviation are affected by the units in which a variable is measured, the sample covariance will also reflect the absolute size of the units used to measure the two variables of interest. However, Pearson's correlation coefficient, *r*, removes this dependence on the unit of measurement; it does so because the correlation coefficient is calculated by dividing the sample covariance by the product of the sample standard deviations of *X* and *Y*. The sample correlation coefficient is thus unit-free and takes on only values between -1 and 1. If *r* is approximately 0, then there is no evidence of linear correlation. On the other hand, a value of 1 indicates a perfect positive linear association, and a value of –1 indicates a perfect negative linear association. It is essential to remember that Pearson's correlation coefficient assesses only the linear association of two variables and is not a measure of non-linear relationships. For instance, two variables that clearly have a very strong curved relationship, as seen in a scatterplot, might have a sample correlation of 0. Thus, the sample correlation of two variables should be viewed with caution. As an example of why, R.A. Fisher presented 4 very different looking scatterplots for two variables, all of which had the same sample correlation; obviously, for these cases, the sample correlation does not reflect the different relationships between variables.

Pearson's correlation is not at all robust to outliers, which is not surprising given that it incorporates sample standard deviations, which are not robust either. In order to find a more robust measure of association, one could order the values of each variable from smallest to largest and then rank them from 1 (smallest) to *n* (largest). (If a variable has two identical values, the usual procedure for assigning ranks is to assign to both values the average of the two ranks that should be assigned to those values.) After ranking the values for both variables, one could calculate Pearson's *r* using the two variables' ranks rather than their original observed values. The resulting correlation coefficient is known as Spearman's rank correlation coefficient and is robust to outliers. Another robust correlation coefficient is Kendall's, which simplifies the information contained in the two variables even more than Spearman's does. Kendall's coefficient involves examining all possible pairs of observations, where an observation is defined as both the *X*-value and the *Y*-value for a given unit. A pair of observations, such as (X_i, Y_i) and (X_j, Y_j), is termed 'concordant' if and only if (1) $X_i > Y_i$ & $X_j > Y_j$ or if (2) $X_i < Y_i$ & $X_j < Y_j$; otherwise, the pair is 'discordant.' Each possible pair of observations is assigned a score of 1 if case (1) holds, a score of -1 if case (2) holds, and a score of 0 if it is discordant. Kendall's rank correlation coefficient is then the sum of these scores, standardised so that it falls within -1 and 1.

If we denote any of these three correlation coefficients by *c*, then we have the following properties:

a) *c* does not depend on the units of measurement of *X* and *Y*.
b) *c(X,Y)* = *c(Y,X)* (this property is known as "symmetry in the arguments").
c) $-1 \le c(X,Y) \le 1$
d) If *c(X,Y)* = ± 1, then the data lie exactly on a straight line.
e) *c(X,Y)* is an adequate measure of association only for linear relationships.

(VI.c.) Graphical EDA methods

Graphical methods make it very easy to discover trends and patterns in a data set; in addition, some of these methods are particularly good at revealing outliers, or departures from these trends.

(VI.c.i.) Univariate Methods

Frequency Plots and True Histograms

In Section VI.b, we noted that it is possible to calculate absolute and/or relative category frequencies for a variable of the categorical variety and, if the possible continuum of values for the variable is divided into intervals, for a variable of the numerical variety. Instead of presenting these frequencies in tabular form, we could use them to plot a frequency chart for the data. In the following graph, we plot this type of chart for the die roll data using the relative frequencies of the faces.

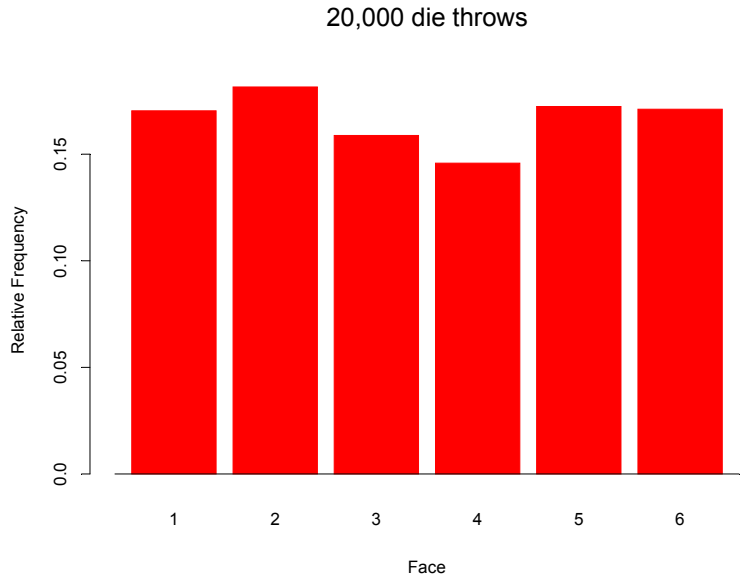

Figure 1: *Frequencies of 20,000 die throws*

In order to see an example of a frequency plot using absolute, rather than relative frequencies, consider the following table containing data from 144 horse races on a circular track. More specifically, this table contains the number of wins for each starting position, where Position 1 is closest to the rail on the inside of track.

Position	1	2	3	4	5	6	7	8
Num. of wins	29	19	18	25	17	10	15	11

The following graph shows the absolute frequencies for each starting lane.

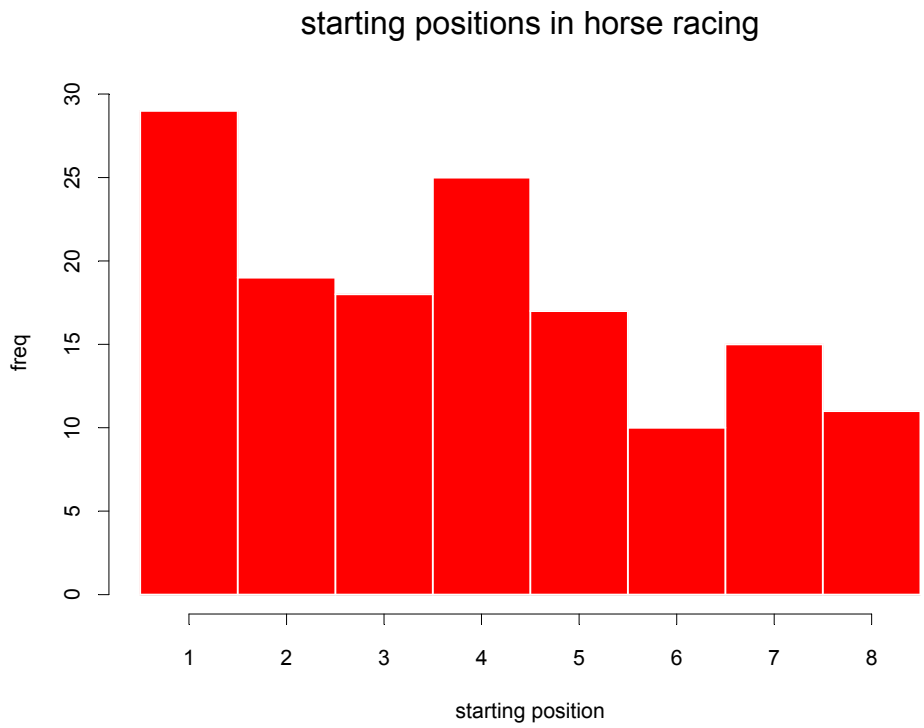

Figure 2: *Histogram of starting positions of winners in horse racing*

These absolute and relative frequency plots are often referred to as "histograms." According to this definition, a histogram is a bar graph in which each bar corresponds to a category created by grouping the variable's values into intervals, classes, or "bins" (unless, of course, the variable is already categorical), and where each bar's height (along the Y-axis) is proportional to the (absolute or relative) frequency of its corresponding class. Here, it is important to distinguish between a histogram or frequency plot and a "true histogram." The latter is similar to a frequency plot, except that in a true histogram, the *area* of a bar, rather than the *height* of a bar, is proportional to the frequency of the interval class to which it corresponds. More specifically, in a true histogram, the area of each bar is equal to the relative frequency of the interval class to which it corresponds. As a result of the way in which a true histogram is defined, the total area of all its bars is 1; this property makes a true histogram a candidate for an estimate of the true population density (i.e., distribution) of a variable, as we will see in Lecture 2. In a strict sense, a frequency plot is identical to a true histogram only when relative frequencies and bins of equal length are used for the frequency plot. However, since the use of absolute vs. relative frequencies affects only the Y-axis scale and not the appearance of the plot, for our purposes, the distinction between frequency plots (histograms) and true histograms will only matter in cases where the interval classes are not all of equal length.

For both frequency plots and true histograms, having a greater number of bins corresponds to having a smaller interval length (assuming that all intervals have the same length). The number of bins used can greatly affect the appearance of both types of plots.

As an example of this phenomenon, consider the following four absolute frequency plots (each with a different number of bins) for the forearm length data:

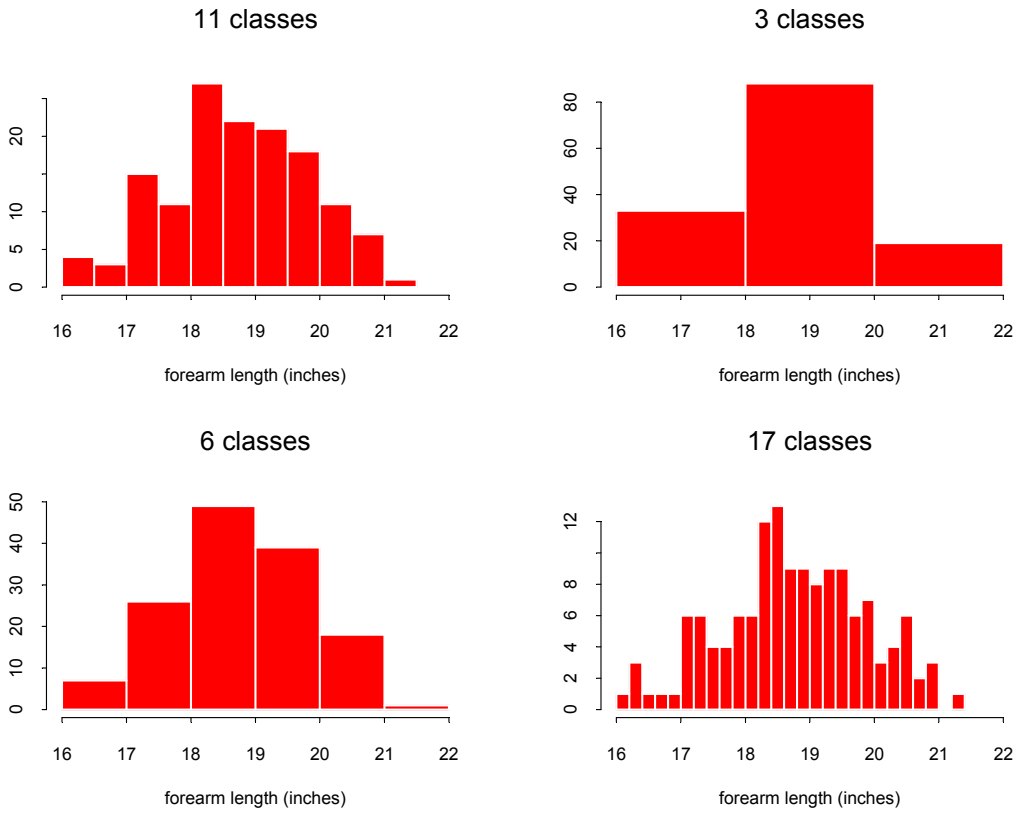

Figure 3: *Histograms for forearm lengths*

Looking at the above plots, we should note that, as a result of using different numbers of bins, the absolute frequencies marked on the Y-axis change from plot to plot. More importantly, we notice that the more classes we use, the more details of the data we can perceive, but the more empty classes there are. If the number of bins is too small, or, correspondingly, the length of the categories is too great, observations that perhaps should be distinguished as members of different classes may be clumped together and, thus, particular features of the data may be lost. On the other hand, using too many intervals, or, correspondingly, using category widths that are too narrow, may allow distracting details to be represented and therefore obscure the overall picture. Thus, in choosing the number of bins or, identically, the width of the bins, we face a trade-off, and, as a result, we should try to find a good balance between these two extremes. Unfortunately, there is no universally accepted rule for choosing the number of bars in a histogram. For example, S-Plus chooses the 'optimal' number of bins to be proportional to the logarithm (in base 2) of the number of observations. However, other statistical packages may employ different rules, and the data analyst should make sure that he/she is familiar with the rule employed in his/her package of choice.

In addition to being affected by the number of bins (of equal width) that are used, the appearance of frequency plots and true histograms can also be influenced by where the breakpoints between interval classes are located and by whether a value that occurs at a category breakpoint is considered to belong to the right-hand bin or to the left-hand bin. Also, since it is not required that all intervals classes be of equal length, using varying interval class lengths can affect the appearance of frequency plots and true histograms. For some data sets, the use of different lengths for different interval classes can give a better overall picture of the data. To illustrate this last fact, we first present an absolute frequency plot with interval classes of equal length for the coal mine data.

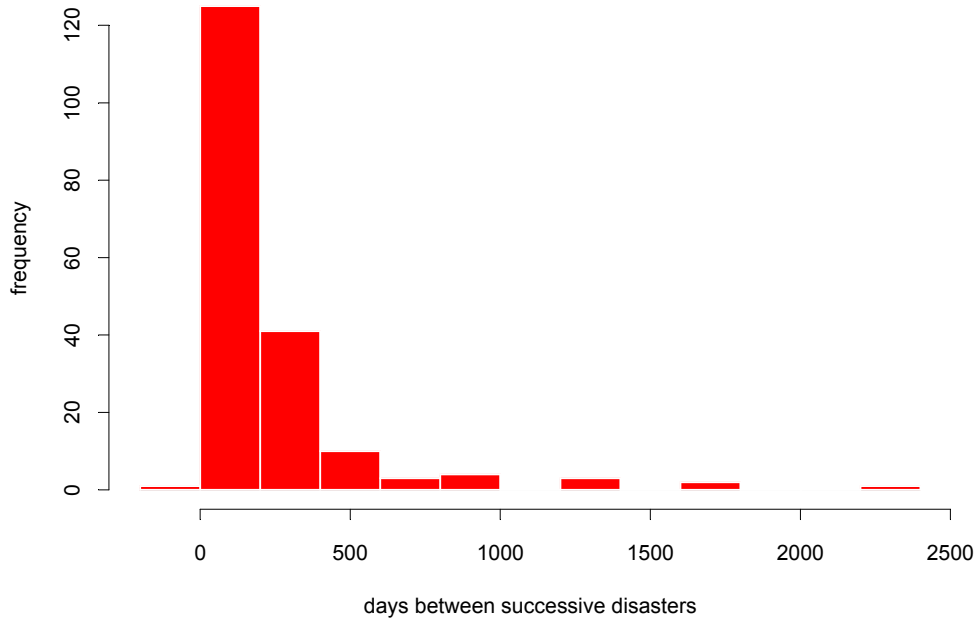

Figure 4: *Histogram for coal mine disasters data*

Because most of the observations occur between 0 and 500, with the exception of a few values that are very far outside this range, using equal interval class lengths results in a loss of some important detail for values in the high frequency (low value) area as well as a waste of space in representing the high value classes with low frequencies. In this case, it would certainly be convenient to have interval classes of different lengths, as is shown in the following two plots.

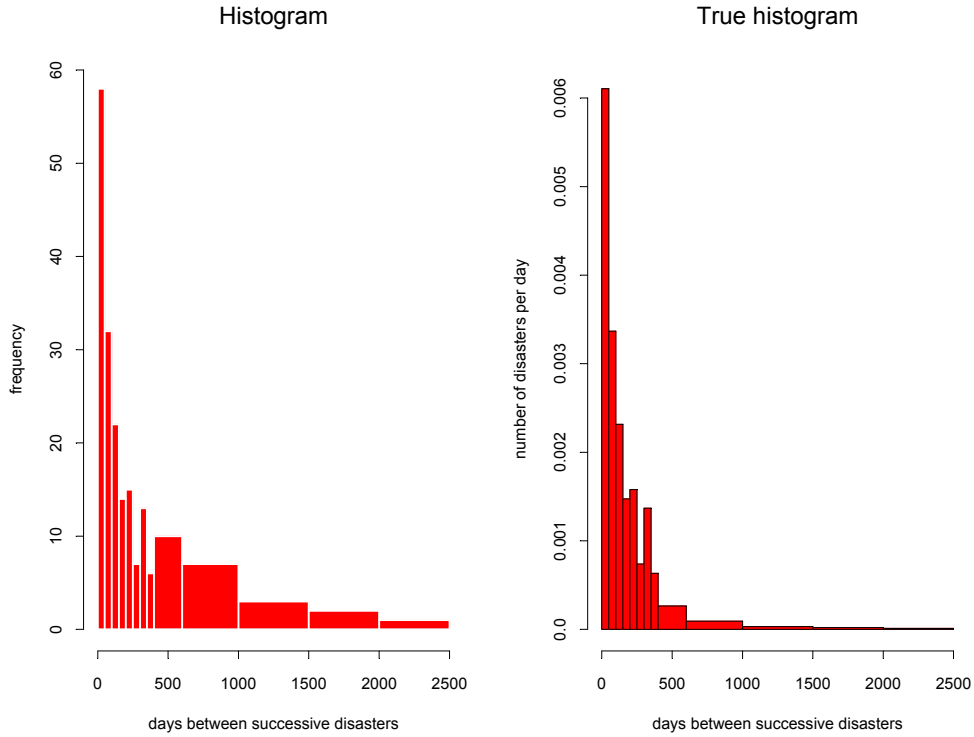

Figure 5: *Histogram and True histogram for coal mine disasters data*

The plot on the left-hand side of Figure 5 is an absolute frequency plot with interval classes of variable length. Note that this representation is distorted because the longer class intervals that are used for large day values have greater heights even though they have smaller frequencies *per day*. This distortion is compensated for in the true histogram that appears on the right-hand side. This compensation occurs because in a true histogram, the areas, and not the heights, are equal to the class frequencies, which means that for two interval classes with the same number of events, the longer class interval has a lower height. In other words, in the true histogram, a fair comparison amongst interval classes is allowed because the Y-axis in this graph represents the number of disasters per day in each class.

Before proceeding, we should note that frequency plots and histograms are particularly useful for getting an idea of the distribution of a variable, and, in particular, where its centre is located, how spread out it is, whether it is symmetric, right-, or left-skewed, and how fat and long its tails are.

Boxplots

Another useful graphical device for describing interval and numerical variables (but not ordinal or nominal variables) is the "boxplot," which is sometimes known as a "box-and-whiskers plot." This plot is based on the five number summary and is particularly useful for identifying outliers and extreme outliers and for comparing the distributions of variables within two or more classes. The ends of the box (the 'hinges') are the lower and upper sample quartiles, and thus, the length of the box is the variable's IQR; further, the sample median for the variable is marked by a line inside the box. The lines extending

from the box (the 'whiskers') extend up to the smallest and largest observation within the interval $\left(Q_1 - 1.5\ \text{IQR},\ Q_3 + 1.5\ \text{IQR}\right)$. Points that fall within the interval $\left(Q_1 - 3\ \text{IQR}, Q_1 - 1.5\ \text{IQR}\right)$ are designated as "negative outliers," and points that fall in the interval $\left(Q_3 + 1.5\ \text{IQR}, Q_3 + 3\ \text{IQR}\right)$ are designated as "positive outliers." Those points located outside the interval $\left(Q_1 - 3\ \text{IQR},\ Q_3 + 3\ \text{IQR}\right)$ are considered to be "extreme outliers."

As an illustration, consider the boxplot for the forearms data:

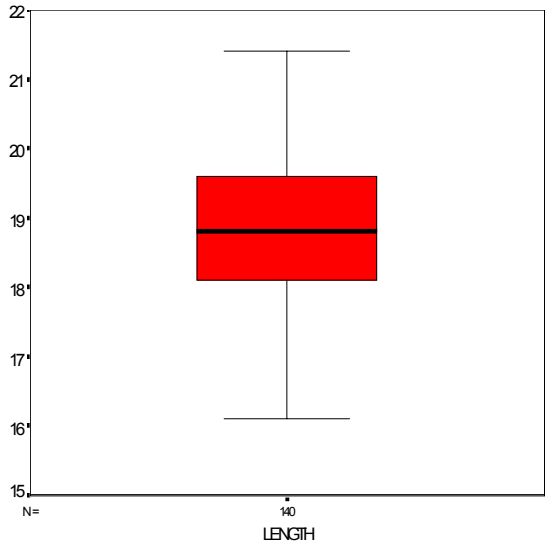

Figure 6: *Boxplot for forearm lengths data*

There are no outliers in this data set. In addition, we can tell that the forearms length variable is roughly symmetric since the sample median line is more or less at the middle of the box and since the two whiskers are equivalent in length. However, it is not possible to tell whether the distribution of the forearms variable is platy-, meso-, or leptokurtic from a boxplot.

As another example, the boxplot for the coal mine disasters data appears in the following graph. The outliers appear as circles, and extreme outliers are marked as asterisks. The number next to each of these symbols indicates the observation number (within the data set) to which the point corresponds.

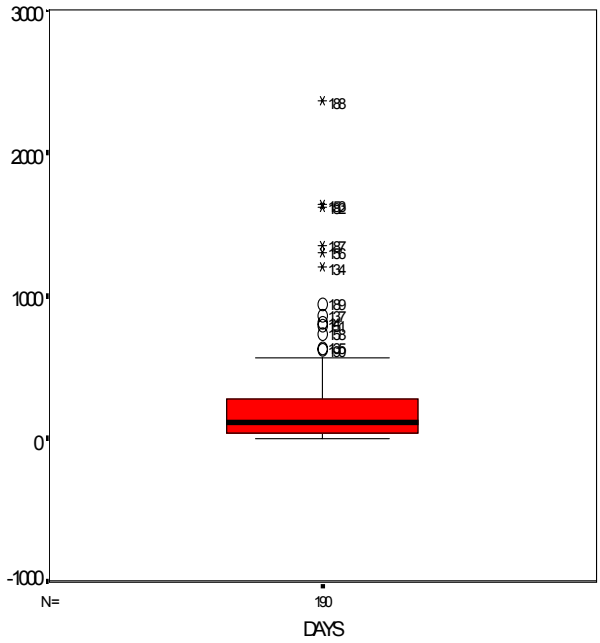

Figure 7: *Boxplot for coal mine disasters data*

Looking at the above plot, we see immediately that the coal mine data set has a large number of positive outliers (i.e., values that are much larger than the majority of points in the data set). This plot demonstrates how easy a boxplot makes it to identify outliers. As was stated before, outliers may represent measurement or recording errors, in which case, the data analyst should seriously consider either removing these observation from the data set or, if possible, correcting them. Alternatively, outliers may merely be anomalous members of the population. In this case, the outliers may be of particular interest for the study; for instance, in the above example, the analyst may want to investigate the circumstances surrounding the 2,366 day period with no mining disasters.

(VI.c.ii.) Multivariate Methods

Boxplots

The boxplot, as defined above, can be used for bivariate analysis in the specific case where one desires to investigate the association between a categorical variable and a non-ordinal and non-nominal variable (i.e., an interval or numerical variable). In this specific case, boxplots make it very easy to compare the distributions of the second variable within each of two or more classes (levels) of the first variable. As an example, the following graph shows the boxplots for the calcium concentration in the drinking water supply for large English and Welsh towns that are north of Derby and for towns that are south of Derby. Here, 'calcium concentration' is a ratio continuous variable and 'location relative to Derby' is a nominal categorical variable with two levels (binary). In the data set, which contains 61 towns in total, calcium is measured in parts per million, and the higher the calcium concentration, the harder the water. The measurement for each town corresponds to its average calcium concentration over the years 1958-1964. Looking at these plots, we

instantly perceive a substantial difference in the positions of the two medians. In addition, the distribution of the calcium concentration averages for towns north of Derby is considerably more asymmetrical than the distribution for the towns south of Derby. This can be seen by the median not being in the centre of the box, as well as the whiskers not having the same length, for the Northern towns, but not for the Southern towns. Lastly, there seems to be only one outlier in the data.

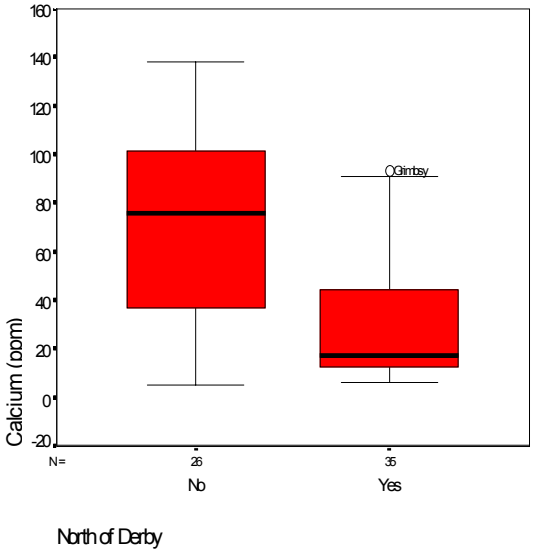

Figure 8: *Boxplots for concentration of calcium in large towns data*

Alternatively, it is possible to generalise the univariate boxplot defined in the previous section in order to use it for the bivariate analysis of two numerical variables. These generalised boxplots are useful tools for examining the bivariate distribution of two such variables. The following graph shows an example of a generalised two-variable boxplot for a data set that comprises the body and brain weights of 28 animals, some of them extinct. For reasons to be discussed later, both variables are represented on a logarithmic scale. In the plot below, the inner ellipse corresponds to the box in a univariate boxplot, with the median centre (a bivariate analogue of the median that is the pair of medians for the two variables) of the data being marked by a cross. The outer ellipse below corresponds to the whiskers in a univariate boxplot. The graph shows that there are four outliers (the three extinct species and the human), which would not be apparent from a scatterplot of the two variables.

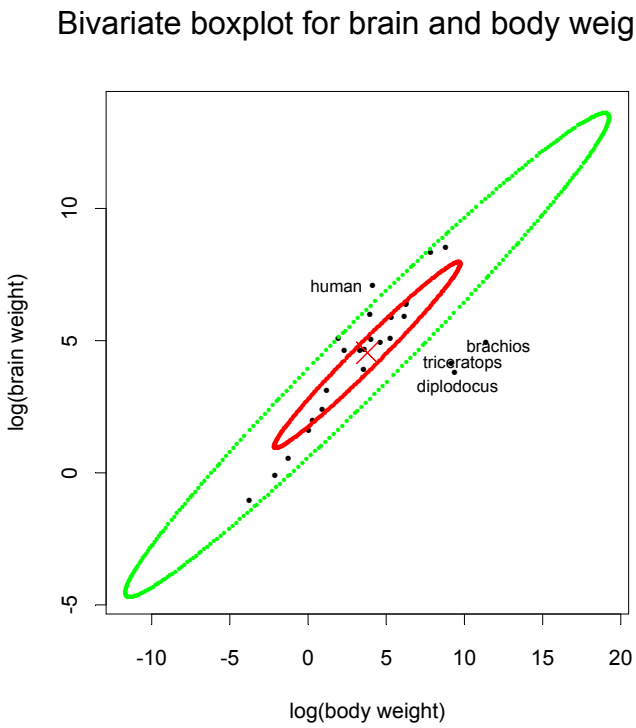

Figure 9: *Bivariate boxplot for body and brain data*

Two-variable Histograms

If one is interested in examining the association between two variables of any type, the variables' (absolute or relative) joint frequencies, which correspond to the classes that result from crossing the natural or constructed categories of the two variables (Section VI.b), can be used to produce a two-variable histogram. In this type of histogram, the magnitude of the absolute or relative frequencies can be indicated either by using colour gradations in a 2-D plot or by using bar height in a 3-plot.

As an example, consider the observations made at the Old Faithful geyser in Yellowstone National Park, Wyoming, U.S.A. The variables of interest are, for each of 272 eruptions, the duration of the eruption and the time elapsed since the previous eruption, both of which are ratio variables. The sample range for each variable is divided into 9 bins of equal length, and the resulting two-variable histogram appears in Figure 9. In this plot, the darkness of each crossed class is proportional to the frequency for that class, and the white points are the actual data points that were used to create the histogram.

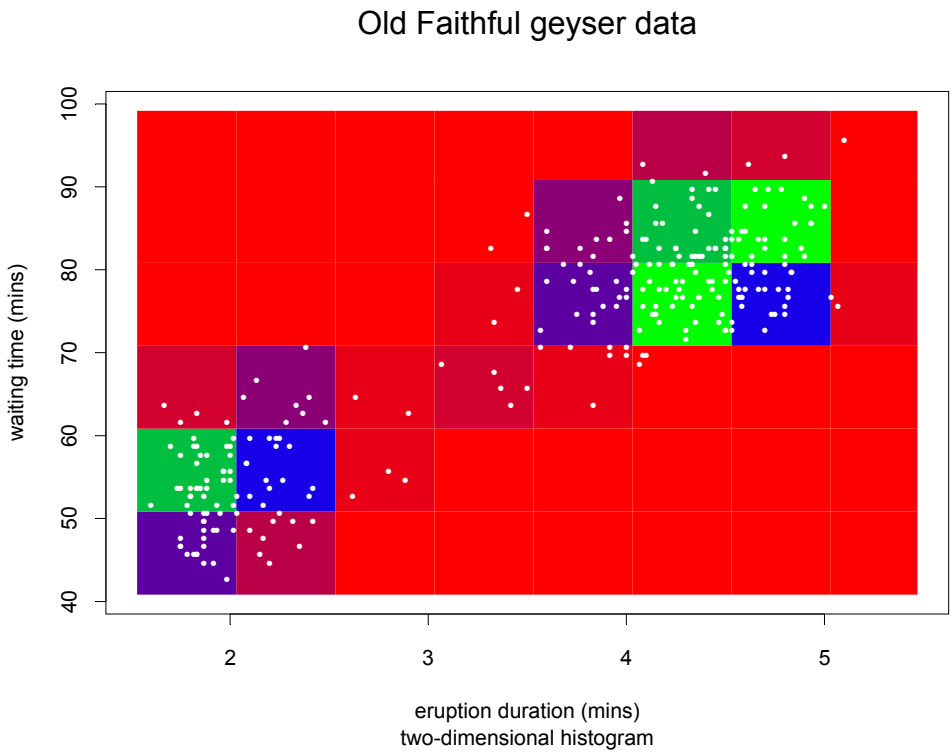

Figure 10: *Two-dimensional histogram of Old Faithful geyser data*

Scatterplots

If the data analyst wants to explore the relationship between two variables that are both of the numerical variety, then a scatterplot can be used. Further, if a data analyst wants to explore the relationship between two numerical variables and one categorical variable, he/she could employ a scatterplot in which different symbols indicate the various levels of the categorical variable. For either of these varieties of scatterplots, the analyst should take care to label the axes clearly, and, if necessary, to provide a legend that states the categorical variable level to which each symbol corresponds.

As an example of this latter variety of scatter plot, let us return to the calcium concentration data set, which already contained one nominal and one ratio variable, and introduce another ratio variable: the average male mortality rate for the same towns, where the mortality rate average for a town corresponds to the same years as the calcium concentration average. In this case, a scatterplot provides interesting insight into the relationship between mortality, water hardness, and geographical location:

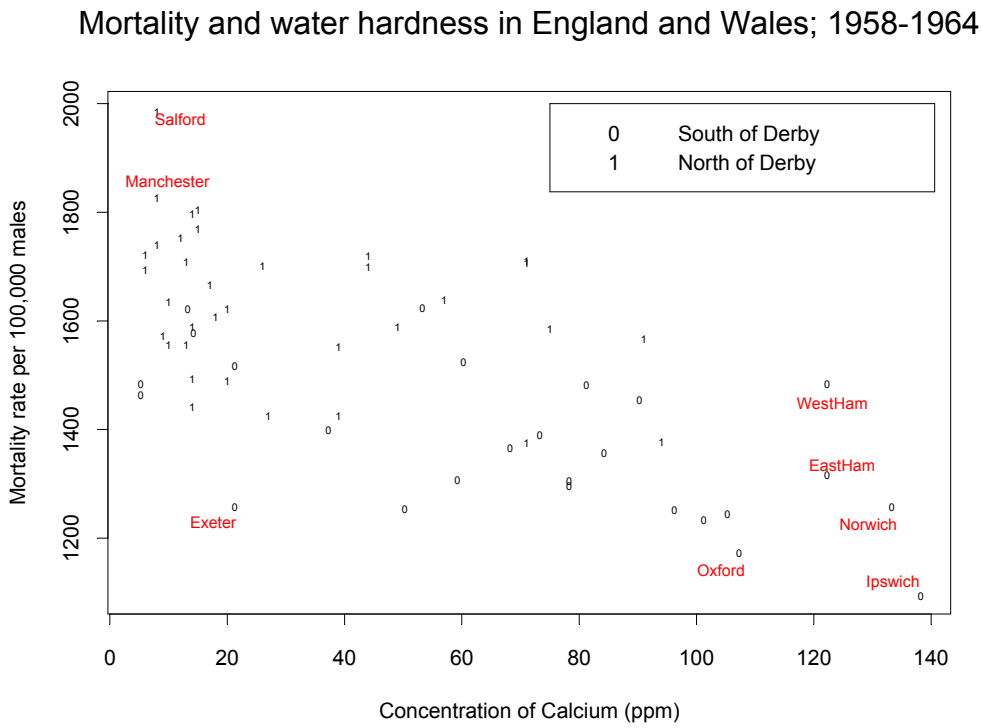

Figure 11: *Scatterplot of mortality and water hardness*

From this graph, we see that there seems to be a geographical factor in the relationship between mortality and concentration of calcium, with Southern towns having generally lower mortality as well as harder water. Note, however, that an association between two variables, such as mortality and water hardness, does not necessarily imply causality. For example, here, we cannot say at this stage that lower mortality is caused by higher levels of calcium in drinking water; in fact, such a proposition seems somewhat silly. So far, the graph simply states that there is an association between these two variables that is worth studying with more care. This issue of causality will be discussed in greater detail in Lecture 8.

(VII.) Transformations of Variables

As a final note before proceeding to a discussion of probability, we will address the issue of "transforming" continuous variables. "Transforming" a variable refers to applying the same mathematical function, such as $\ln(x)$, $\exp(x)$, or x^2, to all the observed values of a variable. Clearly, since the levels of categorical variables do not have a strict numerical interpretation, it is not possible or appropriate to apply mathematical transformations to these variables. For continuous variables, $\ln(x)$ is a particularly common transformation that is often applied to variables that only can or do take on positive values, such as height or earnings.

There are a number of reasons why one might want to transform a continuous variable. In some cases, there is theoretical motivation for such a transformation; as an example, if we

are examining the number of pounds in a continuously compounded savings account over time, we might want to take the natural logarithm of the pound amounts since we know that continuous compounding of a sum of money results in exponential growth. Alternatively, even if there is no theoretical motivation for transforming a variable, there may be practical reasons for doing so. For an example of a practical reason for transformation that occurs in EDA, we return to the animal body and brain weight data, which demonstrates that performing a transformation of the data can sometimes make it easier to explore the relationship between two variables. Two scatterplots for the brain weight and body weight variables are presented below: in the first plot, both variables are on their original scale of grams and kilograms, respectively, and in the second plot, both variables have been transformed to the (base 10) logarithmic scale (i.e., log(g) and log(kg)). Although there is a definite pattern of association between the two variables, as well as some deviations from this pattern (i.e., outliers), it is virtually impossible to detect either the pattern or the outliers if the data are plotted on their original scale. Using logarithmic transformations pulls in both axes so that we observe a clear linear relationship between the two variables, as well as three obvious outliers (the three extinct species).

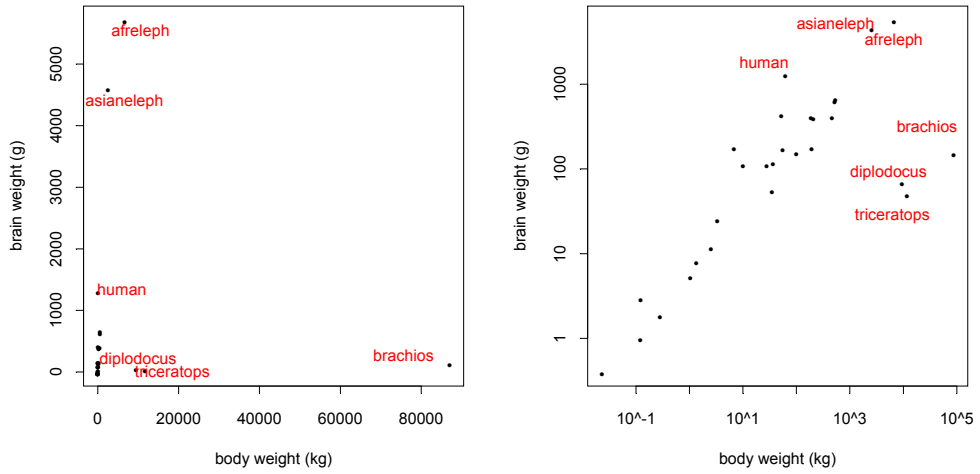

Figure 12: *Use of logarithmic scales in scatterplots*

A final reason for transforming a continuous variable might be that doing so renders the assumptions required by statistical inference methods (used in the phase of statistical analysis following EDA) more reasonable. For instance, many methods of statistical inference assume that the variable of interest has an underlying (population) distribution that is normal. A normal variable can theoretically take on any value in the interval $(-\infty, \infty)$, which is obviously not the case for variables, such as income, that can only take on positive values. Thus, in this case, using a logarithmic transformation, which takes a

number in the interval $(0,\infty)$ to a different number in the interval $(-\infty,\infty)$, may result in a transformed variable for which the assumption of normality is more reasonable.

(VIII.) References

The following books on applied statistics comprise a list of references. The list is by no means exhaustive and is only a sample of recent texts.

AGRESTI, A AND B FINLAY (1997). *Statistical Methods for the Social Sciences, 3rd ed*. Prentice-Hall Inc., Upper Saddle River, New Jersey.

ALTMAN, DG (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London.

CHATFIELD, C (1995). *Problem Solving: A statistician's guide*. Chapman & Hall, London.

CLARK, WAV AND PL HOSKING (1986). *Statistical Methods for Geographers*. John Wiley & Sons, New York.

CLEVELAND, WS (1993). *Visualizing Data*. Hobart Press, Summit, New Jersey.

COCHRAN, WG (1977). *Sampling Techniques*. John Wiley & Sons, New York.

COX, DR AND EJ SNELL (1981). *Applied Statistics*. Chapman & Hall, London.

CRAMER, D (1998). *Fundamental Statistics for Social Research*. Routledge, London.

HAHN GJ AND SS SHAPIRO (1967, republished in 1994). *Statistical Methods in Engineering*. John Wiley & Sons, New York.

HAND, DJ AND BS EVERITT (EDS.) (1987). *The Statistical Consultant in Action*. Cambridge University Press.

HANKE, JE AND AG REITSCH (1994). *Understanding Business Statistics, 2nd ed*. Irwin Publishing co., Burr Hills, Illinois.

HANSEN, MH, WN HURWITZ, AND WG MADOW (1953, reprinted in 1993). *Sample Survey Methods and Theory (2 vols.)*. John Wiley & Sons, New York.

HOWELL, DC (1995). *Statistical Methods for Psychologists (3rd edition)*. PWS-Kent Publishing Co., Boston, Massachusetts.

KALBFLEISCH, JG (1986). *Probability and Statistical Inference (2 vols.)*. Springer-Verlag, New York.

MILLER, RG, B EFRON, BW BROWN, AND LE MOSES (EDS.) (1980). *A Biostatistics Casebook*. John Wiley & Sons, New York.

PURI, BK (1996). *Statistics in Practice: An illustrated guide to SPSS*. Arnold, London.

REES, DG (1995). *Essential Statistics (3rd edition)*. Chapman & Hall, London.

RICE, JA (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, California.

SNEDECOR, GW AND WG COCHRAN (1980). *Statistical Methods (7th edition)*. Iowa State University Press, Ames, Iowa.

VENABLES, WN AND BD RIPLEY (1997). *Modern Applied Statistics with SPLUS (2nd edition)*. Springer-Verlag, New York.

MCB (I-2000), *KNJ* (III-2001), *JIM* (III-2001)