To appear in *Handbook of the Philosophy of Statistics* (eds.) P.S. Bandyopadhyay and M. Forster
North Holland: Elsevier. (2010)

Elementary Probability and Statistics: A Primer

Prasanta S. Bandyopadhyay
Dept. of History & Philosophy
Montana State University
Bozeman. MT 59717

Steve Cherry
Department of Mathematical Sciences
Montana State University
Bozeman. MT 59717

Elementary Probability and Statistics: A Primer[1]

1. Introduction

Some of the chapters of this volume, albeit thorough, are technical and require some familiarity with probabilistic and statistical reasoning. We believe that an introduction to some of the basic concepts of probability and statistics will be helpful for our general reader. Probability and statistics provide necessary tools to *capture* the *uncertain state* of our *knowledge*. In addition, probability/statistics will also provide necessary tool-kits to *quantify* our uncertainties. Our discussion will be mostly quantitative. Patient readers who work through some of those technicalities will be rewarded in the end by seeing how the technicalities contribute to a better appreciation of several of the papers which on many occasions lend themselves to philosophical issues of great importance.

Below we provide a brief introduction the theory of probability and statistics. The introduction is at the level found in introductory statistics classes and discussed in many statistical textbooks. Typical texts used over the past few years at Montana State University include Devore (2008), Moore and McCabe (2006), and Deveaux, Velleman, and Bock (2008). Much of the material and many of the examples presented below are taken from Devore.

We begin our discussion on the nature of probabilistic/statistical reasoning by contrasting probabilistic/statistical reasoning/inference with deductive reasoning/inference. The property of deductive validity of an argument is central to understanding the distinction between deductive inference and inductive inference, when deductive validity could be understood in terms of the monotonic property of reasoning. First, we define "deductive validity" followed by an understanding of the property of monotonicity. An argument is deductively valid if and only if it is logically impossible for its premises to be true, but its conclusion to be false. A sentence is a *deductive consequence* of others when it is logically impossible that they should be true but that sentence is false. Consider the set S_1 which consists of the sentence P, and the sentence P→Q, where "→" captures any if P then Q sentences and the symbol "→" is known as "material conditional." Here, Q is a deductive consequence of S1, which we write as "$S_1 \models$ Q." The symbol "$\models$" denotes deductive consequence relation, which is a relation between a set of sentences and a sentence. Monotonicity is a property of certain types of inferences and is appreciated in terms of the deductive consequence relation. A relation between sets of sentences and a sentence is *monotonic* if and only if when it holds between a set and a sentence, it also holds between any superset of the set and that sentence. S_2 can be taken to be the superset of S_1 consisting of S_1 and ~P. The symbol "~" means "it is false that." The relation "$\models$" represents deductive consequence relation and any deductive consequence relation is by definition monotonic. For example, if S_2 is any superset of the set S_1 above (for which we have $S_1 \models$ Q) then $S_2 \models$Q must hold too.

To appreciate this point, consider the rule of *modus ponens*, which is a well-known rule of deductive logic.

(i) If it rains, then the ground will be wet.

(ii) It has rained. Therefore, the ground will be wet.

[1]We would like to thank Abhijit Dasgupta, Malcolm Forster, and John Bennett for their comments and suggestions regarding our chapter which have helped to improve the chapter considerably. PSB's research has been supported by Montana State University's *NASA* Astrobiology research center grant (4w1781).

"P" which is called "the antecedent" of the if-then sentence, represents the proposition "it rains." "Q", which is called "the consequent" of the if-then sentence, represents the proposition "the ground will be wet." The standard rules of deductive logic including *modus ponens* imply that adding new premises to a store of information can only *increase* the class of conclusions that can be deduced from them. If one adds, for example, ~P to the above set comprising of premise 1 and premise 2, then one would be able to deduce q along with other additional conclusions from the set, {~P, P, P→Q}, although the entire set would turn out to be inconsistent. Non-monotonic reasoning, which underlies much of our probabilistic/statistical reasoning, by contrast, allows the possibility that adding new information can actually result in *dispensing* with conclusions previously held. Many of our daily experiences are characteristic of non-monotonic reasoning in which we lose information in light of new evidence.

Consider the following example.

1. All *observed* crows are black.
2. Therefore, all crows are black.

Suppose we have observed so far that all crows are black. Seeing a picture of an albino crow (and assuming we have no reason to be skeptical about that picture) in the local newspaper, our belief that all crows are black is undermined. So an addition of information to our belief set could force us to lose some of our cherished beliefs. Here, the new premise about the presence of an albino crow has led to the rejection of its conclusion that all crows are black. This loss of information is solely a feature of non-monotonic reasoning that underlies much of probabilistic/statistical reasoning. Although we have introduced the idea of non-monotonic and monotonic reasoning intuitively in terms of whether we lose our existing information in one case and not in another, we need to be aware that it is the deductive validity that goes at the heart of monotonic reasoning which can't be undermined by adding a new premise. The purpose of our examples is just to illustrate this theme behind monotonic reasoning. Probability theory provides a better tool for handling inductive arguments via which non-monotonic reasoning has primarily been expressed. We will begin with probability theory. At the final section of this chapter, after learning about probability theory and statistics, we will, however, return to the theme of non-monotonic reasoning to evaluate whether arguments used in statistical/probabilistic inference *necessarily* involve non-monotonic reasoning.

2.1. Basic and derived rules of the probability calculus.

Probability theory provides a tool for handling inductive arguments. Typically the first step is to define some needed terms. We define a probabilistic experiment (broadly) to be any process that produces outcomes which are not predictable in advance. The toss of a single six-sided die is a simple example. We know that one of six numbers between one and six will occur on any given toss but prior to a toss we cannot state with certainty which outcome will be observed. A list of all the possible outcomes of an experiment is called the sample space. We may be interested in the individual outcomes themselves or in sets of outcomes, i.e. we will be interested in subsets of the sample space and will refer to such subsets as events. In tossing a single six-sided die once the sample space is

$$S = \{1,2,3,4,5,6\}.$$

We may be interested in the event of observing an even number. This is the subset $\{2,4,6\}$. If any of the individual outcomes in this event occur then the event itself is said to have occurred. This is true in general. For example, if we toss the die and observe a 3 then the events $\{1,3,5\}$ and $\{2,3\}$ have both occurred as has any other event containing 3 as an outcome. For any sample space the empty or null set is considered to be an event along with the entire sample space itself.

The union of two events $A \cup B$ is the event that occurs if either *A* or *B* or both occur. The intersection of two events $A \cap B$ is the event that occurs if both *A* and *B* occur. The complement of any event *A* is denoted A^c is the event comprised of all outcomes that are not in *A*. A collection of events $A_1, A_2, \cdots, A_n$ is said to be mutually exclusive (or disjoint) if none of the pairs have any outcomes in common.

The probability of an event *A*, denoted *P(A)* is a quantitative measure of how likely we are to observe *A* in a single trial of an experiment. Initially, we define or interpret probability as is typically done in introductory statistics text books. The probability of event is the long run relative frequency with which the event occurs over very many independent trials of the experiment. This definition is controversial in part because it does not always make sense (What is the probability the Denver Broncos will win the Super Bowl in 2010?). For further discussion and additional interpretations of probability see Hájek (2007). We note that the mathematical rules and properties of probability described below do not depend on the specific interpretation of probability.

The mathematical development of probability starts with three basic rules or axioms:

1. For any event A; $0 \le P(A) \le 1$. If A has probability 0 then it is impossible and if its probability is 1 then it is certain to occur. Any event of practical interest will never have probabilities of 0 or 1.
2. Denoting the sample space by S, $P(S) = 1$ (something has to happen). S could be considered a tautology.
3. For a sequence of mutually exclusive events $A_1, A_2, A_3, \cdots$,

$$P(A_1 \cup A_2 \cup A_3 \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

All other rules and properties of probability can be derived from these three axioms.

Two simple rules that are usually established quickly are that the probability of the empty set $\varnothing$ is 0 and the so-called Complement Rule: for any event *A*

$P(A^c) = 1 - P(A)$.

Another rule that can be derived quickly is a modification of axiom 3. For a finite sequence of mutually exclusive events $A_1, A_2, \cdots, A_n$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i).$$

Most events are not mutually exclusive. It can be shown that for any two events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Intuitively this makes sense. The probability that event A or event B (or both) occur is equal to the probability that A occurs plus the probability that B occurs minus the probability that both occur (which has been double counted by the addition of the probabilities of A and B). Note that if *A* and *B* are mutually exclusive then their intersection is the empty set with probability 0 and we are back to the addition rule for mutually exclusive events given above.

This rule can be extended to more than two events: for any three events A, B, and C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

In any given experiment we need to be able to assign probability to outcomes in some way. There are potentially many different ways of doing so as long as they do not violate the axioms and other rules of probability. We have seen, for example, that the sample space for the experiment of tossing a six-sided die once is

$$S = \{1,2,3,4,5,6\}.$$

We note that the outcomes are six mutually exclusive events and so

$$P(S) = P(1) + P(2) + \cdots P(6) = 1.$$

As long as the probabilities we assign to the individual outcomes are all between 0 and 1 and sum to 1, we will have a mathematically valid probability model for the experiment. There are an infinity of such possibilities. An empirical method of assigning probabilities, based on our frequentist interpretation of probability, is to toss the die a large number of times and record the outcomes. The assignment of probabilities would be based on the proportion of times each outcome occurred. Or we might reason that the die is fair in that we do not expect to see any one outcome more often than any other, i.e. each outcome is equally likely and taken together must sum to 1. Thus, each of the outcomes would be assigned a probability of 1/6.

2.2. Conditional probability and marginal probability:

The sample space of an experiment contains all the possible outcomes. Sometimes additional information becomes available that constrains the set of possible outcomes. The probability of an event then is computed based on the outcomes in the constrained sample space and such a probability is called a conditional probability. Before giving a mathematical definition we look at a simple example.

A geneticist studying two genes has access to a population of 1000 individuals. The table below shows a categorization of the individuals with respect to the genes.

Gene 1	Gene 2 Dominant	Gene 2 Recessive
Dominant	560	240
Recessive	140	60

The experiment will consist of drawing an individual at random from the population. By drawing "at random" we mean that all individuals have the same chance of being drawn, 1/1000. If the person drawn has two dominant genes we label her *D1D2*. If the person has a dominant Gene 1 and a recessive Gene 2 we label her *D1R2* and so on. The sample space is

$$S = \{D1D2, D1R2, R1D2, R1R2\}.$$

Based on what we expect to see over the long run a reasonable assignment of probabilities is

$$P(D1D2) = 0.56, P(D1R2) = 0.24, P(R1D2) = 0.14, P(R1R2) = 0.06.$$

This assignment was made by simply counting the number of individuals in each category and dividing by the total population size. Each of the numbers is between 0 and 1 and sum to 1 so the assignment is mathematically valid. Suppose we are interested in the probability that the person selected has a dominant Gene 2. We can directly apply the addition rule for mutually exclusive events and compute this as

$$P(D2) = P(D1D2 \cup D1R2) = P(D1D2) + P(D1R2) = 0.56 + 0.24 = 0.80.$$

In this simple case we can also note that 800 of the 1000 individuals have a dominant Gene 1 and conclude that the probability is 800/1000=0.80. This is the unconditional (or marginal) probability of selecting an individual with a dominant Gene 2.

Now suppose we are told that we will be selecting our individual from the group with a recessive Gene 1. We now know that we are only dealing with the subpopulation of 200 individuals with such a gene. The sample space is now

$$S = \{R1D2, R1R2\}$$

Of the 200 individuals in this subpopulation 140 are *R1D2* and 60 are *R1R2* and so the probability of a dominant Gene 2 is 140/200 = 0.70. We say that the conditional probability of a dominant Gene 2 given a recessive Gene 1 is 0.70. Mathematically we write

$$P(D2 \mid R1) = 140/200 = 0.70.$$

In general, for any two events *A* and *B* with $P(B) \neq 0$, we define the conditional probability of *A* given *B* to be

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Considering the genetics example we see that

$$P(D2 \mid R1) = \frac{P(D2 \cap R1)}{P(R1)} = \frac{140/1000}{200/1000} = \frac{140}{200} = 0.70.$$

Failure to recognize the appropriate restriction of the sample space leads to common errors in calculating probability. Suppose we are told that a family has two children and that at least one is a girl. What is the probability that both are girls? The most common intuitive answer to this question is, assuming boys and girls occur with equal probability, 0.5. This answer is wrong. Let *GG* denote the outcome of a girl born first and a girl born second, *GB* denote the outcome of a girl born first and a boy born second and so on. The sample space is

$$S = \{GG, GB, BG, BB\}$$

and, assuming equal chances for a boy or girl, a reasonable assignment of probabilities to the outcomes is 0.25, i.e. each is equally likely. We are being asked to find the conditional probability of *GG* given or conditional on the information that at least one of the children is a girl. Denoting this event by *B*, we have $B = \{GG, GB, BG\}$. The probability that *GG* occurs is 0.25 and the probability that *B* occurs is 0.75. Note also that $(GG \cap B) = GG$. Thus,

$$P(GG \mid B) = \frac{P(GG \cap B)}{P(B)} = \frac{P(GG)}{P(B)} = \frac{0.25}{0.75} = 1/3.$$

Suppose we had phrased the question as: What is the probability of two girls given that the oldest is a girl. Denoting the latter event by *B*, we have $B = \{GG, GB\}$, and

$$P(GG \mid B) = \frac{P(GG \cap B)}{P(B)} = \frac{P(GG)}{P(B)} = \frac{0.25}{0.50} = 1/2.$$

The definition of conditional probability leads immediately to the so-called Multiplication Rule:

$$P(A \cap B) = P(A \mid B)P(B)$$

The Law of Total Probability is an important result. Let $A_1, \cdots, A_k$ be a sequence of mutually exclusive and exhaustive sets. By exhaustive we mean that the union of these sets is the sample space. For any other event *B*, the unconditional probability of *B* is

$$P(B) = P(B \mid A_1)P(A_1) + \cdots + P(B \mid A_k)P(A_k).$$

This rule can be used to prove Bayes's Rule (or Theorem).:

Let $A_1, \cdots, A_k$ be a sequence of mutually exclusive and exhaustive sets, and B be event such that $P(B) \neq 0$. Then

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j) P(A_j)}{\sum_i^k P(B \mid A_i) P(A_i)}.$$

The middle term is just an application of the definition of conditional probability. The numerator of the last term follows from the Multiplication Rule and the denominator is the Law of Total Probability.

Bayes' Rule provides insight into another type of problem for which our intuition leads to incorrect probability calculations. Suppose a test exists for a rare (only 1 in 1000 adults has the disease) but serious disease and that the test is quite accurate in that when the disease is present it will return a positive result 99% of the time and when the disease is absent a positive result will occur only 2% of the time. You are given the test and receive a positive result. How worried should you be? Your first inclination would be to focus on how likely the test is to return a positive result when the disease is present and many people would choose that figure (0.99) as the probability of having the disease. But the event of interest is not observing a positive result given the disease but the event of having the disease given a positive test result. Let *D* denote the event of having the disease and + denote the event of testing positive. Based on the above information we know the following:

$$P(D) = 0.001, P(+ \mid D) = 0.99, P(+ \mid D^c).$$

By Bayes' Rule

$$P(D \mid +) = \frac{P(+ \mid D) P(D)}{P(+ \mid D) P(D) + P(+ \mid D^c) P(D^c)} = \frac{0.99(0.001)}{0.99(0.001) + 0.02(0.999)} = \frac{0.00099}{0.02097} = 0.047.$$

.

2.3. Probabilistic independence and generalized probability conjunction rule

Two events are said to be probabilistically independent if and only if the probability of the occurrence or non-occurrence of one event in no way affects the probability of the occurrence or non-occurrence of the other. Mathematically, we define two events to be independent if

$$P(A \mid B) = P(B).$$

It can be shown that this equality also implies $P(B \mid A) = P(A)$. Note that under an assumption of probabilistic independence the Multiplication Rule becomes

$$P(A \cap B) = P(A \mid B)P(B) = P(A)P(B) .$$

Some textbooks define independence in this way and then show that independence implies

$$P(A \mid B) = P(B) .$$

Typically in practical problems independence is an assumption that is made. It may be reasonable in some cases such as in assuming that repeated tosses of a six-sided die yield outcomes that are independent of one another (getting a 6 on one toss does not affect the chances of getting a 6 on the next, or any other toss). But in other situations it may not be a reasonable assumption. Subsequent to the explosion of the space shuttle Challenger in 1988 it was determined that the computed risk of catastrophic failure had been seriously underestimated due, in part, to unwarranted assumptions of independence of failure of related parts.

Mutual independence of more than two events is more complex. Mathematically it is defined as follow. A sequence of events $A_1, \cdots, A_k$ are said to be mutually independent if for every $k (k = 2,3,\cdots,n)$ and every set of indices $i_1, \cdots, i_k$

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}) .$$

For example, three events *A*, *B*, and *C* are mutually independent if

$$P(A \cap B) = P(A)P(B)$$
$$P(A \cap C) = P(A)P(C)$$
$$P(B \cap C) = P(B)P(C)$$
$$P(A \cap B \cap C) + P(A)P(B)P(C)$$

Pairwise independence does not imply mutual independence. Consider tossing two coins, a quarter and a dime. Let *H1* denote the event of observing a heads on the quarter, *T2* denote the event of observing a tails on the dime, and *C* denote the event of observing either two heads or two tails. We assume the tosses are independent of one another. The sample space of the experiment consists of four equally likely outcomes of

$$S = \{H1H2, H1T2, T1H2, T1T2\} .$$

We have

$$P(H1 \cap T2) = P(H1T2) = 1/4 = P(H1)P(T2)$$
$$P(H1 \cap C) = P(H1H2) = 1/4 = P(H1)P(C)$$
$$P(T2 \cap C) = P(T1T2) = 1/4 = P(T2)P(C)$$
$$P(H1 \cap T2 \cap C) = P(\varnothing) = 0 \neq P(H1)P(T2)P(C)$$

Thus, although the events are pairwise independent they are not mutually independent.

2.4. Probabilistic/logical independence and mutual exclusiveness of propositions.

Two events may be logically independent without being probabilistically independent. Let *A* be the event than a randomly chosen individual drinks coffee and *B* be the event that a randomly chosen individual smokes. These two events are logically independent of one another because one does not imply the other nor does one imply the negation of the other. However, social research has shown that smokers are more likely to drink coffee than non-smokers and thus these two events are probabilistically dependent.

Students in introductory probability courses frequently struggle with the relationship between two events being mutually exclusive of one another and being probabilistically dependent/independent of one another. It seems that it is a natural first inclination to consider two mutually exclusive events to be independent, but in fact if two events are independent of one another they must overlap in some way. If two events *A* and *B* are mutually exclusive (and both have non-zero probability of occurrence.) then knowing one has occurred automatically rules the other out because

$$P(A \mid B) = 0 \neq P(A) .$$

For example, if we let *H* denote getting a heads on a single toss of a fair coin and *T* denote getting a tails, then these two events are clearly mutually exclusive and dependent;

$$P(H \mid T) = 0 \neq P(H) .$$

Formally, we say that mutual exclusivity implies dependence but the converse is not true. Thus, dependent events may or may not be mutually exclusive. The negation (independence implies two events are not mutually exclusive) is, of course, also true.

For an example of two overlapping independent events consider the experiment of tossing a fair coin twice. Let *H1* denote the event of getting a heads on the first toss and *B* denote the event of two heads or two tails. These two events overlap but

$$P(H1 \mid B) = 1/2 = P(H1) .$$

3. From probability to statistics and a road-map for the rest of the chapter
3.1. The fundamental difference between probability and statistics.

One way to appreciate the difference between statistics and probability is to see how these two disciplines distinctively make inference based on data/sample. By "inference", we mean the procedure by which one goes from the known data to the unknown value of the parameter which we are interested in knowing/estimating. The style of inference applied in statistics is an inference that goes from a sample to the population. It is the paradigm example of uncertain inference. Uncertainty of statistical inference stems from the fact that the inference is made from

the known sample to the unknown population. In contrast, there are other kinds of inference where we make inferences about the unknown samples based on our information about the population. The latter kind of inference is known as non-statistical inference or inference involving probability. In short, drawing an inference from a sample to a population is statistics and drawing an inference from a population to a sample is mathematics.

Consider an urn consisting of a large number of balls of the same size, but of different colors. Suppose we know that 25% of the balls are red. We are going to draw a sample of 100 balls randomly with replacement from the urn. What is the probability that the proportion of red balls in the sample will lie between 0.20 and 0.30? Using the laws of probability we can answer this question exactly. Suppose we know nothing about the proportion of colors in the urn and we sample balls as described above. We wish to use the information in the sample to estimate the proportion of red balls in the population. What is the best way of doing this? If the observed proportion is 0.25 what conclusions can be drawn about the true but unknown proportion of red balls in the population? Is the observed proportion in the sample consistent with a true proportion of say 0.45, i.e. how likely are the observed data if the true proportion is 0.45? These are the types of questions of interest to statisticians (and to scientists using data to draw inferences to populations or processes from which the data were drawn).

3.2. Five Types of questions in statistics.

Although we have already touched some of the central issues of statistics above, we will elaborate how we could understand the enterprise of statistics in terms of five types of questions. Since data/sample are at the heart of statistics and often statistics is regarded as the science of data analysis, our questions will begin with data/sample first. These five questions are as follows:

(i) How could one describe the sample effectively?
(ii) From the sample, how one could make an inference about the total population?
(iii) How reliable will be the conclusion of our inference?
(iv) (a) Is there is a relation between two or more variables?
(b) If so, then what could we say about the relation? Is it a simple association, or is there a causal relation between them? And finally,
(v) How should one collect the sample so that it will help to produce the most reliable estimate?

For the sake of discussion, we are going to outline how different sections revolve around one or two of these questions although in this short review we won't be able to cover all key questions about statistics. (i) pertains to what we ordinarily mean by "statistics" which we take to be a collection of data. We want to know what those data tell us about a single variable representative of some object of interest. Section 4, among other topics, responds to (i). Section 6 which discusses two distinct ways of doing statistical inference, estimation and hypothesis testing, are devoted to addressing (ii). In most colleges and universities, these two approaches to doing statistical inference are an essential part of learning statistics. To be able to do statistical inference we need to develop probability models. In one sense, a probability model provides the tools by connecting data to how one should be doing inference in a systematic fashion. Section 5 discusses additional probability results needed to make reliable inferences based on the sample and a probability model. The ideas behind sampling distributions and the central limit theorem

will be expanded in section 5. Section 6 will address (iii). Statistics is an ever growing discipline with multi-layered complexity. We cannot address all aspects of this complexity in this short section. As a result, we won't address (v) in detail although it is important.

4. Data represented and described.

Data/ are the brick and mortar of statistics, but a data set does not have to be particularly large before a simple listing of data values becomes overwhelming and effective ways of summarizing the information in a data set are needed. The first chapter or two in most introductory textbooks deal with graphical and numerical methods of summarizing data. First, we discuss how the bridge between the observation and data we have, and the world they represent could be built in terms of some non-ambiguous terms, like objects, variables and scales which constitute some basics of mathematical statistics. Second, we introduce three measures of central tendency and point out their implications in critical reasoning. Third, we discuss the variance and the standard deviation, the two key units of measuring dispersion of data.

4.1. Understanding data in terms of objects, variables, and scales.

In deductive logic, we attribute truth-values to propositions. In probability theory, we attribute probability values both to events and propositions. In statistics, data which stand for our observations about the world lie at its core. In order for data to be converted into some language which is free from any ambiguity, so that what the data could furnish us with reliable information about the world, we take recourse to language of mathematical statistics.

The discussion of data in most introductory statistics textbooks typically starts with the definition of a population as a collection of objects of interest to an investigator. The investigator wishes to learn something about selected properties of the population. Such properties are determined by the characteristics of the individuals who make up the population and these characteristics are referred to as variables because their values vary over the individuals in the population. These characteristics can be measured on selected members of the population. If an investigator has access to all members of a population then he has conducted a census. A census is rarely possible and an investigator will select instead a subset of the population called a sample. Obviously, the sample must be representative of the population if it is to be used to draw inferences to the population from which it was drawn.

An important concept in statistics is the idea of a data distribution which is a list of the values and the number of times (frequency) or proportion of the time (relative frequency) those values occur.

Variables can be classified into four basic types – nominal, ordinal, interval, and ratio. Nominal and ordinal variables are described as qualitative while interval and ratio scale variables are quantitative.

Nominal variables differ in kind only. For example, political party identification is a nominal variable whose "values" are labels; e.g. Democrat, Republican, Green Party. These values do not

differ in any quantitative sense. This remains true even if we represent Democrats by 1, Republicans by 2 and so on. The numbers remain just labels identifying group membership without implying that 1 is superior to 2. Because this scaling is not liable to quantification does not mean that it has no value. In fact, it helps us to summarize a large amount of information into a relatively small set of non-overlapping groups of individuals who share a common characteristic.

Sometimes the values of a qualitative variable can be placed in a rank order. The latter might stand for the quality of toys received in different overseas cargos. Each toy in a batch receives a quality rating (Low, Medium, and High). They could also be given numerical codes (e.g. 1 for high quality, 2 for medium quality, and 3 for low quality). This ordinal ranking implies a hierarchy of quality in a batch of toys received from overseas. This ranking must satisfy the law of transitivity implying that if 1 is better than 2 and 2 is better than 3 then 1 must be better than 3. Since both nominal and ordinal scales are designated as qualitative variables, they are regarded as non-metric scales.

Interval scale variables are quantitative variables with an arbitrarily defined zero value. Put another way, a value of 0 does not mean the absence of whatever is being measured. Temperature measured in degrees Celsius is an interval scale variable. This is a metric scale in which for example the difference between 2 and 5 is the same as the difference between 48 and 51.

In contrast to interval scale data, in "ratio" scale data, zero is actually a pointer of "nothing" scored on the scale just as we see zero on a speedometer which signifies no movement of a car. Temperature measured in degrees Kelvin is a ratio scale variable because a value of 0 implies the absence of all motion at the atomic level.

Mathematical operations make sense with quantitative data whereas this is not true in general of qualitative data. This should not be taken to mean that qualitative data cannot be analyzed using quantitative methods, however. For example, gender is a qualitative variable and it makes no sense to talk about the "average" gender in a population but it makes a lot of sense to talk about the proportions of men and women in a population of interest.

4.2. Measures of central tendency

The data distribution can be presented graphically (e.g. in a histogram) or tabularly. Graphical analyses are an important and often overlooked aspect of data analysis itself in part because it seems so simple. We do not go into much detail here because we are more interested in statistical inference but we emphasize the importance of graphical summaries as an overall part of an initial data analysis.

Graphical summaries can provide a quick overall impression of the data but we need more. One important property of a sample (and by extension of the population from which it was drawn) is the location of its "center".

We suppose we have a sample of size *n* from some population of interest. We denote the values of the observations in the data set by $x_1, x_2, \cdots, x_n$ .There are at least three distinct measures of central tendency: (i) the mode, (ii) the mean and (iii) the median. The mode is the most frequent value in the data. The mean is computed by adding up all the numbers and dividing by the size of the sample. We denote the mean by

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

The median is the middle value in the sense that half the values lie at or above the median and half lie at or below it. Computation of the median starts with ordering the data, generally from lowest to highest. If the sample size is odd the median is the middle value. If the sample size is even, the median is the mean of the two middle values.

By way of example consider the number of home runs hit by Babe Ruth during his 15 years with the New York Yankees (1920 to 1934),

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

Forty six home runs appear most often (3 times) and this is the mode. The mean number of home runs hit by Ruth is

$$\bar{x} = (1/15)(54 + 59 + \cdots + 22) = 659/15 = 43.9 .$$

We compute the median by first ordering the data from low to high

22, 25, 34, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60.

Since *n* is odd, we choose the middle value which is 46 (the same as the mode). The median is not unique in this data set.

The question of which measure is most appropriate arises immediately. The answer depends on context. The mode is generally easy to identify if one has already carried out a graphical assessment and may be perfectly adequate in some cases. However, it is rarely used in practical statistics anymore. Indeed it is rare to see the mode given in research papers in the biological and ecological sciences and we do not discuss it further here.

The mean and median are commonly used as numerical summary statistics, however. Often both are provided. The mean has the advantage that it uses all the data; however it is not resistant to unusual values in the data set. A common example used by one of the authors when he teaches introductory statistics is to imagine the effect on the mean annual income of the people in the classroom if Bill Gates was to walk in the door. The median is resistant to unusual or atypical observations but it only requires one or two data points and thus ignores a lot of information in the data set.

We can illustrate these points by considering the graphical representation of idealized data distributions below. The graphs are histograms with smooth curves added. The data in Fig. 1 are skewed to the right (they have a long right tail)., the distribution in Fig. 2 is skewed to the left , and Fig. 3 is symmetric.

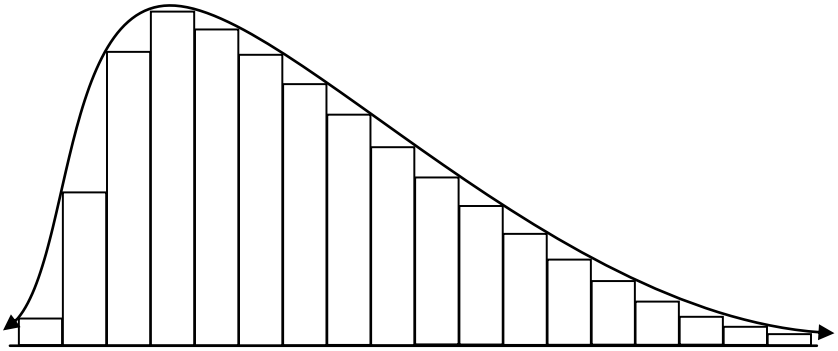

Fig. 1. Right Skewed

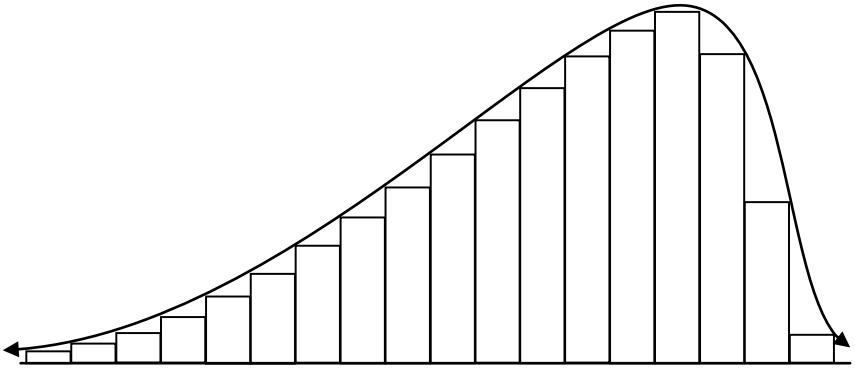

Fig. 2. Left skewed

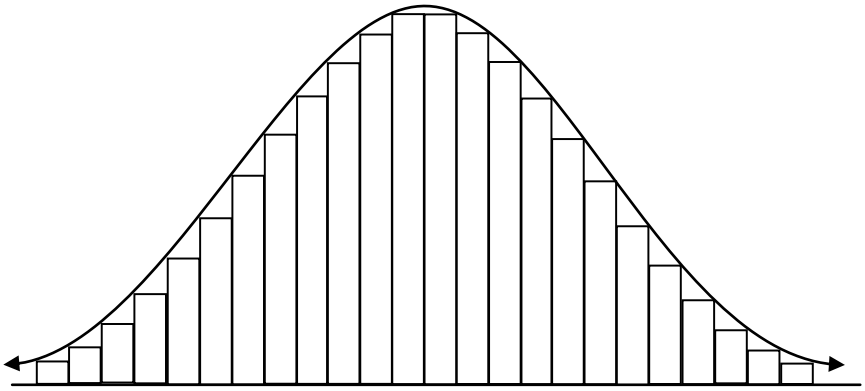

Fig. 3. Symmetric.

In each figure the median is the value that has half the area under the curve above it and half the area below it. The mean of the data set depicted in Fig. 1 will be greater than the median because the values in the long right tail "pull" the mean that direction. Similarly the mean in Fig. 2 will

be less than the median. The mean and median will be equal to one another in Fig.3. The implication then is that the median is a better measure of central tendency (in the sense of identifying the most typical value) when data distributions are skewed and that it is comparable to the mean when the data distribution is symmetric. This would seem to imply that the median is then superior to the mean, and this argument may well be valid for descriptive data analysis. But the mean is more important for inferential analysis as will be discussed below. Note that the mode would generally not be interpretable as the most typical value when data are skewed. The mode, median, and mean are all equivalent for descriptive purposes when the data distribution is symmetric. However, the mean is more commonly used.

Which measure to use depends in large part on how it will be used. Generally, an investigator is interested in drawing an inference from a sample to a population. An investigator may be interested in estimating an unknown population mean defined to be (for a finite population of size *N*)

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \ .$$

The sample mean $\bar{x}$ would seem to be the natural estimator to use in this case. However if the population distribution is symmetric and the sample is drawn in such a way that it is representative of the population (in which case the sample distribution is approximately symmetric) the population mean could be estimated by any of the central tendency measures described above. Which is the best? Answering this question is one reason why probability theory is so important to the theory of statistic and we will return to this topic below (see section 7.1).

Strictly speaking these measures of center are only appropriate for interval or ratio scale data. They are sometimes applied to ordinal data under the (strong) assumption that the rankings can be treated as numerical. Nominal data are sometimes summarized by providing the proportion of individuals in each category. Proportions can be viewed as means. For example, we have *n* observations of people in a population classified as male or female we could give a value of 1 to males and a value of 0 to females. We would have a list of *n* 0s and 1s and the sum of those divided by the sample size would be the proportion of males in the sample.

4.3 Measures of Dispersion

There is an old joke in statistics about a statistician who had his head in an oven and his feet in a freezer. When asked how he felt he replied, "On average I am quite comfortable". How the values in a data set are spread out around the center is an important part of data description. Such variation is commonplace (indeed it is the rationale for the definition of the term variable) and in many cases it may be as interesting if not more interesting than the center.

The appropriate measure of dispersion to use is determined in part by the measure of center used. When the mean is chosen interest centers on quantifying the deviations of the observations about the mean. The sample variance is typically defined to be

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

This is a kind of average squared deviation about the mean. The units of measurement of the sample variance are awkward, (if the variable is cost of a home then the units of the variance are in squared dollars) and, especially for descriptive purposes, the positive square root of the variance is often given. This is referred to as the standard deviation. Both the variance and standard deviation are non-negative and will be equal to 0 if and only if all the values in a data set are the same. This will never or at least very rarely happen in practice.

The sample mean and variance are both sensitive to atypical observations because of their dependence on the sample mean.

Why do we divide by n -1?. Not every text does and if the goal is to describe variability in a sample then either one can be used. However, the ultimate goal is to use the sample numerical summary statistics to estimate their unknown population counterparts. Consider a finite population of size N . The population variance is defined to be

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2 .$$

If we are to use the sample variance to estimate the population variance and if we knew the population mean μ then dividing by n in the formula for the sample variance would be appropriate. But as a general rule we will not know the population mean and will estimate it using the sample mean. The variability of the n observations in the sample about the sample mean will tend to be less than their variability about the population mean. Thus, dividing by the sample size will tend to underestimate the population variance. Dividing by n-1 provides the necessary correction to this underestimation. We will return to this topic below when we discuss bias in statistical estimators.

As an example consider a simple data set: 10, 12, 14 and 16. The table below summarizes computation of the sample variance and standard deviation.

Observation	Value	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	10	10-13=-3	9
2	12	12-13=-1	1
3	14	14-13=1	1
4	16	16-13=3	9
Total	52	0	20

The sample mean is 52/4=13. The sample variance is 20/3=6.66 and the sample standard deviation is the square root of 6.66 or 2.58. Note that the sum of the deviations in column 3 is 0. This is true in general and is the reason why we cannot just "average" the deviations from the mean.
The variance and standard deviation are computed using the mean and are thus naturally paired with the mean. It is not correct, for example, to summarize location using the median and summarizing variability using the standard deviation. If the median is chosen as the measure of center another measure of variability is needed. One such is based on the sum of the absolute deviations from the median. Another cruder measure is the Interquartile Range which is the difference between the third quartile (75^{th} percentile) and first quartile (25^{th} percentile). We will not discuss such measures further here.

For most distributions the bulk of the possible values will lie within a couple of standard deviations of the mean. For symmetric distributions like the idealized example in Fig. 4, about 68% of the values will lie within one standard deviation of the mean, about 95% will lie within two standard deviations of the mean and over 99% will lie within three standard deviations of the mean. These figures are roughly accurate for skewed distributions as long as they are not badly skewed. The figures are exact for so-called normal distributions (also referred to as bell-shaped or Gaussian distributions). We will discuss this more below.

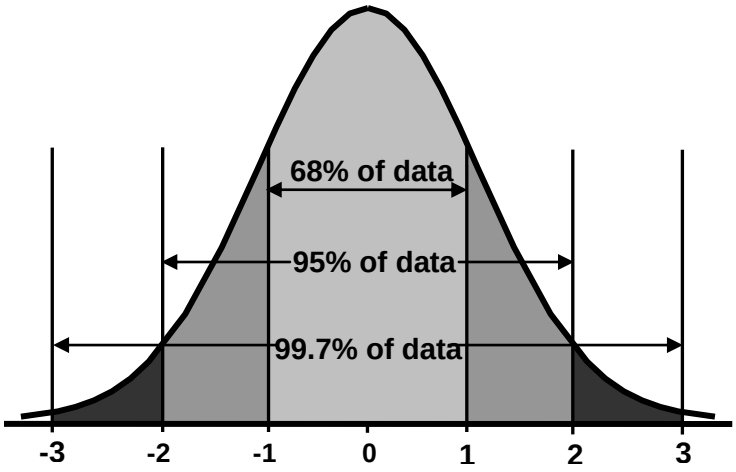

Fig. 4. 68-95-99.7 rule.

Figure 4. A graphical representation of the values in an idealized symmetric population with mean 0 and standard deviation 1.

4.4 From descriptive statistics to inference

The material presented in the preceding section is typically referred to as descriptive statistics and no probability is required to understand and to use the methods described above. However, at points in the discussion we alluded to the use of descriptive summary statistics as estimators of unknown population quantities, e.g. using the sample mean to estimate a population mean. The

sample mean is based on values observed in a subset of the population and will not in general be equal to the population mean. If we observe a sample mean of say 400 what can we say about the population mean? Is it exactly equal to 400? This is not very likely. Could the population mean be 425 or 350 or 675?

In determining how to quantify the uncertainty in the observed value of a statistic such as the sample mean, we ask ourselves the question, "What would we expect to see if the basic sampling process was repeated over and over again, independently, and under the same conditions"?. That is, we conduct the thought experiment of repeating our sampling plan many times, computing the value of the statistic of interest each time and looking at the resulting distribution of those values. We are interested in the long run behavior of the statistic. Obviously we cannot do this in practice. To carry out this thought experiment we must make an assumption about the distribution of the values in the population and about how we draw values from the population. For the latter we must assume that we have a sample that is representative of the population in which we have an interest. Formally, we say that we have a simple random sample from the population, by which we mean that all possible samples of the same size have the same probability of being selected. Once we have a reasonable probability model for the process that generated our data we can use the results of probability theory to investigate the probabilistic behavior of sample statistics of interest. In the next section, we continue our discussion of probability theory with the goal of doing just that.

5. Random variables and probability distributions

There will be many experimental situations in which interest is primarily focused on numerical characteristics of various outcomes. We might be interested in questions such as "What is the probability of observing 6 or more heads in 10 tosses of a fair coin"?, or perhaps more practically, "What is the probability of surviving five years after being diagnosed with cancer"? We will need to be able to associate each outcome with a number. The rule which determines the association is called a random variable. In some ways this is an unfortunate terminology. Strictly speaking, a random variable is neither random nor a variable – it is a mathematical function mapping outcomes in a sample space to the real line. Typically however, introductory statistics textbooks steer clear of this technical definition and motivate the name as, for example, Devore (2009, p 87) does when he writes that a random variable is "a variable because different numerical values are possible and random because the observed value depends on which of the possible experimental outcomes results". We will need to assign probability to the values a random variable can assume. This assignment is done by means of a probability model or probability distribution.

We will divide this section into three short subsections. First, we will introduce the notion of a random variable, the expectation of a random variable and some basics of probability distribution. Once we become acquainted with random variables, we will discuss two types of popular probability distributions. They are, (i) normal distribution and (ii) binomial distribution. We spend some time on the nature of normal distribution and how we could convert it to its standard normal form.

5.1. Random variables and the basics of probability distributions.

In simple language, a random variable is a rule for assigning a number to the outcome of a random process. By convention random variables are denoted by upper case letters from the Roman alphabet (X, Y, W, etc). Realizations (i.e. observed values of a random variable are denoted by the corresponding lower case letters (x,y,w, etc).

Consider tossing a fair coin three times and recording the outcomes in terms of heads or tails. The sample space is comprised of eight equally likely outcomes

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$

Let X = the number of heads observed on any single trial of this experiment. X is a random variable. It can take on one of four values; 0, 1, 2, or 3. The probability that X takes on these values is determined by the probabilities associated with the outcomes in the sample space. For example X=2 if we observe any one of three possible outcomes *HHT*, *HTH*, or *TTH*. Thus, the probability X=2 is

$$P(X = 2) = P(HHH \cup HHT \cup THH) = 3/8$$

by the addition rule for mutually exclusive events. We summarize all the possibilities in the probability distribution for *X:*

X	0	1	2	3
$p(x) = P(X = x)$	1/8	3/8	3/8	1/8

Note that all values of $p(x)$ are between 0 and 1 and sum to 1 so this is a valid assignment of probabilities to the values of the random variable. The probability distribution $p(x)$ of the random variable X = the number of heads observed in three tosses of a fair coin describes how probability is assigned to the possible values X can assume.

Random variables can be discrete or continuous. A discrete random variable is either finite or countably infinite. The random variable X = the number of heads in three tosses of a fair coin is discrete and finite. As an example of an infinite discrete random variable, consider the experiment of tossing a coin until a head is observed. Let X be the number of failures recorded until the heads occurs. Of course, in practice we would never toss the coin forever but there is no well-defined upper limit and it may make sense to model the outcomes of this experiment by assigning X the possible values 0, 1, 2, 3,…. Most commonly used discrete random variables are integer valued but they do not need to be. Continuous random variables can take on uncountably many values in an interval on the real line. By uncountable we mean that the values of the random variable cannot be placed in a one-to-one correspondence with the natural numbers. Consider the random variable, X = the length of a randomly selected phone call. Although one could measure the call in terms of its minutes and seconds, the latter still fail to measure it in

terms of its other shorter intervals. As a result, it is better to regard it as a continuous random variable capable of taking on any value on the positive real line.

In this chapter, we will mostly confine ourselves to discrete random variables. Each random variable is associated with a probability distribution (function), which must satisfy the rules of the probability theory. We saw a simple example above. In general a probability distribution for a discrete random variable (sometimes also called a probability mass function) is defined as follows. A discrete random variable X takes a finite or countably infinite number of values. The values it can assume are determined by the outcomes in the associated sample space, i.e. for every outcome o in the sample space there is an associated value for X. Mathematically then X is a function mapping outcomes o to numbers x, $X(o) = x$ although for convenience we usually suppress the functional notation. The probability distribution (or model) for X is then

$$p(x) = P(X = x) = P(o \in S, X(o) = x).$$

A function $p(x)$ is a valid probability distribution if

(i) $0 \leq p(x) \leq 1$ for all $\boldsymbol{x}$

(ii) $\sum_x p(x) = 1$ where the sum is taken over all possible values of x

The probability $P(X \in A)$ is found by summing the probabilities associated with the values x in the event A.

Consider tossing a fair die once and recording the number on the upturned face, X. Clearly there are six different values this random variable can take on and each one is equally likely, i.e.

$$P(X = 1) = \cdots = P(X = 6) = 1/6.$$

The probability that we observe a value of X greater than or equal to 3 is

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = 2/3.$$

We will observe different values of a random variable over many repeated trials of its associated experiment. The expected value of a random variable is a weighted average of its possible values. Formally, for a discrete random variable X the expected value of X is defined to be

$$E(X) = \mu_X = \sum_x xp(x).$$

There are several ways to think about expected values. Given a long-run relative frequency interpretation of probability the expected value can be thought as the long-run average. That is, if we imagine repeating the experiment a large number of times and computing the average of the observed values of the random variable that average should be close to the expected value. Another way to think about it is to imagine a "perfect" sample, i.e. a sample that contains the values we expect it to contain given the probability distribution. The average of that perfect sample is equal to the expected values.

Consider a random variable *X* with the following probability distribution

X	0	1	2	3
p(x)	0.48	0.39	0.12	0.01

Now imagine the experiment is repeated 100 times so that we observe 100 values of *X*. Based on the probability distribution we expect to see 48 values of 0, 39 values of 1, 12 values of 2, and 1 value of 3. The mean of these 100 numbers will be

$$\frac{48(0)+39(1)+12(2)+3(1)}{100} = 0.66$$

which we can rewrite as

$$0.48(0)+0.39(1)+0.12(2)+3(1) = 0.66$$

which is the expected value of *X*

We need some way to quantify the variability in those possible values of a random variable. We define the variance of a random variable to be

$$V(X) = \sigma_X^2 = \sum_x p(x)\left(x - E(x)\right)^2.$$

There is a short cut formula which is easier for computational purposes when computing variances by hand

$$\sigma_X^2 = \sum_x p(x)x^2 - \mu_x^2.$$

The variance depends on how often one expects each value of *X* to occur and how far away the *X* values are from the expected value of *X*. One could interpret the variance of *X* as the weighted average squared distance from *E(X)*, using the probabilities of each value of *X* as the weights. The standard deviation σ_X of the random variable is the positive square root of the variance of *X*.

Recall the example above involving tossing a six-sided die once. We let X denote the number of dots on the upturned face and noted that under an assumption that the die was fair we had the following probability distribution

$$P(X = 1) = \cdots = P(X = 6) = 1/6.$$

The expected value is

$$\mu_X = (1/6)(1 + 2 + 3 + 4 + 5 + 6) = 3.5.$$

The variance, computed using the short cut formula is

$$\sigma_X^2 = (1/6)(1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 = 2.92.$$

The standard deviation is 1.71.

Two discrete random variables X and Y are said to be independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) = p(x)p(y)$$

for all possible values of x and y. Essentially what this says is that two discrete random variables are independent if the joint probability mass function factors into the product of the marginal mass functions.

Modifications of the above arguments are needed for continuous random variables. Whereas probability distributions for discrete random variables assign probability to single values probability distributions for continuous random variables assign probability to intervals of real numbers. The probability distribution for a continuous random variable is called a probability density function. A mathematically valid probability density function is any function f satisfying the following properties,

1) $f(x) \geq 0$ for all x

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

The probability that X takes on a value in the interval (a,b) is

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

The expected value of a continuous random variable is defined to be

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx$$

The variance of a continuous random variable is defined to be

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx .$$

Unlike discrete random variables the probability that a continuous random variable will take on a single value is 0

$$P(X = a) = \int_a^a f(x)dx = 0 .$$

Two continuous random variables *X* and *Y* are independent if their joint density factors into the product of the marginals

$$f(x,y) = f_X(x)f_Y(y) .$$

There are an infinite number of both discrete and continuous probability distributions. In practice however a relative handful of families of distributions are used. We will briefly consider one discrete distribution and one continuous distribution important in the practice of statistics.

5.2. The normal distribution

The normal distribution is the most important in statistics. It is the distribution people have in mind when they referred to the "bell-shaped" curve and it is also often referred to as the Gaussian distribution after the mathematician Karl Friedrich Gauss.

A continuous random variable *X* is said to have a normal distribution (or be normally distributed) with mean μ and variance σ^2 if its probability density function is

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

The mean can be any real number and the variance any positive real number. We say that *X* is $N(\mu,\sigma^2)$.

There is a different normal distribution for each pair of mean and variance values and it is mathematically more appropriate to refer to the family of normal distributions but this distinction is generally not explicitly made in introductory courses.

The history of the normal distribution is fascinating (Stigler 1986). It seems to have first appeared in the work of Abraham DeMoivre in the mid-18th century and Gauss found it useful for work he was doing in the late 18th and early 19th centuries. It was imbued with semi-mystical significance initially. Some were impressed by the fact that this one distribution contained the three most famous irrational numbers, *e*, $\sqrt{2}$, and π. Normally distributed variables were considered to be a law of nature. Although not viewed quite as reverentially today it is still important for reasons which will be discussed in more detail below.

The graph of the density function is symmetric about μ. There is a different curve for every μ and σ^2. In any normal distribution 68% of the data fall within σ (one sigma) of the mean μ, 95% of the data fall within 1.96σ of μ, and 99.7% of the data fall within 3σ of μ. These proportions are the same for any normally distributed population. For simplicity, we frequently convert the values from the units in which they were measured to unitless standard values. Figure 4 shows an example of the so-called standard normal distribution with mean 0 and variance equal to 1 along with an illustration of the 68-95-99.7 rule.

To transform a normally distributed random variable into a standard normal random variable we subtract the mean and divide by the standard deviation. The result is typically referred to as a Z score,

$$Z = \frac{X - \mu}{\sigma}$$

The random variable *Z* has a *N*(0,1) distribution.

As an example, suppose the heights of American young women are approximately normally distributed with μ = 65.5 inches and σ = 2.5 inches. The standardized height

$$Z = \frac{height - 65.5}{2.5}$$

follows a standard normal distribution. A woman's standard height is the number of standard deviations by which her height differs from the mean height of all American young women. A woman who is 61 inches tall, for example, has a standard height of

$$Z = \frac{61 - 65.5}{2.5} = -1.8$$

or 1.8 standard deviations less than the mean height.

The standard normal distribution is important in introductory classes because it simplifies probability calculations involving normally distributed random variables. Because the normal distribution is a continuous distribution probabilities can be computed as areas under the density curve. But the probability density function does not have a closed form integral solution and those areas must be determined numerically. Further, many introductory courses in statistics do

not require calculus as a prerequisite and so integration is not an assumed skill. Tables of probabilities (areas) associated with the standard normal distribution are provided in introductory statistics texts. Finding the probability that a normally distributed random variable X with mean μ and variance σ^2 falls in some interval (a,b) is solved by converting to standard units and using the tabled values.

Using the standard normal distribution to solve probability problems is no longer of much practical importance because probabilities can now be determined using computer software but the standard normal random variable still plays a major role in statistical inference as we will see.

The family of normal distributions has some nice mathematical properties not shared by other probability distributions. These mathematical properties explain, in part, the importance of the normal distribution in statistics. Users of statistics are often interested in linear transformations of their data or in combining data through the use of linear combinations. For example, the sample mean is a linear combination of observed data. The application of probability theory to data analysis must take this into account. Linear functions and linear combinations of normally distributed random variables have a property that is not shared by other probability distributions that might serve as a model for a data generating process.

Suppose X is a normally distributed random variable with mean μ and variance σ^2. Then any linear transformation of the form $Y = a + bX$ (with $b \neq 0$) will be normally distributed with mean $a + b\mu$ and variance $b^2\sigma^2$.

Suppose we have a sequence of independent random variable $X_1, X_2, \cdots, X_n$ with means $\mu_1, \mu_2, \cdots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. Let $a_1, a_2, \cdots, a_n$ be constants. What is the probability distribution of the linear combination

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \, ?$$

It can be shown that if each of the X_i s is normally distributed then the linear combination Y is also normally distributed with mean $a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$ and variance

$$a_2^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \, .$$

The important point of the above results is not the resulting means and variances. Those results hold for any probability distribution. What is important and, mathematically speaking, remarkable is that linear transformations and linear combinations of normally distributed random variables are themselves normally distributed.

Many commonly used statistical methods start with an assumption that observed data are a representative sample drawn from a population of individuals. A frequent goal is to use summary information in the sample to draw inferences to unknown corresponding quantities in the population. For example, the sample mean of a data set is commonly used to estimate an unknown population mean. Quantifying the uncertainty associated with the estimate requires a

probability model. The data are viewed as realizations of a sequence of independent random variables $X_1, X_2, \cdots, X_n$. The sample mean, viewed as a random variable is

$$\overline{X} = (1/n)(X_1 + X_2 + \cdots + X_n).$$

Given the additional assumption that the values in the population can be approximated by a normal distribution with mean μ and variance σ^2 then $\overline{X}$ will be normally distributed with mean μ and variance σ^2/n. We discuss the implications of this result in more detail below.

The normal distribution is important in statistics for another reason, a truly remarkable and fascinating result: . The Central Limit Theorem (CLT). There are different versions of the CLT but we will consider it as it pertains to the probability distribution of the particular linear combination of random variables called the sample mean. Suppose we have a sequence of independent random variables $X_1, X_2, \cdots, X_n$ all sharing the same finite mean μ and finite variance σ^2. In the context of statistics we think of these random variables as constituting a random sample from a population with mean μ and variance σ^2. We make no other distributional assumptions about the random variables, i.e. about the distribution of values in the population. The sample mean is

$$\overline{X} = (1/n)(X_1 + X_2 + \cdots + X_n).$$

The Central Limit Theorem says that if the sample size *n* is large enough then $\overline{X}$ will be approximately normally distributed as $N(\mu, \sigma^2/n)$.

How large is "large enough"? A value frequently seen in introductory statistics texts is that $n \geq 30$. But, like all rules of thumb this one should not be applied indiscriminately. For some well behaved distributions (i.e. symmetric with small variances) sample sizes of 5 to 10 may suffice. For other distributions, especially those with high variance (known as fat or heavy-tailed distributions) the required sample size can be significantly greater than 30.

5.3 The binomial distribution

Consider an experiment that has only two possible outcomes which can be labeled a "success" or a "failure". The probability of a success is denoted by *p* and the probability of a failure by 1-*p*. The simplest such experiment is a coin toss with Heads counting as a "success". If the coins is fair then $p = 0.5$. An experiment of this type is called a Bernoulli trial.

We will often be interested in counting the number of successes in a sequence of *n* such trials. The number of successes is a random variable with a probability distribution. We make the following two assumptions:

1) The *n* trials are independent, i.e. the outcome of any given trial does not affect the outcomes of other trials.
2) The probability of success *p* is the same on all the trials.

Then *X* = the number of successes is said to have a binomial distribution with parameters *n* and *p*, or using a common form of mathematical shorthand we say that *X* is *Bin(n,p)*. The probability distribution (or probability mass function) is

$$p(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for $x = 0,1,\cdots,n$. The $n!$ term is called "*n* factorial" and is defined as

$$n! = n(n-1)(n-2)\cdots(3)(2)(1)$$

for any positive integer *n*. We define $0! = 1$. This distribution is derived using the rules of probability. Any given sequence of *x* successes and *n-x* failures in has probability $p^x(1-p)^{n-x}$ from the assumption of independence and constant probability of success. There are

$$\frac{n!}{x!(n-x)!}$$

mutually exclusive sequences with *x* successes and *n-x* failures. Thus to compute the probability of *x* successes we add up the $\frac{n!}{x!(n-x)!}$ probabilities $p^x(1-p)^{n-x}$. We have used the multiplication rule for mutually independent events and the addition rule for mutually exclusive events.

The mean and variance of a binomially distributed random variable *X* are $\mu_X = np$ and $\sigma_X^2 = np(1-p)$, respectively.

The binomial distribution is a common probability model chosen for data analysis problems involving proportions. Examples include estimating the proportion of voters who favor a particular candidate, the proportion of cancer patients cured by a new treatment, the proportion of individuals in a population who have a particular disease, and so on. The proportion of successes in *n* Bernoulli trials is

$$\hat{p} = X / n .$$

The sample proportion is a linear transformation of *X* and is also a random variable. It has mean $\mu_{\hat{p}} = p$ and variance $\sigma_{\hat{p}}^2 = p(1-p)/n$, respectively. Use of $\hat{p}$ as a statistical estimator requires knowledge of its probability distribution and unlike normally distributed random variables linear

transformations of binomial random variables are not binomially distributed. However, if the sample size is large enough then the Central Limit Theorem implies that $\hat{p}$ will be approximately

$$N\left(p, \frac{p(1-p)}{n}\right).$$

This result follows from the fact that the sample proportion can be considered the mean of a random sample of Bernoulli random variables. In this case "large enough" is a value of *n* such that both *np* and *n(1-p)* are greater than 10.

5.4 Sampling distributions

We know by now that in statistical inference, we make inferences about a population based on information in a sample which is a subset of the population. Mostly we do not know the details of the population and we will use the information in the sample to estimate the unknown population quantities of interest. Obviously, we require that the sample be representative of the population. We also need a probability model.

A population of interest has been identified and interest centers on a numerical characteristic of the population. We call such characteristics parameters and these are constants. A common parameter of interest is the population mean, μ. A sample will be drawn and the sample mean will be used to estimate the population mean. The distribution of the values associated with the individuals in the population is called the population distribution. Prior to data collection we consider the sample to be a sequence of independent random variables. $X_1, X_2, \cdots, X_n$. Each of these random variables has the same probability distribution with mean μ (equal to the population mean) and variance σ^2 (equal to the population variance), The sample mean is of course

$$\overline{X} = (1/n)(X_1 + X_2 + \cdots + X_n).$$

A common assumption is that the population distribution is normal with mean μ and variance σ^2. We know from the above that $\overline{X}$ is normally distributed with mean μ and variance σ^2/n. This is the probability distribution of the random variable $\overline{X}$. In the field of statistics this random variable will be put to a particular use and statisticians have their own terminology to describe this quantity. It is a statistic and its probability distribution is called a sampling distribution. In general the sampling distribution of a statistic is a description of how probability is assigned to the possible values the statistic can take on. What does the sampling distribution of $\overline{X}$ tell us? We see that the expected value or mean of this statistic is equal to the parameter it will be used to estimate. We say that $\overline{X}$ is an unbiased estimator of μ. We see that the variability in the values the sample mean can assume is a function of the population variance and the sample size. In particular as the sample size gets larger the variance of $\overline{X}$ decreases. This is a desirable characteristic; the larger the sample, i.e. the more information we have about the population, the more we can trust that $\overline{X}$ will produce a good estimate of μ. The sample mean then has two

properties we would like to see in any statistic; we want statistics to be unbiased (at least approximately) and we want the variability in the possible values the statistic can assume to decrease with increasing information (larger sample size).

The rules of probability mean that an assumption of a normal population distribution results in a normal sampling distribution for the sample mean. For example, we know that prior to data collection there is a 95% probability that the calculated value of the sample mean will lie with 1.96 $\sigma / \sqrt{n}$ units of the population mean.

If the assumption of normality for the population is not valid then the sampling distribution of the sample mean will no longer be normal although the sample mean still has expected value (mean) of μ and variance of σ^2 / n. Thus, it is still unbiased and it still has a variance that is decreasing as the sample size increases. Further, if the sample size is large enough, the sampling distribution of the sample mean will still be approximately normal by the Central Limit Theorem and prior to data collection there is an approximate 95% probability that the calculated value of the sample mean will lie within 1.96 $\sigma / \sqrt{n}$ units of the population mean.

A related and common application is estimation of a population proportion. The population of interest can be visualized as a population of two values, 0 or 1 with 1 denoting a "success". A simple random sample of size *n* will be taken and *X* = the number of successes will be counted. The sample proportion

$$\hat{p} = X / n$$

will be computed and used to estimate the population proportion *p*. Given these assumptions *X* will have a binomial distribution with mean *np* and variance *np(1-p)* and $\hat{p}$ will have mean *p* and variance *p(1-p)/n*. The statistic $\hat{p}$ is an unbiased estimator of the parameter *p* and its variance decreases with increasing sample size. Its sampling distribution is known but is awkward to work with. From the Central Limit Theorem we know that if *n* is large enough then the sampling distribution of $\hat{p}$ will be approximately normal with mean *p* and variance *p(1-p)/n.*

6. Statistical inference.

Statistical inference involves drawing conclusions about an entire population based on the information in a sample. For the sake of discussion, we divide this section into three subsections. First, we broach the two ways of doing inference, (i) estimation, and (ii) the test of significance. Second, we distinguish between two ways of conducting estimation, (i) point estimation and (ii) confidence interval estimation. Third and finally, we introduce the concept of the test of significance.

6.1. Two types of ways of doing theory of inference: estimation & testing.

Consider first the estimation problem and the test of significance. In the problem of estimation, we determine the values of the parameters, and in the test of significance, we determine whether

the result we observe is due to random variation of the sample or due to some other factor. Problems of estimation can be found almost everywhere, in science and in business. Consider an example in physics where Newcomb measured the speed of light. Between July and September 1882, he made 66 measurements of the speed of light. The speed of light is a constant but Newcomb's measurements varied. Newcomb was interested in determining the "true" value of the parameter; the speed of light. In business, a retailer might like to know something about the average income of families living within 3 miles of her store.

Note the difference in the two populations in the examples. The retailer's population of interest is finite and well-defined. Theoretically she could sample every household of interest and compute the true value of the parameter. Logistically, such a sampling plan, called a census, is often practically impossible and a sample is taken from that population. Newcomb's population is hypothetical. It is comprised of all possible measurements he could have theoretically made and estimation of the parameter of interest requires an assumption that the 66 observed measurements are a random sample from that population.

Scientists and other users of statistics often have a different question of interest. Imagine Newcomb in an argument with another scientist. Newcomb believes the speed of light is equal to one value and the other scientist believes the value is less than that. The goal of the study would have been to assess the strength of evidence in the data for or against these hypotheses. The retailer could be considering an expansion of her store but will not do it unless the mean income of families in the nearby area exceeds a specified threshold. She takes a sample, not so much with the intent of estimating the income but of determining if there is sufficient evidence that it is high enough to justify expansion.

Statisticians have argued and continue to argue over the best ways to do these two types of inference. Testing in particular is controversial. Below we will continue with the approach we have taken above and present the basics as they are often presented in introductory courses in statistics. Our goal is to present the concepts, not practically useful methods.

6.2. Two types of estimation problems: point and interval estimation.

An investigator is interested in a population, in particular, in the value of some constant numerical quantity θ, called a parameter. The value of θ is unknown. A representative sample of size *n* will be taken (denoted $X_{1,} \cdots, X_n$) and information in the sample will be used to estimate the parameter. This is done by distilling the information in the *n* values in the data set into a single summary quantity called a point estimator or statistic, $\hat{\theta}$. Under an assumption of random sampling the estimator (statistic) is a random variable that is a function of the data,

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n).$$

For convenience the functional notation is generally suppressed but it is important to understand that point estimators are random variables that are functions of data.

In practice, a single sample is taken and a single value of the point estimator is obtained. That value, a realization of the estimator, is called an estimate. In practice, we only have a single sample and a single estimate but we know that if we took another sample from the same population that we would not get the same value. Recall from above the thought experiment of taking very many samples and observing very many realizations of the estimator. The distribution of those values is the sampling distribution of the estimator. Of course, we cannot take very many samples to determine the sampling distribution and this is why probability theory is so important – under some assumptions the sampling distribution of an estimator can be derived using the results of probability.

Simon Newcomb took 66 measurements of the time it took light to travel a known distance. The values are shown below. The values are the number of nanoseconds each measurement differed from 0.000024800 seconds. Thus, the first measurement was actually 0.000024828 seconds. Although the speed of light was believed to be constant (at least according to theory) there is variability in the measurements. Denote the true time it takes light to travel the distance in Newcomb's experiment by μ.

Newcomb's measurements of the speed of light data

28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
−44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
−2	24	25	27	24	16
29	20	28	27	39	23

Using the information in the sample requires several assumptions. We assume that μ is the mean of a hypothetical population of all possible measurements. We assume the 66 observed measurements are representative of that population. The measurements are independent of one another in the sense that knowing the value of the i^{th} measurement provides no information about the $(i+1)^{st}$. Stated another way, we could permute the values of the observations in the table without impacting the analysis. We also assume that the population of possible measurements is normally distributed.

This latter assumption is questionable in light of the two outliers -44 and -2. Such atypical values are not unusual in data sets and the question of what to do with them is not easily answered. Here, for simplicity, we omit them and proceed with an analysis of the remaining 64 observations.

The sampling distribution of the sample mean $\overline{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{64}$. The point estimate, i.e. the realized value of the sample mean is $\bar{x} = 27.75$. In a sense

this is our best guess, given the data, about the value of μ. However, there is uncertainty about this estimate because we know the value of the sample mean varies from sample to sample, That is, if Newcomb had taken another 64 observations it is highly unlikely that he would have observed the same value of the sample mean. The variability is quantified by the standard deviation, $\sigma/\sqrt{64}$, which in practice is itself an unknown parameter which must be estimated by the sample standard deviation which is 5.08 for this data set. We will assume, unrealistically, that this, in fact, is the population standard deviation and that $\sigma/\sqrt{64} = 5.08/8 = 0.64$.

We have a point estimate and we have quantified the uncertainty associated with this estimate. We go further however and compute an interval estimate of μ, called a confidence interval. Prior to data collection we know, from the properties of the normal distribution, that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution and that

$$P(-1.96 < Z < 1.96) = P\left(-.196 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

A relatively simple set of algebraic manipulations leads to the conclusion that

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Thus, prior to data collection we know that the random interval $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ will capture or contain μ with probability 0.95. Over the long run, 95% of all such intervals will contain the population mean but 5% will not. When we plug in the values from Newcomb's data we have the realized value of the interval,

$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 27.75 \pm 1.96(5.08)/8 = (26.5, 29.0).$$

The probability that μ lies between 26.5 and 29.0 is either 0 or 1 because μ is a fixed constant. If μ=23 then the probability it lies between 26.5 and 29 is 0. If μ = 28 then the probability it lies between 26.5 and 29 is 1. However, the process that produced the interval is reliable in that over the long run 95% of all intervals produced in this way will contain μ so we say that we are 95% confident that μ lies in the interval (26.5,29.0).

It is possible to compute intervals for other levels of confidence. Let $z_{\alpha/2}$ be the value from a standard normal distribution with $\alpha/2$ of the area under the normal curve lying above it, i.e.

$P(Z > z_{\alpha/2}) = \alpha/2$.

Then a $100(1-\alpha)\%$ confidence interval for the mean μ of a normally distributed population when the value of σ is know is given by

$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$.

The most common confidence levels used in practice are 90% ($z_{0.05} = 1.645$), 95% ($z_{0.025} = 1.96$), and 99% ($z_{0.005} = 2.575$).

Three key assumptions are required for such intervals to be valid:

(i) The data must represent a simple random sample from the population of interest.
(ii) The population must be normally distributed.
(iii) The standard deviation of the population is known.

By far the most important assumption is (i). One implication of the assumption is that observations can be considered independent of one another. This assumption along with assumption (ii) is required for the sampling distribution of $\bar{X}$ to be normal with the indicated mean and variance. It is the normality of $\bar{X}$ that was necessary for the above derivation of the confidence interval procedure. As long as the sample size is large enough the procedure is fairly robust to violations of assumptions (ii) and (iii). If the population is not normal but the sample size is large enough then the Central Limit Theorem will result in an approximate normally distributed $\bar{X}$ and the confidence levels of intervals will be approximately correct. If the population standard deviation is not known then we estimate it with the sample standard deviation. Given a large enough sample size this estimate should be reasonably close to the true standard deviation, close enough that the intervals will have a true confidence level close to the advertised level.

One might wonder why 100% confidence intervals are not computed. The answer is that they are so wide as to be practically useless (a demographer can be 100% confident that the population of the earth is between 0 and one trillion people). Construction of confidence intervals is a balancing act between achieving a high enough level of confidence and a narrow enough interval to be useful. The half-width of a confidence interval is $z_{\alpha/2}\sigma/\sqrt{n}$. Clearly the larger the population standard deviation the wider an interval will be. Investigators generally have little control over the population standard deviation but do have control over the level of confidence and the sample size. However, specification of a higher level of confidence requires use of a larger $z_{\alpha/2}$ leading to a wider interval. Larger sample sizes will lead to narrower intervals. However, collecting samples can be expensive and there will often be limits on how large a sample can be selected. Also, because the half-width of the interval is inversely proportional to the square root of the sample size increasing the sample size by a factor of say *k* does have the same effect of decreasing width. For example, increasing the sample size by 4 times only cuts the interval width in half. It is possible in the simple setting described here to manipulate the

confidence interval formula to determine a sample size needed to achieve a specified width and level of confidence. However, this is more difficult in general.

Confidence intervals are widely used but remain controversial among many statisticians. The interpretation of confidence intervals in particular is difficult for many students and even more sophisticated users to grasp. It is easy to say that one is 95% confident that an interval contains a parameter of interest but when pressed as to exactly what that means many cannot explain it satisfactorily. They want to say, indeed they almost surely believe, that the interval contains the parameter with 95% probability. Some statisticians empathize with such users. They object to the invention of another way of quantifying uncertainty, especially one difficult to understand. To these statisticians the natural language of uncertainty is probability, but they tend to interpret probability differently from frequentists. The approach taken by these statisticians is typically referred to as Bayesian because of the prominent role by Bayes' Rule in their methodology. This is perhaps an unfortunate description because Bayes' Rule works quite well regardless of one's interpretation of probability but it has been used to too long to change now. We do not discuss this approach to interval estimation here as our goal is to present a primer of statistics as it is taught in most introductory courses. However, some of the chapters in this volume will discuss the Bayesian approach in more detail.

6.3. The test of significance and hypothesis testing.

The question of interest in point and interval estimation is, "What is the value of an unknown parameter and how do we quantify our uncertainty about our estimate of the parameter?" Testing is motivated by another question; "How consistent are the data with a stated claim or hypothesis about the value of a parameter?"

The hypothesis takes the form of specifying a particular value or of specifying a range of values within which the true value lies. Implicit in the specification of a hypothesis about the value of a parameter is another hypothesis (the negation of the first) that the value of a parameter is something else. The two hypotheses are called the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis specifies a value or range of values of a parameter and the alternative is the value or range of values deemed plausible if the null hypothesis is not true.

Confusingly there are two basic approaches to testing encountered in the introductory statistical literature; the test of significance and the hypothesis test. The first approach is attributed to the English statistician and population geneticist Sir Ronald A. Fisher. Hypothesis testing is attributed to the Polish mathematician Jerzy Neyman and the English statistician Egon Pearson. Different texts treat the topic in different ways. Devore (2008) discusses Neyman-Pearson hypothesis testing. DeVeaux, Velleman, and Bock (2008) place more emphasis on significance testing although they briefly describe the Neyman-Pearson approach. In general, the discussion of testing in such textbooks takes a middle ground between the two approaches that neither Fisher nor Neyman-Pearson would have approved of. We will start with significance testing and have a little to say about hypothesis testing at the end.

The following simple example illustrates the basic reasoning in a test of significance. An hypothesis is stated along with an alternative. Data are collected and steps are taken to determine if the data are consistent with the null hypothesis. In this example, it is possible that the data are

consistent with the null hypothesis and the decision is to fail to reject the null. If the data are not consistent with the null hypothesis then it might be decided to reject the null. In practice, the data may be equivocal and no decision is warranted. Or it may be that different investigators make a different decision given the same data.

We now describe the procedure for tests of significance about population means. We will assume that we have a simple random sample of size *n* from a normally distributed population with unknown mean μ and variance σ^2 (or standard deviation σ). For definiteness, suppose that another physicist working at the same time as Newcomb believed that the speed of light was such that the mean time it took light to travel the distance in Newcomb's experiment was 29 (using the scaled units from above). We suppose (in our fictional account) that Newcomb did not believe this physicist was correct. Being a scientist he was willing to put his supposition to the test. Further, he was willing to start by assuming that he was wrong and then use the data to show that this assumption is not tenable. This is a classic argument by contradiction. Newcomb did not have the tools to carry out the test that we have but we will see what would have happened if he had. Denoting the true time it takes light to travel the distance in the experiment by μ the null and alternative hypotheses are

$$H_0 : \mu = 29$$
$$H_a : \mu \neq 29.$$

We need a test statistic; a means of quantifying the evidence against the null hypothesis. It is natural to base the test statistic on the sample mean $\overline{X}$. Note that both large and small values of the sample mean would provide evidence against the null and for the alternative. We will use the 64 measurements given above. Recall that we assumed that the standard deviation was known, $\sigma = 5.08$ and that we computed the realized value of the sample mean to be $\bar{x} = 27.75$. Is this consistent with the null hypothesis, i.e. could a value of 27.75 plausibly be observed if the true mean is indeed 29? We know that the sampling distribution of the sample mean $\overline{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n} = 5.08/8 = 0.635$. If the null hypothesis is true then the mean is 29. Thus, probability theory gives us the sampling distribution of the sample mean under an assumption that the null hypothesis is true. We can use this information to see how unusual the observed value of the sample mean is given an assumed population mean of 29. The test statistic commonly used is

$$Z = \frac{\overline{X} - 29}{0.635}.$$

Under the assumptions *Z* will have a standard normal distribution. We compute the value of the test statistic

$$z = \frac{27.75 - 29}{0.635} = -1.97.$$

If the null hypothesis is true then the observed value of the sample mean lies 1.97 standard deviations below the hypothesized mean of 29. In our example, Newcomb did not seem to have a direction in mind initially, i.e. he simply stated his belief that the true mean was not equal to 29. This is an example of a two-sided (or two-tailed) test because as noted above both large and small values of the sample mean would be unusual under the null hypothesis. Of course, any single observation is probabilistically uncertain (with our underlying continuous probability model any single observation has probability 0). We quantify the evidence against the null by computing the probability of observing a value of the test statistic as extreme or more extreme than the one we actually observed COMPUTED UNDER AN ASSUMPTION THAT THE NULL HYPOTHESIS IS TRUE. This probability is referred to as a *P*-value. Given the two-sided nature of the test we compute

$$P(Z \leq -1.97) + P(Z \geq 1.97) = 0.049.$$

This calculation was carried out assuming that Z is a standard normal random variable which is true if the null hypothesis is true. Thus, if the null hypothesis is true we would expect to see a value of the sample mean of 27.75 or something more "extreme" (less than 27.75 or greater than 30.25) only about 4.9% of the time if the experiment were repeated over and over again, independently, and under the same conditions. Is this evidence strong enough to reject the null hypothesis? A frequently used cut-off value between "strong enough" and "not strong enough" is 0.05. If the *P*-value is less than or equal to 0.05 then reject the null hypothesis and if the *P*-value is greater than 0.05 then fail to reject the null hypothesis. If one adheres strictly to this criterion then the evidence would be strong enough to reject.

The selection of a cut-point between reject and fail-to-reject is called fixed-level testing. The cut-point itself is referred to as an α-level. Other commonly chosen levels are $\alpha = 0.01$ and $\alpha = 0.10$. These three values (0.01, 0.05, and 0.10) have become so ingrained in statistical practice that many people assume there is some scientific justification for them but that is not true.

Two different people could come to different conclusions about the results of a test based on their particular selection of a significance level. If Newcomb chose $\alpha = 0.05$ he would reject the null hypothesis but his imaginary opponent could have demanded stronger evidence say by choosing $\alpha = 0.01$in which case he would fail to reject the null hypothesis.

We have seen a specific example of a two-sided test. More generally a two-sided test consists of the following steps.

1) State the null and alternative hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0 .$$

2) Compute the value of the test statistic;

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} .$$

3) Compute the *P*-value,

$$P-value = 2\mathrm{P}\left(Z \geq \left| z_{obs} \right| \right).$$

where Z is assumed to be standard normal and z_{obs} is the observed value of the test statistic.

There are two other one-sided versions of significance tests. The upper-tailed version has null and alternative hypotheses of

$H_0 : \mu \leq \mu_0$
$H_a : \mu > \mu_0$

And the lower-tailed version has null and alternative hypotheses of

$H_0 : \mu \geq \mu_0$
$H_a : \mu < \mu_0$

The test statistic is the same in all cases. The *P*-value for the upper-tailed version is the probability of getting a value of the test statistic equal to or greater than the one actually observed computed under an assumption that the true mean is μ_0 . If the evidence against the null is strong enough to reject the null when the true mean is μ_0 it will be even stronger when $\mu < \mu_0$. Similarly, the *P*-value for the lower-tailed version is the probability of getting a value of the test statistic equal to or less than the one actually observed computed under an assumption that the true mean is μ_0 .

If fixed-level testing is desired then an appropriate significance level is chosen and the decision of whether to reject or fail to reject the null hypothesis is based on a comparison of the *P*-value to the significance level.

Hypothesis testing is carried out in a manner that is mechanically identical to the above. A key difference is that interest centers on controlling the rates at which various errors are made. Hard and fast rules are specified for whether or not to reject the null hypothesis. The probability of rejecting the null hypothesis when $\mu = \mu_0$ is called a Type I error and the significance level is the rate at which that error will occur. The probability of failing to reject the null hypothesis when it is false is called a Type II error. The power of a test is 1 – *P(Type II error)*, In practice an investigator decides how much of a difference between the true mean and the hypothesized value μ_0 he is willing to accept, i.e. he decides how false the null has to be before he needs a high probability of detecting it.

Obviously the ideal hypothesis test will have low rates of both types of errors but specifying a small chance of a Type I error leads to a larger chance of a Type II error. This tension between the two types of errors is frequently misunderstood. An investigator will choose a low level of

significance (a small chance of committing a Type I error), say 1%, failing to understand that the probability of a Type II error can be quite high.

Hypothesis testing is frequently characterized as a decision making process. As such a null hypothesis may be "accepted", not because it is believed to be true but because it is believed to be true enough to justify acting as if it is true. It may make sense in applications in quality control, for example. A shipment of raw material arrives at a plant and it must be accepted or rejected. Type I and Type II error rates are determined based on economic considerations and on purity considerations. It is known they over the long run a certain proportion of acceptable shipments will be rejected and a certain proportion of unacceptable shipments will be accepted but those error rates are economically justifiable.

As mentioned above Fisher is generally credited with inventing significance testing and Neyman and Pearson credited with hypothesis testing. Fisher, who was a scientist, was critical of the Neyman-Pearson approach. He did not concern himself with Type II errors. He did not see science as a decision making process. Neyman-Pearson did not use *P*-values in their original work. They were interested in controlling error rates. The test statistics in tests of population means were sample means themselves. They used the specified values of error rates to find so-called rejection and acceptance regions for values of the sample mean. If an observed value fell in a rejection region the null was rejected. If a value fell in an acceptance region the null was "accepted". Later, it was noticed that the *P*-value could be used to determine if the sample mean was in or out of a rejection region. If the *P*-value was less than the significance level then the value of the sample mean fell in the rejection region and vice versa. But the size of the *P*-value meant little in Neyman-Pearson hypothesis testing. If the significance level was chosen to be 0.05 then a *P*-value of 0.049 meant the same as a *P*-value of 0.00000006, and a *P*-value of 0.0495 would lead to a decision to reject the null while one of 0.0505 would lead to a decision to fail to reject the null. Fisher believed the observed size of the *P*-value was important.

We will not go into the details of the distinction between the two methods of testing. Royall (1997) has a good discussion of the differences. We will also not concern ourselves with the controversies that still surround this topic. Some of those controversies are addressed in chapters to follow. Our goal here has been merely to provide an overview of the main results from probability theory and mathematical statistics to readers who may not have seen this material before or to those who saw it a long time ago.

Also, we have presented the basic concepts of estimation and testing hypothesis about population means and with the unreasonable assumption that the population standard deviation is known. There are many more parameters of interest (variables and standard deviations, correlation coefficients, differences in means, slopes and intercepts of regression lines, etc), and different assumptions are needed in many other cases (the population standard deviation is rarely if ever known in practice). The details differ but the broad concepts presented here remain the same and those are the concepts we wanted to address.

7. Conclusion

It is time to wrap up our primer on probability and statistics. Before we conclude we revisit where we began our journey. We defined a notion called "monotonic reasoning" that underlies

deductive consequence relation which, in fact, captures the core of deductive reasoning. We have seen that a deductive consequence relation is by definition monotonic. The way we have understood monotonic reasoning is this: A relation between a set of sentences and a sentence is monotonic if and only if when it also holds between a set and a sentence, it also holds between any superset of the set and the sentence. We mentioned that probability theory provides tools for both understanding and handling inductive arguments. However, there is a possibility of misunderstanding that whenever the expression" probability" occurs in statistical or philosophical literature, it necessarily means that the context is inductive, and therefore, cannot be monotonic. We will show that there are contexts in which although the expression "probability" occurs they do not have anything to do with inductive inferences. To drive this point home, we make a distinction between "probabilistic inference" and "probabilistic conclusion". By "probabilistic inference," we mean an inference engine that takes statements as inputs and spits out as its output a conclusion where the inference itself is probabilistic. All inductive inferences are probabilistic inferences in this sense. Consequently, they are non-monotonic. We mean by "probabilistic conclusion" an argument whose conclusion contains a probabilistic statement. An argument with a probabilistic conclusion need not be inductive. Hence, it need not involve non-monotonic reasoning.

In the case of an inference involving probability, its conclusion could be a categorical statement in which the conclusion does not follow necessarily from the premises of the argument. Since it is concerned with inference itself the inference is sometimes called "uncertain inference." However, in other cases of inference, the conclusion of that inference could be a probabilistic statement, but it could follow deductively from its premises. It is a case of inference about uncertainty (See, Hempel, 1965; and Kyburg, 2000). We get four possible combinations between probabilistic conclusion and probabilistic inference. They are as follows:

Probabilistic conclusion, but not probabilistic inference.
Probabilistic inference, but no probabilistic conclusion.
Probabilistic inference and probabilistic conclusion. And
Neither probabilistic conclusion nor probabilistic inference.

(i) Probabilistic conclusion, but not a case of probabilistic inference.

P1: This die has six faces, labeled 1, 2, 3, 4, 5, 6.
P2: Each face is equally probable.
C: The probability of rolling a 3 on this die is 1/6

The conclusion makes a statement about the probability of a certain event and the conclusion follows deductively from the premises. Given the probability calculus, it is a deductively valid argument in which if the premises are true, then the conclusion must be true, too. If we add a new premise to this argument then we won't be able to undermine its deductive validity. Therefore, it involves monotonic reasoning. At this point, we want to mention that Simpson's paradox which is a well-known paradox in probability and statistics in fact falls under this category. Simpson's Paradox involves the reversal of the direction of a comparison or the cessation of an association when data from several groups are combined to form a single whole (see the "Introduction" together with Hájek's paper in the volume for this position). This shows that an argument with a probabilistic conclusion does not necessarily make it a probabilistic inference; therefore, the argument in question does not stand for an inductive inference.

(ii). Probabilistic inference, but no probabilistic conclusion.

P1. 95% of the balls in the urn are red.
P2: A ball is randomly selected from the urn.
C: This ball drawn will be red.

The conclusion does not follow deductively from its premises, although it is a categorical statement about a particular ball, which may be false. Given a long-run relative frequency interpretation of probability we expect to see a red ball about 95% of the time with repeated draws. We might be willing to bet a large sum of money that a particular draw will be red, but we cannot state this unequivocally. It will be fair to say that even though the premises are true, the conclusion could be very well false. Hence, the inference is not deductive. As a result, the reasoning involved is non-monotonic and therefore, this is an example of inductive inference.

(iii) Probabilistic inference and probabilistic conclusion

The following is an example in which the inference is probabilistic as well as its conclusion.

P1. The die has six faces, labeled 1, 2, 3, 4, 5, 6
P2. In a sequence of 250 rolls, a 3 was rolled 50 times.
C: The probability of rolling a 6 with this die is about 1/5.

In this situation, we happen to observe 50 threes in 250 rolls of a fair die. The probability of obtaining a three on a single roll is 1/5 (10/50). It is an instance of inductive inference since even though the premises are true the conclusion could be false. Therefore, the reasoning the argument employs is non-monotonic. To understand the reasoning involved in (iii) compare (iii) with (i). In (iii) we have a specific outcome of a random process and we are reasoning from the sample to the population. In this sense, we are going from the specific to the general. In (i), however, we are reasoning from the population process to the behavior of samples over the long run. Here, we are making an inference about a specific outcome from a general claim.

(iv) Neither probabilistic conclusion nor probabilistic inference.

Consider an example which does not involve probabilistic inference nor does it involve probabilistic conclusion.

P1. Wherever there is smoke, there is fire.
P2: The hill has smoke.
C: There is fire on the hill.

Neither its premises nor its conclusion contain a probabilistic claim. This example is a case of non-probabilistic inference (deductive argument) with a non-probabilistic (categorical) conclusion. Therefore, from the perspective of our probabilistic curiosity, it becomes uninteresting. Although this specific argument is deductively valid and therefore, involves monotonic reasoning, any example under this rubric need not be deductively valid. Since our present interest lies in probabilistic reasoning, an example of this specific argument is of no concern for us.

This ends our primer where we have discussed basic ideas and theories behind probability and statistics. We have also discussed some of the fundamental differences between inductive and deductive arguments. We hope that this will provide a background for the general reader to read and appreciate many of the papers in the volume.

Bibliography

DeVeaux, R., P. Vellman., and D. Bock (2008): *Stats*: *Data and Models, Second Edition.* Pearson-Addison Wesley, New York, NY. USA.

Devore, J (2008): *Probability and Statistics for Engineering and the Sciences, Seventh Edition.* Brooks-Cole, Belmont, CA. USA.

Hájek, A (2007): "Interpretations of Probability" in *Stanford Encyclopedia of Philosophy*

Hempel, C. G. (1965): *Aspects of Scientific Explanation*. New York: Free Press.

Kyburg, H (2000): "Probable Inference and Probable Conclusion." In *Science, Explanation, and Rationality*. J. Fetzer (ed.) Oxford University Press., New York: New York

Moore, D., and G. McCabe (2006): *An Introduction to the Practice of Statistics, Fifth Edition.* W.H. Freeman & Company, New York, NY. USA.

Royall, R. (1997): *Statistical Evidence*: *A likelihood paradigm*. Chapman & Hall: London.

Stigler, S (1986): *History of Statistics*: *The Measurement of Uncertainty Before 1900*. Harvard University Press: Cambridge: Mass.