Chapter 1

The Big Picture

Why experimental design matters.

Much of the progress in the sciences comes from performing experiments. These may be of either an exploratory or a confirmatory nature. Experimental evidence can be contrasted with evidence obtained from other sources such as observational studies, anecdotal evidence, or "from authority". This book focuses on design and analysis of experiments. While not denigrating the roles of anecdotal and observational evidence, the substantial benefits of experiments (discussed below) make them one of the cornerstones of science.

Contrary to popular thought, many of the most important parts of experimental design and analysis require little or no mathematics. In many instances this book will present concepts that have a firm underpinning in statistical mathematics, but the underlying details are not given here. The reader may refer to any of the many excellent textbooks of mathematical statistics listed in the appendix for those details.

This book presents the two main topics of experimental design and statistical analysis of experimental results in the context of the large concept of scientific learning. All concepts will be illustrated with realistic examples, although sometimes the general theory is explained first.

Scientific learning is always an iterative process, as represented in Figure 1.1. If we start at Current State of Knowledge, the next step is choosing a current theory to test or explore (or proposing a new theory). This step is often called "Constructing a Testable Hypothesis". Any hypothesis must allow for *different*

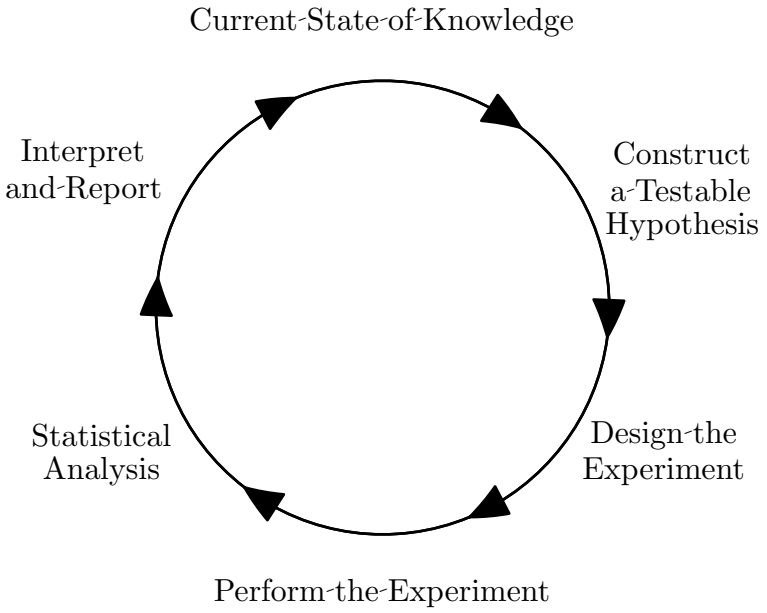

Figure 1.1: The circular flow of scientific learning

possible conclusions or it is pointless. For an exploratory goal, the different possible conclusions may be only vaguely specified. In contrast, much of statistical theory focuses on a specific, so-called "null hypothesis" (e.g., reaction time is not affected by background noise) which often represents "nothing interesting going on" usually in terms of some effect being exactly equal to zero, as opposed to a more general, "alternative hypothesis" (e.g., reaction time changes as the level of background noise changes), which encompasses any amount of change other than zero. The next step in the cycle is to "Design an Experiment", followed by "Perform the Experiment", "Perform Informal and Formal Statistical Analyses", and finally "Interpret and Report", which leads to possible modification of the "Current State of Knowledge".

Many parts of the "Design an Experiment" stage, as well as most parts of the "Statistical Analysis" and "Interpret and Report" stages, are common across many fields of science, while the other stages have many field-specific components. The focus of this book on the common stages is in no way meant to demean the importance of the other stages. You will learn the field-specific approaches in other courses, and the common topics here.

1.1 The importance of careful experimental design

Experimental design is a careful balancing of several features including "power", generalizability, various forms of "validity", practicality and cost. These concepts will be defined and discussed thoroughly in the next chapter. For now, you need to know that often an improvement in one of these features has a detrimental effect on other features. A thoughtful balancing of these features in advance will result in an experiment with the best chance of providing useful evidence to modify the current state of knowledge in a particular scientific field. On the other hand, it is unfortunate that many experiments are designed with avoidable flaws. It is only rarely in these circumstances that statistical analysis can rescue the experimenter. This is an example of the old maxim "an ounce of prevention is worth a pound of cure".

Our goal is always to actively design an experiment that has the best chance to produce meaningful, defensible evidence, rather than hoping that good statistical analysis may be able to correct for defects after the fact.

1.2 Overview of statistical analysis

Statistical analysis of experiments starts with graphical and non-graphical exploratory data analysis (EDA). EDA is useful for

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

- preliminary selection of appropriate models of the relationship between an outcome variable and one or more explanatory variables.

EDA always precedes formal (confirmatory) data analysis.

Most formal (confirmatory) statistical analyses are based on **models**. Statistical models are ideal, mathematical representations of observable characteristics. Models are best divided into two components. The structural component of the model (or **structural model**) specifies the relationships between explanatory variables and the mean (or other key feature) of the outcome variables. The "random" or "error" component of the model (or **error model**) characterizes the deviations of the individual observations from the mean. (Here, "error" does *not* indicate "mistake".) The two model components are also called "signal" and "noise" respectively. Statisticians realize that no mathematical models are perfect representations of the real world, but some are close enough to reality to be useful. A full description of a model should include all assumptions being made because statistical inference is impossible without assumptions, and sufficient deviation of reality from the assumptions will invalidate any statistical inferences.

A slightly different point of view says that models describe how the *distribution* of the outcome varies with changes in the explanatory variables.

Statistical models have both a structural component and a random component which describe means and the pattern of deviation from the mean, respectively.

A statistical test is always based on certain model assumptions about the population from which our sample comes. For example, a t-test includes the assumptions that the individual measurements are independent of each other, that the two groups being compared each have a Gaussian distribution, and that the standard deviations of the groups are equal. The farther the truth is from these assumptions, the more likely it is that the t-test will give a misleading result. We will need to learn methods for assessing the truth of the assumptions, and we need to learn how "robust" each test is to assumption violation, i.e., how far the assumptions can be "bent" before misleading conclusions are likely.

Understanding the assumptions behind every statistical analysis we learn is critical to judging whether or not the statistical conclusions are believable.

Statistical analyses can and should be framed and reported in different ways in different circumstances. But all statistical statements should at least include information about their level of uncertainty. The main reporting mechanisms you will learn about here are confidence intervals for unknown quantities and p-values and power estimates for specific hypotheses.

Here is an example of a situation where different ways of reporting give different amounts of useful information. Consider three different studies of the effects of a treatment on improvement on a memory test for which most people score between 60 and 80 points. First look at what we learn when the results are stated as 95% confidence intervals (full details of this concept are in later chapters) of $[-20, 40]$ points, $[-0.5, +0.5]$, and $[5, 7]$ points respectively. A statement that the first study showed a mean improvement of 10 points, the second of 0 points, and the third of 6 points (without accompanying information on uncertainty) is highly misleading! The third study lets us know that the treatment is almost certainly beneficial by a moderate amount, while from the first we conclude that the treatment may be quite strongly beneficial or strongly detrimental; we don't have enough information to draw a valid conclusion. And from the second study, we conclude that the effect is near zero. For these same three studies, the p-values might be, e.g., 0.35, 0.35 and 0.01 respectively. From just the p-values, we learn nothing about the magnitude or direction of any possible effects, and we cannot distinguish between the very different results of the first two studies. We only know that we have sufficient evidence to draw a conclusion that the effect is different from zero in the third study.

p-values are not the only way to express inferential conclusions, and they are insufficient or even misleading in some cases.

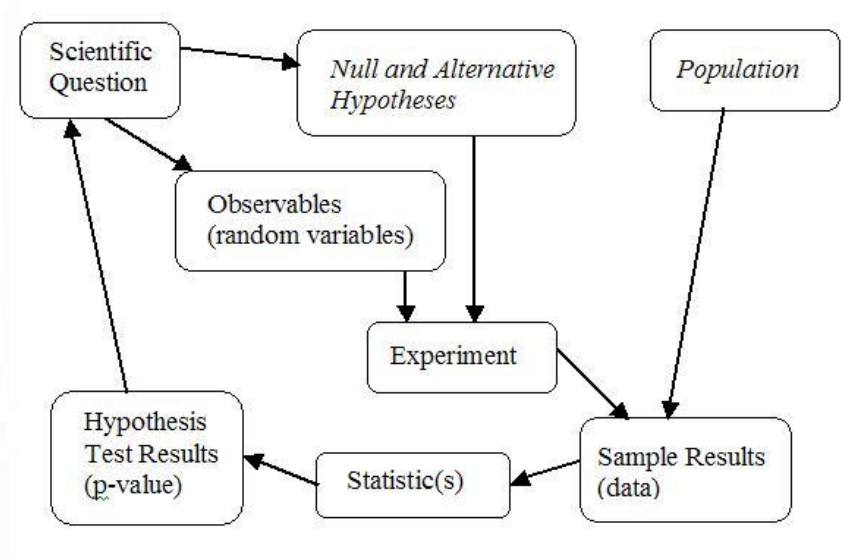

Figure 1.2: An oversimplified concept map.

1.3 What you should learn here

My expectation is that many of you, coming into the course, have a "concept-map" similar to figure 1.2. This is typical of what students remember from a first course in statistics.

By the end of the book and course you should learn many things. You should be able to speak and write clearly using the appropriate technical language of statistics and experimental design. You should know the definitions of the key terms and understand the sometimes-subtle differences between the meanings of these terms in the context of experimental design and analysis as opposed to their meanings in ordinary speech. You should understand a host of concepts and their interrelationships. These concepts form a "concept-map" such as the one in figure 1.3 that shows the relationships between many of the main concepts stressed in this course. The concepts and their relationships are the key to the practical use of statistics in the social and other sciences. As a bonus to the creation of your own concept map, you will find that these maps will stick with you much longer than individual facts.

By actively working with data, you will gain the experience that becomes "data-sense". This requires learning to use a specific statistical computer package. Many excellent packages exist and are suitable for this purpose. Examples here come

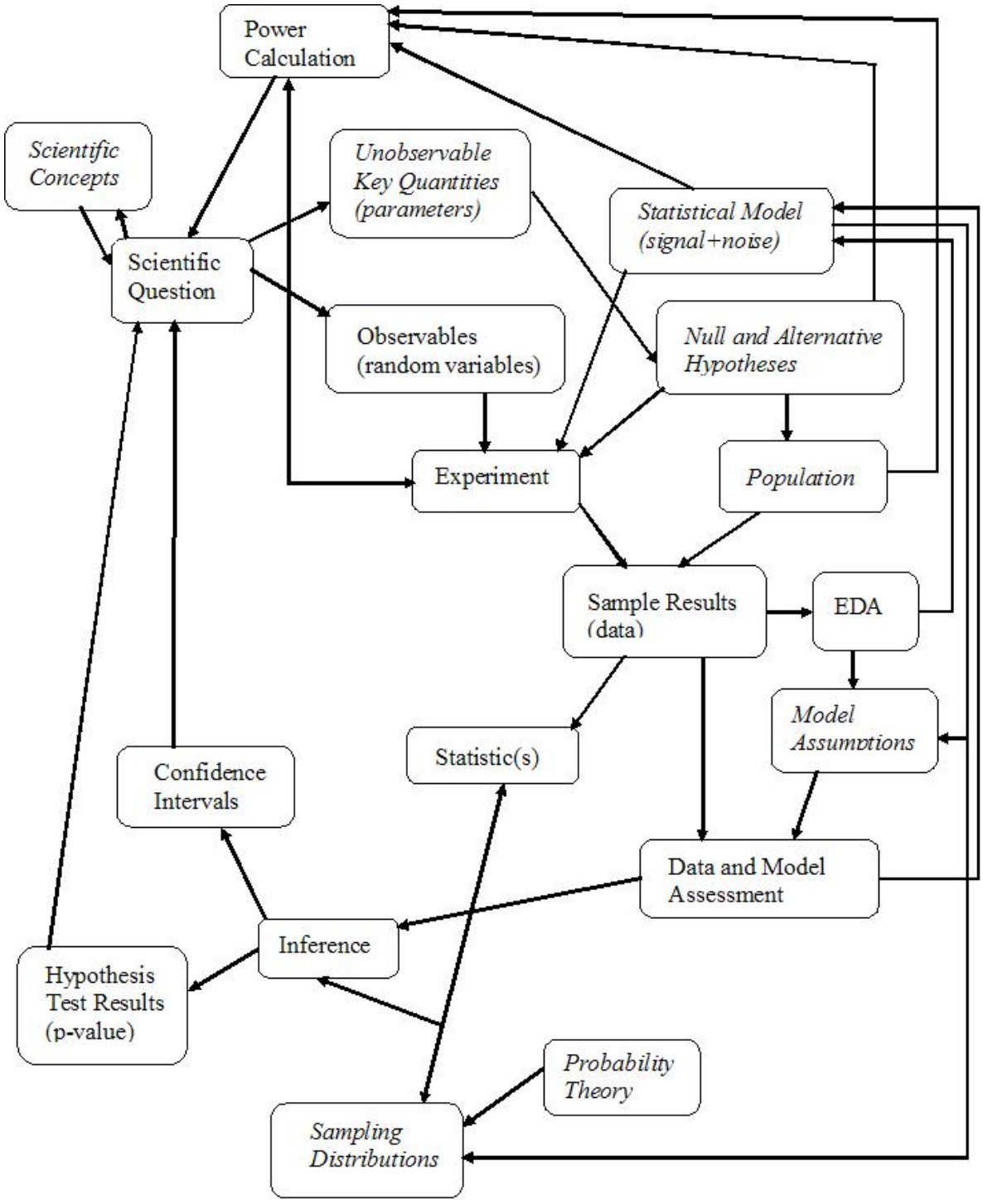

Figure 1.3: A reasonably complete concept map for this course.

from SPSS, but this is in no way an endorsement of SPSS over other packages.

You should be able to design an experiment and discuss the choices that can be made and their competing positive and negative effects on the quality and feasibility of the experiment. You should know some of the pitfalls of carrying out experiments. It is critical to learn how to perform exploratory data analysis, assess data quality, and consider data transformations. You should also learn how to choose and perform the most common statistical analyses. And you should be able to assess whether the assumptions of the analysis are appropriate for the given data. You should know how to consider and compare alternative models. Finally, you should be able to interpret and report your results correctly so that you can assess how your experimental results may have changed the state of knowledge in your field.

Chapter 2

Defining and Classifying Data Variables

The link from scientific concepts to data quantities.

A key component of design of experiments is **operationalization**, which is the formal procedure that links scientific concepts to data collection. Operationalizations define **measures** or **variables** which are quantities of interest or which serve as the practical substitutes for the concepts of interest. For example, if you have a theory about what affects people's anger level, you need to operationalize the concept of anger. You might measure anger as the loudness of a person's voice in decibels, or some summary feature(s) of a spectral analysis of a recording of their voice, or where the person places a mark on a visual-analog "anger scale", or their total score on a brief questionnaire, etc. Each of these is an example of an operationalization of the concept of anger.

As another example, consider the concept of manual dexterity. You could devise a number of tests of dexterity, some of which might be "unidimensional" (producing one number) while others might be 'multidimensional"' (producing two or more numbers). Since your goal should be to convince both yourself and a wider audience that your final conclusions should be considered an important contribution to the body of knowledge in your field, you will need to make the choice carefully. Of course one of the first things you should do is investigate whether standard, acceptable measures already exist. Alternatively you may need to define your own measure(s) because no standard ones exist or because the

existing ones do not meet your needs (or perhaps because they are too expensive).

One more example is cholesterol measurement. Although this seems totally obvious and objective, there is a large literature on various factors that affect cholesterol, and enumerating some of these may help you understand the importance of very clear and detailed operationalization. Cholesterol may be measured as "total" cholesterol or various specific forms (e.g., HDL). It may be measured on whole blood, serum, or plasma, each of which gives somewhat different answers. It also varies with the time and quality of the last meal and the season of the year. Different analytic methods may also give different answers. All of these factors must be specified carefully to achieve the best measure.

2.1 What makes a "good" variable?

Regardless of what we are trying to measure, the qualities that make a good measure of a scientific concept are high reliability, absence of bias, low cost, practicality, objectivity, high acceptance, and high concept validity. **Reliability** is essentially the inverse of the statistical concept of variance, and a rough equivalent is "consistency". Statisticians also use the word "precision".

Bias refers to the difference between the measure and some "true" value. A difference between an *individual* measurement and the true value is called an "error" (which implies the practical impossibility of perfect precision, rather than the making of mistakes). The bias is the *average* difference over many measurements. Ideally the bias of a measurement process should be zero. For example, a measure of weight that is made with people wearing their street clothes and shoes has a positive bias equal to the average weight of the shoes and clothes across all subjects.

Precision or reliability refers to the reproducibility of repeated measurements, while bias refers to how far the average of many measurements is from the true value.

All other things being equal, when two measures are available, we will choose the less expensive and easier to obtain (more practical) measures. Measures that have a greater degree of subjectivity are generally less preferable. Although devis-

ing your own measures may improve upon existing measures, there may be a trade off with acceptability, resulting in reduced impact of your experiment on the field as a whole.

Construct validity is a key criterion for variable definition. Under ideal conditions, after completing your experiment you will be able to make a strong claim that changing your explanatory variable(s) in a certain way (e.g., doubling the amplitude of a background hum) causes a corresponding change in your outcome (e.g., score on an irritability scale). But if you want to convert that to meaningful statements about the effects of auditory environmental disturbances on the psychological trait or construct called "irritability", you must be able to argue that the scales have good construct validity for the traits, namely that the operationalization of background noise as an electronic hum has good construct validity for auditory environmental disturbances, and that your irritability scale really measures what people call irritability. Although construct validity is critical to the impact of your experimentation, its detailed understanding belongs separately to each field of study, and will not be discussed much in this book beyond the discussion in Chapter 3.

Construct validity is the link from practical measurements to meaningful concepts.

2.2 Classification by role

There are two different independent systems of classification of variables that you must learn in order to understand the rest of this book. The first system is based on the role of the variable in the experiment and the analysis. The general terms used most frequently in this text are explanatory variables vs. outcome variables.

An experiment is designed to test the effects of some intervention on one or more measures, which are therefore designated as **outcome variables**. Much of this book deals with the most common type of experiment in which there is only a single outcome variable measured on each experimental unit (person, animal, factory, etc.) A synonym for outcome variable is dependent variable, often abbreviated DV.

The second main role a variable may play is that of an explanatory variable. **Explanatory variables** include variables purposely manipulated in an experiment and variables that are not purposely manipulated, but are thought to possibly affect the outcome. Complete or partial synonyms include independent variable (IV), covariate, blocking factor, and predictor variable. Clearly, classification of the role of a variable is dependent on the specific experiment, and variables that are outcomes in one experiment may be explanatory variables in another experiment. For example, the score on a test of working memory may be the outcome variable in a study of the effects of an herbal tea on memory, but it is a possible explanatory factor in a study of the effects of different mnemonic techniques on learning calculus.

Most simple experiments have a single dependent or outcome variable plus one or more independent or explanatory variables.

In many studies, at least part of the interest is on how the effects of one explanatory variable on the outcome depends on the level of another explanatory variable. In statistics this phenomenon is called **interaction**. In some areas of science, the term **moderator variable** is used to describe the role of the secondary explanatory variable. For example, in the effects of the herbal tea on memory, the effect may be stronger in young people than older people, so age would be considered a moderator of the effect of tea on memory.

In more complex studies there may potentially be an intermediate variable in a causal chain of variables. If the chain is written A⇒B⇒C, then interest may focus on whether or not it is true that A can cause its effects on C only by changing B. If that is true, then we define the role of B as a mediator of the effect of A on C. An example is the effect of herbal tea on learning calculus. If this effect exists but operates only through herbal tea improving working memory, which then allows better learning of calculus skills, then we would call working memory a **mediator** of the effect.

2.3 Classification by statistical type

A second classification of variables is by their statistical type. It is critical to understand the type of a variable for three reasons. First, it lets you know what type

of information is being collected; second it defines (restricts) what types of statistical models are appropriate; and third, via those statistical model restrictions, it helps you choose what analysis is appropriate for your data.

Warning: SPSS uses "type" to refer to the storage mode (as in computer science) of a variable. In a somewhat non-standard way it uses "measure" for what we are calling statistical type here.

Students often have difficulty knowing "which statistical test to use". The answer to that question always starts with variable classification:

Classification of variables by their roles and by their statistical types are the first two and the most important steps to choosing a correct analysis for an experiment.

There are two main types of variables, each of which has two subtypes according to this classification system:

Quantitative Variables
- **Discrete Variables**
- **Continuous Variables**

Categorical Variables
- **Nominal Variables**
- **Ordinal Variables**

Both categorical and quantitative variables are often recorded as numbers, so this is not a reliable guide to the major distinction between categorical and quantitative variables. **Quantitative variables** are those for which the recorded numbers encode magnitude information based on a true quantitative scale. The best way to check if a measure is quantitative is to use the **subtraction test**. If two experimental units (e.g., two people) have different values for a particular measure, then you should subtract the two values, and ask yourself about the meaning of the difference. If the difference can be interpreted as a *quantitative* measure of difference between the subjects, and if the meaning of each quantitative difference

is the same for any pair of values with the same difference (e.g., 1 vs. 3 and 10 vs. 12), then this is a quantitative variable. Otherwise, it is a categorical variable.

For example, if the measure is age of the subjects in years, then for all of the pairs 15 vs. 20, 27 vs. 32, 62 vs. 67, etc., the difference of 5 indicates that the subject in the pair with the large value has lived 5 more years than the subject with the smaller value, and this is a quantitative variable. Other examples that meet the subtraction test for quantitative variables are age in months or seconds, weight in pounds or ounces or grams, length of index finger, number of jelly beans eaten in 5 minutes, number of siblings, and number of correct answers on an exam.

Examples that fail the subtraction test, and are therefore categorical, not quantitative, are eye color coded 1=blue, 2=brown, 3=gray, 4=green, 5=other; race where 1=Asian, 2=Black, 3=Caucasian, 4=Other; grade on an exam coded 4=A, 3=B, 2=C, 1=D, 0=F; type of car where 1=SUV, 2=sedan, 3=compact and 4=subcompact; and severity of burn where 1=first degree, 2=second degree, and 3=third degree. While the examples of eye color and race would only fool the most careless observer into incorrectly calling them quantitative, the latter three examples are trickier. For the coded letter grades, the average difference between an A and a B may be 5 correct questions, while the average difference between a B and a C may be 10 correct questions, so this is not a quantitative variable. (On the other hand, if we call the variable quality points, as is used in determining grade point average, it can be used as a quantitative variable.) Similar arguments apply for the car type and burn severity examples, e.g., the size or weight difference between SUV and sedan is not the same as between compact and subcompact. (These three variables are discussed further below.)

Once you have determined that a variable is quantitative, it is often worthwhile to further classify it into discrete (also called counting) vs. continuous. Here the test is the **midway test**. If, for *every* pair of values of a quantitative variable the value midway between them is a meaningful value, then the variable is **continuous**, otherwise it is **discrete**. Typically discrete variables can only take on whole numbers (but all whole numbered variables are *not* necessarily discrete). For example, age in years is continuous because midway between 21 and 22 is 21.5 which is a meaningful age, even if we operationalized age to be age at the last birthday or age at the nearest birthday.

Other examples of continuous variables include weights, lengths, areas, times, and speeds of various kinds. Other examples of discrete variables include number of jelly beans eaten, number of siblings, number of correct questions on an exam,

and number of incorrect turns a rat makes in a maze. For none of these does an answer of, say, $3\frac{1}{2}$, make sense.

There are examples of quantitative variables that are not clearly categorized as either discrete or continuous. These generally have many possible values and strictly fail the midpoint test, but are practically considered to be continuous because they are well approximated by continuous probability distributions. One fairly silly example is mass; while we know that you can't have half of a molecule, for all practical purposes we can have a mass half-way between any two masses of practical size, and no one would even think of calling mass discrete. Another example is the ratio of teeth to forelimb digits across many species; while only certain possible values actually occur and many midpoints may not occur, it is practical to consider this to be a continuous variable. One more example is the total score on a questionnaire which is comprised of, say, 20 questions each with a score of 0 to 5 as whole numbers. The total score is a whole number between 0 and 100, and technically is discrete, but it may be more practical to treat it as a continuous variable.

It is worth noting here that as a practical matter most models and analyses do not distinguish between discrete and continuous *explanatory* variables, while many do distinguish between discrete and continuous quantitative *outcome* variables.

Measurements with meaningful magnitudes are called quantitative. They may be discrete (only whole number counts are valid) or continuous (fractions are at least theoretically meaningful).

Categorical variables simply place explanatory or outcome variable characteristics into (non-quantitative) categories. The different values taken on by a categorical variable are often called **levels**. If the levels simply have arbitrary names then the variable is **nominal**. But if there are at least three levels, and if every reasonable person would place those levels in the same (or the exact reverse) order, then the variable is **ordinal**. The above examples of eye color and race are nominal categorical variables. Other nominal variables include car make or model, political party, gender, and personality type. The above examples of exam grade, car type, and burn severity are ordinal categorical variables. Other examples of ordinal variables include liberal vs. moderate vs. conservative for voters or political parties; severe vs. moderate vs. mild vs. no itching after application of a skin irritant; and disagree vs. neutral vs. agree on a policy question.

It may help to understand ordinal variables better if you realize that most ordinal variables, at least theoretically, have an underlying quantitative variable. Then the ordinal variable is created (explicitly or implicitly) by choosing "cut-points" of the quantitative variable between which the ordinal categories are defined. Also, in some sense, creation of ordinal variables is a kind of "super-rounding", often with different spans of the underlying quantitative variable for the different categories. See Figure 2.1 for an example based on the old IQ categorizations. Note that the categories have different widths and are quite wide (more than one would typically create by just rounding).

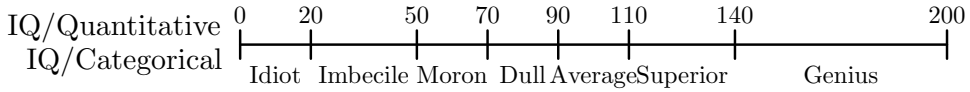

Figure 2.1: Old IQ categorization

It is worth noting here that the best-known statistical tests for categorical *outcomes* do not take the ordering of ordinal variables into account, although there certainly are good tests that do so. On the other hand, when used as *explanatory variables* in most statistical tests, ordinal variables are usually either "demoted" to nominal or "promoted" to quantitative.

2.4 Tricky cases

When categorizing variables, most cases are clear-cut, but some may not be. If the data are recorded directly as categories rather than numbers, then you only need to apply the "reasonable person's order" test to distinguish nominal from ordinal. If the results are recorded as numbers, apply the subtraction test to distinguish quantitative from categorical. When trying to distinguish discrete quantitative from continuous quantitative variables, apply the midway test and ignore the degree of rounding.

An additional characteristic that is worth paying attention to for quantitative variables is the range, i.e., the minimum and maximum possible values. Variables that are limited to between 0 and 1 or 0% and 100% often need special consideration, as do variables that have other arbitrary limits.

When a variable meets the definition of quantitative, but it is an explanatory

variable for which only two or three levels are being used, it is usually better to treat this variable as categorical.

Finally we should note that there is an additional type of variable called an "order statistic" or "rank" which counts the placement of a variable in an ordered list of all observed values, and while strictly an ordinal categorical variable, is often treated differently in statistical procedures.