Smoothing the Transition from Data to Statistical Inference

Joseph B. Lang[1]

To curb the human appetite for jumping to unwarranted data-based conclusions about a target of inference, statistics courses set out to formalize the transition from data to statistical inference. Unfortunately, this transition typically unfolds in a series of fits and starts. This paper sets out to rectify this situation. Instead of following the standard sequence of seemingly unrelated topics (descriptive statistics, probability, random variables, sampling distributions, ...), this paper argues that it is better to introduce formal links between data and the targets of inference early on in any elementary course. To accomplish this, an explicit "colon notation" is used to clearly compare and contrast objects based on samples, populations, and processes. The colon notation, along with the important concepts of sample, population, and process distributions, allows the instructor to present linking results, such as laws of large numbers and "the fundamental theorem of statistics," within the first few class periods. These data-target links lead naturally to well-formulated questions about target populations and processes, and serve as a conduit to statistical inference. Equipped with the concepts already used to formulate the data-target links, the instructor can highlight the basic ideas underlying statistical inference earlier in the course. This leaves more time to cover other important statistical concepts in a survey course.

Key words: Colon Notation; The Fundamental Theorem of Statistics; Linking Data to Targets of Inference; Sample, Population, and Process Distributions; Statistics Education.

1 Introduction

The transition from data to *informal* statistical inference is a smooth and natural one that occurs automatically in the human brain. These informal inferences can be useful on occasion, but too often they are baseless owing to a lack of any formal link between the data and the perceived target of inference, typically a population or process. Of course there are other reasons that we are led astray with informal, automatic inferences (cf. Kahneman, 2011), but this paper focuses on the "data-target link" issue. To curb the human appetite for jumping to unwarranted data-based conclusions about a target of inference, statistics courses set out to formalize the transition from data to statistical inference. Unfortunately, this formal transition is not always a smooth one, especially from the perspective of post-secondary students in an elementary survey course. The vast literature on statistics education makes this abundantly clear (cf. delMas et al. 1994, Moore et al. 1995, Garfield and Ben-Zvi 2007, Thompson et al. 2007, Wessels and Nieuwoudt 2013, and references therein). To the student, the transition unfolds in a series of fits and starts. It is a march through a sequence of seemingly unrelated topics with the apparent final goal of learning how to use formulas for t tests and confidence intervals. The student learns about descriptive statistics, then abandons this topic and abruptly changes course to learn about probability and random variables. The student is then told about the importance of sampling distributions for statistical inference and is bombarded with confusing statements such as "the sample mean has an approximate Normal distribution," or

[1]Joseph B. Lang is a professor in the Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242 (email: joseph-lang@uiowa.edu). March 27, 2014.

"the mean of the sample mean is equal to the population mean." Having just computed a sample mean of $\overline{x} = 105.8$, the student is justifiably perplexed by such statements. Finally, the student is introduced to the wonderfully mysterious formulas of hypothesis testing and confidence intervals. In applying these formulas, students find themselves alternating between treating symbols such as $\overline{x}$ and $\sqrt{n}(\overline{x} - \mu_0)/s$ as observed values and random variables. From the student's perspective, this nebulous, shifting treatment of symbols is the extent to which data and statistical inference are linked.

The current paper sets out to smooth the formal transition from data to statistical inference. Instead of following the standard sequence of disjoint topics, we argue that it is better to introduce formal links between data and the targets of inference early on in the course. In particular, the important concept of a process (aka random experiment), along with suggestive notation, allows the instructor to introduce linking results such as the laws of large numbers and "the fundamental theorem of statistics" within the first few class periods. Our presentation stresses that the validity of these linking results depends critically on the sample generation or selection process. These data-target links lead naturally to well-formulated questions about target populations and processes, and serve as a conduit to statistical inference. Equipped with the concepts already used to formulate the data-target links, the instructor can highlight the basic ideas underlying statistical inference earlier in the course. This leaves more time to cover other important statistical concepts in a survey course (see for example, concepts listed in the American Statistical Association's GAISE report, 2012).

Space limitations preclude an exhaustive treatment of all topics encountered on the journey from data to statistical inference. Instead, we highlight only the main concepts, especially those that are best described using less conventional, more explicit, notations. Whereas many of the concepts covered herein are not new (they are discussed in any introductory statistics textbook), the ordering and emphases are different and the presentation approach is novel. We introduce a "colon notation" that highlights differences and commonalities among data (X:$\underline{s}$), populations (X:$\underline{P}$), and processes (X:RE). Confusion about statistical inference concepts often stems from the conflation of the three distinct distribution types: the sample, population, and process distribution of a variable (cf. delMas et al. 2004). The colon notation mitigates such conflation problems by promoting an explicit accounting of these three distribution types and summaries thereof. For example, the sample, population, and process mean of X are denoted $mean(X{:}\underline{s})$, $mean(X{:}\underline{P})$, and $mean(X{:}RE)$, rather than the more conventional $\overline{x}$, μ, and μ.

Most elementary statistics textbooks restrict attention to inferences about populations. This usually forces the instructor and students to confront awkward notions such as "random sampling from an infinite population" and targets of inference described as "the population mean of an infinite number of rolls of a die." Herein, we accommodate inferences about both populations and processes, and clearly distinguish between the two. This is arguably a more natural approach (cf. Frick 1998), which leads to clearer definitions of inference targets and avoids discussions of infinite populations.

Process-related concepts such as process outcomes and process distributions are common threads underlying many of the ideas in this paper. Data and variable values, such as the observed sample mean, are process outcomes, and inferences are typically based on process distributions of variables.

For emphasis, we introduce process outcome notation and explicitly refer to the underlying process when describing a process distribution for statistical inference. On a related note, the reader will notice that we mostly avoid using the phrase "sampling distribution" in this paper. Section 6.1 gives several reasons for this avoidance.

The balance of this paper gives more details on the topics described above. Section 2 gives a definition of data and variables and introduces the colon notation. Section 3 describes and compares the ultimate goals of descriptive and inferential statistics. Population and process targets of inference are defined. The three distribution types and corresponding summaries are described in Section 4. Section 5 gives results that formally link data and targets of inference. These results, which include the fundamental theorem of statistics and laws of large numbers, are given in a form that can be presented within the first few lectures of an elementary course. In particular, there is no need to cover topics such as IID random variables. Section 6 represents a prelude to statistical inference. Process distributions and the Central Limit Theorem along with variants, such as the Studentized Mean Approximation, are described. For economy of space, this section touches on interval estimation of a process or population mean, but does not delve into the important topic of statistical tests of hypotheses. Section 7 gives two canonical examples with questions that can be addressed using the ideas in this paper, and Section 8 gives a brief discussion. Finally Appendix A gives details on a general, less elementary, version of the fundamental theorem of statistics, Appendix B gives some figures that correspond to several of the concepts and notations introduced in this paper, and Appendix C gives detailed solutions to the examples of Section 7.

2 Data and Variables

Arguably, data is the most important ingredient in statistics. We use data to describe samples and to make inferences about populations and processes. We use data to make informed decisions in the face of uncertainty and we use it to bolster or discredit an argument. On the computational side, we summarize, tabulate, graph, munge, and mine data. Because data feature so prominently in statistics (a discipline also known as data science), it is important to have a clear definition and a useful notation.

Technically, data are characteristic values for a collection of entities, such as a sample of people or outcomes of a random experiment. Hair color is a characteristic; hair colors for a sample of people are data. Grade point average (GPA) is a characteristic; GPA values for a sample of undergraduates are data. The number of up-dots is a characteristic; the number of up-dots on each of five rolls of a die are data. A variable in statistics is a formal or symbolic representation of a characteristic (cf. Mcnaughton, 2002). For example, the variable X might be used to represent the characteristic hair color. For convenience, we follow convention and use the terms 'characteristic' and 'variable' interchangeably. More formally, a variable in statistics, as opposed to mathematics, can be conveniently viewed as a function that maps an entity to a characteristic value.[2]

[2]An entity is anything that can be described, such as a person, a sample of people, a place, a process outcome, or a scenario. Characteristic values may be numbers or they may be other more exotic objects such as matrices, word

In an elementary course, it is usually sufficient to restrict attention to data that can be viewed as a collection of X or (X, Y) values for a sample of entities. The upper case letters X and Y are variables (equivalently, characteristics) that measure the entities in the sample. Such data can be represented using the "colon" notation $X{:}\underline{s}$ and $(X, Y){:}\underline{s}$. For example, $X{:}\underline{s}$ (pronounced "X of $\underline{s}$") is the collection of X values for the entities in the sample $\underline{s} = (s_1, \ldots, s_n)$. Formally,

$$X{:}\underline{s} \equiv X{:}(s_1, \ldots, s_n) \equiv (X(s_1), X(s_2), \ldots, X(s_n)),$$

where $X(s_i)$ is the X value for entity s_i, see Figure 1 in Appendix B. Similarly,

$$(X, Y){:}\underline{s} = ((X(s_1), Y(s_1)),\ (X(s_2), Y(s_2)),\ \ldots,\ (X(s_n), Y(s_n)))$$

where the pair $(X(s_i), Y(s_i))$ give the X and Y values for entity s_i.

This suggestive colon notation has several advantages over the more conventional notation for data, e.g. $\underline{x} = (x_1, \ldots, x_n)$. The symbol $X{:}\underline{s}$ reminds us that the data are X values for the sample $\underline{s}$. Clearly, the GPA values for a sample of 25 female students, say $GPA{:}\underline{s}_f$, is different from the GPA values for a sample of 25 male students, say $GPA{:}\underline{s}_m$. We could use $\underline{x}$ and $\underline{y}$ for these two data sets and remind students that they are GPA values for two different samples, but the context is quickly lost and the "little x's and little y's" quickly become "little x's and little y's" and nothing more. Similarly, $GPA{:}\underline{s}_f$ and $AGE{:}\underline{s}_f$ are different data sets that can be used to describe the same sample of 25 females, $\underline{s}_f$. With $\underline{x}$ and $\underline{y}$ notation, the student will quickly forget that both data sets measure the same sample. The colon notation forces the reader to take note of both the variable used and the collection of entities that it is measuring. This notation is designed to prompt questions about what sample is being measured, how the sample was selected or generated, and how the variable is actually defined.

3 Descriptive versus Inferential Statistics

3.1 Descriptive Statistics

The ultimate goal of descriptive statistics is to describe a sample $\underline{s}$ using data, say $X{:}\underline{s}$ or $(X, Y){:}\underline{s}$. To "use data" for such a description, we must in turn describe or summarize the data itself. We make two important observations: (1) The ultimate goal of descriptive statistics is *not* to describe data; rather it is to describe a sample $\underline{s}$ using data. (2) A description of a sample $\underline{s}$ based on data $X{:}\underline{s}$ or $(X, Y){:}\underline{s}$ is necessarily incomplete because, besides X and Y, there are infinitely many other characteristics of the entities in $\underline{s}$. Observation (1) highlights the preeminent role of the sample $\underline{s}$ in descriptive statistics and Observation (2) highlights the incompleteness of any statistical description and opens the door to competing descriptions of the same sample $\underline{s}$.

It is not immediately clear how to describe or summarize data $X{:}\underline{s}$, especially when the sample is large. A useful starting point is the idea of 'distribution' of data $X{:}\underline{s}$, denoted $dist(X{:}\underline{s})$. Technically,

descriptors ($X(e) = red$), intervals ($X(e) = [96.2, 99.8]$), or even functions ($X(e) = f(\cdot; z(e))$, where $f(u; z(e)) \equiv u^{z(e)}$). In this paper, unless otherwise stated, we focus on scalar-valued variables.

$dist(X{:}\underline{s})$ is the collection of distinct values of $X{:}\underline{s}$ along with their relative frequencies. For example, if $X{:}\underline{s} = (5, 3, 4, 5, 4, 4)$ then the collection of distinct values $(3, 4, 5)$ along with the corresponding relative frequencies $(\frac{1}{6}, \frac{3}{6}, \frac{2}{6})$ gives the distribution $dist(X{:}\underline{s})$. The distribution is a useful summary because it, or a coarser summary thereof, can be graphically depicted using a bar graph, a histogram, or a density plot, for example. Equipped with distributions, we can graphically compare multiple data sets, say $X{:}\underline{s}_1$, $X{:}\underline{s}_2$,... $X{:}\underline{s}_K$, and hence graphically compare samples $\underline{s}_1, \underline{s}_2, \ldots, \underline{s}_K$ based on variable X.

Simpler, but less fundamental summaries of data $X{:}\underline{s}$, including sample means, variances, and quantiles, are available when X is scalar-valued. For example, the mean of X for sample $\underline{s}$, denoted $mean(X{:}\underline{s})$, is defined as

$$mean(X{:}\underline{s}) = \frac{X(s_1) + X(s_2) + \cdots + X(s_n)}{n} = \frac{\text{sum of } X \text{ values for } \underline{s}}{\text{number of entities in } \underline{s}}.$$

Two samples, say $\underline{s}_1$ and $\underline{s}_2$, can be compared on the basis of $mean(X{:}\underline{s}_1)$ and $mean(X{:}\underline{s}_2)$. This comparison is clearly incomplete because it measures the entities in $\underline{s}_1$ and $\underline{s}_2$ using only the variable (characteristic) X and it uses only a measure of centrality to describe the data sets $X{:}\underline{s}_1$ and $X{:}\underline{s}_2$.

The typical elementary statistics course goes on to describe many other graphical and numerical summaries of data $X{:}\underline{s}$ and $(X, Y){:}\underline{s}$. We will not pursue this topic any further herein. Instead we close this subsection by pointing out the utility of using notations such as $X{:}\underline{s}$ and $mean(X{:}\underline{s})$, rather than the more conventional $\underline{x}$ and $\overline{x}$. The former reminds the student that it is the sample $\underline{s}$ that is being described using the variable X. The explicit reference to the sample $\underline{s}$ in this notation stresses its preeminence in descriptive statistics and serves to curb the urge to misinterpret a summary such as $mean(X{:}\underline{s})$ as a measure of some larger collection of units or a process.

3.2 Inferential Statistics

The ultimate goal of inferential statistics is to use a sample $\underline{s}$ to reduce uncertainty about an incompletely-observed population or an imperfectly-described process. A population, denoted $\underline{P} = (e_1, \ldots, e_N)$, is a finite collection of all entities of interest. Typically N, the size of the population $\underline{P}$, is so large that it is not feasible to measure or even observe all N entities. A process is generically a sequence of steps that produces outcomes. At some level of precision, process outcomes are not perfectly predictable; that is, a process can never be perfectly described and generally the intrinsic randomness cannot be eliminated. For this reason, we can always view a process as a random experiment and denote it by a symbol such as RE. In this document, we shall use the terms 'process' and 'random experiment' interchangeably. Examples of processes include simple random sampling, rolling a die once, rolling a die n times, the manufacturing of a product, or Nature generating the weather, a financial scenario, or a patient scenario.

Inference about a population $\underline{P}$ is typically simplified by using data such as $X{:}\underline{s}$ to answer questions about a more focused target such as $X{:}\underline{P}$. The X population, $X{:}\underline{P}$ (pronounced "X of $\underline{P}$"), is the collection of X values for all the entities in $\underline{P}$; i.e.

$$X{:}\underline{P} = (X(e_1), X(e_2), \ldots, X(e_N)), \qquad \text{see Figure 1 in Appendix B.}$$

In this population setting, the sample $\underline{s} = (s_1, \ldots, s_n)$ can be thought of as a "subset"[3] of the population $\underline{P} = (e_1, \ldots, e_N)$. The collection $X{:}\underline{P}$ is viewed as a fixed (non-random) collection of N values. The outstanding question is, How can we "use data" $X{:}\underline{s}$ to make inferences about the population of values $X{:}\underline{P}$? It should be clear that, among other things, the answer depends on how we went about choosing the sample $\underline{s}$ from $\underline{P}$. For example, suppose that $X(e) = 1$ or 0 as person e is for or against a law that keeps minors out of bars after 11pm and suppose that $\underline{P}$ is the collection of voting age citizens of Iowa City. On the one hand, if $\underline{s}$ is a sample of n university freshman, it is not clear what $X{:}\underline{s}$ can tell us about $X{:}\underline{P}$, even if n is very large. On the other hand, if $\underline{s}$ is a simple random sample of size n from $\underline{P}$, then it seems that $X{:}\underline{s}$ should tell us something about $X{:}\underline{P}$, especially if n is large.

Inference about a process RE is typically simplified by using data such as $X{:}\underline{s}$ to answer questions about a more focused target such as $X{:}RE$ (pronounced "X of RE"). The process $X{:}RE$ is a random experiment with action-based description (see also Figure 2 in Appendix B),

$X{:}RE$ = "random experiment RE is carried out and the X value of the outcome is reported."

The variable X in the process $X{:}RE$ is conventionally called a random variable because its value depends on the unpredictable outcome of the random experiment RE. In this process setting, the sample $\underline{s} = (s_1, \ldots, s_n)$ can be viewed as an outcome of a sample-generating random experiment RE^s. The outstanding question is, How can we "use data" $X{:}\underline{s}$ to make inferences about the process $X{:}RE$? It should be clear that, among other things, the answer depends on whether and how RE^s is related to RE. For example, suppose that X is the number of up-dots on a die. On the one hand, if RE ="roll this 6-sided die" and RE^s = "roll a different 10-sided die n times," then it is not clear what $X{:}\underline{s}$ can tell us about $X{:}RE$, even if n is very large. On the other hand, if RE^s = "roll this 6-sided die n times," that is, RE^s = "repeat RE n times," then it seems that $X{:}\underline{s}$ should tell us something about the process $X{:}RE$, especially if n is large.

In summary, our simplified goal of inferential statistics is to use data $X{:}\underline{s}$ to reduce uncertainty about the population $X{:}\underline{P}$ or process $X{:}RE$. (This paper focuses on univariate data $X{:}\underline{s}$, but the ideas extend readily to inferences based on multivariate data such as $(X, Y){:}\underline{s}$.) The first step toward realizing this goal is to formally link the data to the population or process. To present these formal links, as in Section 5, we must first understand the three distributions of a variable X.

4 Distributions of Variables

4.1 The Three Distribution Types

Operationally, to address the simplified goal of statistical inference of the previous section, we might use the *distribution* of $X{:}\underline{s}$ to reduce uncertainty about the *distribution* of $X{:}\underline{P}$ or the *distribution* of $X{:}RE$. This last statement correctly hints at the need to consider three distinct distributions

[3]Strictly speaking, $\underline{s}$ is an ordered n-tuple that can have repeat values, so it could technically include every value in $\underline{P}$, at least once; i.e. it need not be a "subset" as defined in set theory. Nonetheless, in practice, it generally does not have repeat values and it is a proper subset of $\underline{P}$, with $n < N$, so we will continue to use the term "subset."

for variable X, the sample distribution $dist(X{:}\underline{s})$, the population distribution $dist(X{:}\underline{P})$, and the process (or probability) distribution $dist(X{:}RE)$. Much confusion in the mind of the student of elementary statistics stems from the conflation of these three distribution types.

We previously defined the sample distribution $dist(X{:}\underline{s})$. Here we re-iterate that definition and give definitions of the other two distributions:

1. $dist(X{:}\underline{s})$ is the distribution of variable X for sample $\underline{s}$, or the sample distribution of X. This sample distribution is characterized by the collection of distinct values in $X{:}\underline{s}$ and their relative frequencies. The sample distribution $dist(X{:}\underline{s})$ can be characterized by sample proportions of the form $prop_{\underline{s}}(X \in A)$, over all A. Here $prop_{\underline{s}}(X \in A)$ is the proportion of entities in $\underline{s}$ that have X values in the set A.[4]

2. $dist(X{:}\underline{P})$ is the distribution of variable X for population $\underline{P}$, or the population distribution of X. This population distribution is characterized by the collection of distinct values in $X{:}\underline{P}$ and their relative frequencies. The population distribution $dist(X{:}\underline{P})$ can be characterized by population proportions of the form $prop_{\underline{P}}(X \in A)$, over all A. Here, $prop_{\underline{P}}(X \in A)$ is the proportion of entities in $\underline{P}$ that have X values in the set A.

3. $dist(X{:}RE)$ is the distribution of variable X for [or wrt] process RE, or the process distribution of X. This process (or probability) distribution is characterized by probabilities of the form $P_{RE}(X \in A)$, over all A. Here, $P_{RE}(X \in A)$ is the probability that RE will generate an outcome that has X value in the set A.[5]

Remark. Taking the formal view of variable X as a function, it makes sense to take $(X \in A)$ as an inverse image notation with definition, $(X \in A) = \{e : X(e) \in A\}$. Then, for example, the proportion and probability notations can be viewed as having the generic forms $prop_{\underline{s}}(\{set\})$, which is the proportion of entities in $\underline{s}$ that fall in $\{set\}$, and $P_{RE}(\{set\})$, which is the probability that RE will generate an outcome in $\{set\}$.

4.2 Distribution Summaries

Statistical inference can be even more focused. Rather than using $dist(X{:}\underline{s})$ to reduce uncertainty about $dist(X{:}\underline{P})$ or $dist(X{:}RE)$, we might use some other, less fundamental, summary of $X{:}\underline{s}$ to reduce uncertainty about a summary of $X{:}\underline{P}$ or a summary of $X{:}RE$. Examples of these less fundamental summaries include the mean, median, variance, and standard deviation. As an example, consider the sample, population, and process means,

$$mean(X{:}\underline{s}) = \frac{X(s_1) + \cdots + X(s_n)}{n} = \frac{\text{sum of } X \text{ values for } \underline{s}}{\text{number of entities in } \underline{s}},$$

[4] In other symbols, $prop_{\underline{s}}(X \in A) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(X(s_i) \in A)$ or $prop_{\underline{s}}(X \in A) = n^{-1} freq_{\underline{s}}(X \in A)$. As an example, if $X{:}\underline{s} = (3, 1, 1, 3, 2, 3)$ then for instance $prop_{\underline{s}}(X = 3) = 3/6$ and $prop_{\underline{s}}(X < 2) = 2/6$.

[5] By the definition of the process $X{:}RE$, the probability $P_{RE}(X \in A)$ is identical to $P_{X:RE}(A)$, the probability that $X{:}RE$ will generate an outcome in the set A. In probability theory, $P_{X:RE}$ is called an induced probability function (cf. Resnick, pp 74-5, 1998). This technical detail is mentioned because it shows the utility of the process notation such as $X{:}RE$.

$$mean(X{:}\underline{P}) = \frac{X(e_1) + \cdots + X(e_N)}{N} = \frac{\text{sum of } X \text{ values for } \underline{P}}{\text{number of entities in } \underline{P}},$$

$$mean(X{:}RE) = \int X(e) dP_{RE}(e).$$

The process mean $mean(X{:}RE)$ is defined as the Lebesgue-Stieltjes integral that gives a probability weighted average of possible X values. There are simple ways to compute this integral for commonly encountered processes $X{:}RE$. For example, the integral can be computed as a sum or a Riemann integral when $P_{X:RE}$ admits a density with respect to certain measures (cf. Resnick, Chapter 5, 1998). Even in these simpler settings, an elementary course should emphasize interpretation rather than computation. Fortunately, the formal link results of Section 5.4 below provide a simple empirical interpretation that can be presented before any discussion of computational formulas.

Conventionally, the simpler, but less informative, symbols $\overline{x}$, μ, and μ have been used to represent $mean(X{:}\underline{s})$, $mean(X{:}\underline{P})$, and $mean(X{:}RE)$. The colon notation helps clarify which distribution or summary is being referred to, which is very important if we are to explain confusing statements such as, "the mean of the sample mean is equal to the population mean" or "the mean of the sample has an approximate Normal distribution." Unfortunately, the experienced instructor of elementary statistics is generally so comfortable with such statements that he or she may not see how confusing they are to the student, or recall how confusing they were to him or her.

The variance and standard deviation are two other important summaries that can be computed using the distribution alone. The variances are defined as follows:

$$var(X{:}\underline{s}) = \frac{(X(s_1) - mean(X{:}\underline{s}))^2 + \cdots + (X(s_n) - mean(X{:}\underline{s}))^2}{n-1},$$

$$var(X{:}\underline{P}) = \frac{(X(e_1) - mean(X{:}\underline{P}))^2 + \cdots + (X(e_N) - mean(X{:}\underline{P}))^2}{N},$$

$$var(X{:}RE) = \int (X(e) - mean(X{:}RE))^2 dP_{RE}(e).$$

The process variance $var(X{:}RE)$ is defined as the Lebesgue-Stieltjes integral that gives a probability weighted average of possible $(X - mean(X{:}RE))^2$ values. As with the process mean, there are simple ways to compute this integral for commonly encountered processes $X{:}RE$. Regardless of how it is computed, the formal link results of Section 5.4 give this process variance a simple empirical interpretation.

Remark: Conventionally, the simpler, but less informative, symbols s^2, σ^2, and σ^2 have been used to represent $var(X{:}\underline{s})$, $var(X{:}\underline{P})$, and $var(X{:}RE)$.

The sample, population, and process standard deviations are defined to be the square roots of the corresponding variances. That is, $sd(X{:}\underline{s}) = \sqrt{var(X{:}\underline{s})}$, $sd(X{:}\underline{P}) = \sqrt{var(X{:}\underline{P})}$ and $sd(X{:}RE) = \sqrt{var(X{:}RE)}$. The process standard deviation, $sd(X{:}RE)$, like the process variance, has a simple empirical interpretation, which is described in Section 5.4.

Of course there are infinitely many other summaries that could be computed. For example, we could consider $median(X{:}\underline{s})$, $median(X{:}\underline{P})$, and $median(X{:}RE)$, or other quantiles. For convenience, the current paper will focus on the mean, variance, and standard deviation.

5 Formal Links between Data and a Population or Process

Recall that our simplified goal of inferential statistics is to use data such as $X{:}\underline{s}$ to reduce uncertainty about a population such as $X{:}\underline{P}$ or a process $X{:}RE$. As an initial step toward this end, the fundamental results in this section can be used to formally link data to a population or process of interest. In particular, these links show that when the sample $\underline{s}$ is generated or selected in a very special way, data such as $X{:}\underline{s}$ can be used to approximate certain aspects of the population $X{:}\underline{P}$ or process $X{:}RE$.

5.1 Samples and Data as Special-Case Process Outcomes

In an elementary statistics course, the fundamental results that link data to a population or process are most easily illustrated by restricting attention to special sample selection/generation methods. For inference about a population $\underline{P}$, we will restrict attention to the special case where the sample $\underline{s} = (s_1, \ldots, s_n)$ can be viewed as an outcome of the sampling process $SRS(n, \underline{P})$, which denotes a simple random sample of size n, taken with replacement, from the population $\underline{P}$. This viewpoint will be represented using the "process-outcome" notation

$$\underline{s} = (s_1, \ldots, s_n) \leftarrow SRS(n, \underline{P}).$$

It is important to keep in mind that $SRS(n, \underline{P})$ is a sampling process, which is a special random experiment that is chosen and carried out by the researcher. (At this point, the instructor could give a more careful definition of a SRS and compare and contrast it to other probability and non-probability sampling processes. One important bit about the $SRS(n, \underline{P})$ is that each entity in $\underline{P}$ has the same chance of being selected.)

For inference about a process RE, we will restrict attention to the special case where the sample $\underline{s} = (s_1, \ldots, s_n)$ can be viewed as comprising the outcomes of n independent[6] replications of RE; that is, $\underline{s} = (s_1, \ldots, s_n)$ is an outcome of a random experiment of the form $RE^s = RE(n)$, where this replicate-process notation is defined as

$$RE(n) = \text{"Random experiment } RE \text{ is replicated } n \text{ times."}$$

Using the process-outcome notation, $\underline{s}$ is viewed as $\underline{s} = (s_1, \ldots, s_n) \leftarrow RE(n)$. This implies that the i^{th} entity in $\underline{s}$, s_i, is the outcome of the i^{th} replicate of RE. Clearly the sample-generating process $RE^s = RE(n)$ is related to RE, the process of interest.

In summary, the fundamental linking results are most easily illustrated by assuming that $\underline{s} \leftarrow SRS(n, \underline{P})$ and $\underline{s} \leftarrow RE(n)$, for population and process inference, respectively. Because $SRS(n, \underline{P})$ can be viewed as n replicates of $SRS(1, \underline{P})$, in symbols $SRS(n, \underline{P}) = SRS(1, \underline{P})(n)$, these assumptions can be combined and stated more simply as $\underline{s} \leftarrow RE(n)$, where for population inference, $RE = SRS(1, \underline{P})$ and for process inference, RE is the process of interest. Here, the "process outcome" notation tells us that the sample $\underline{s}$ is viewed as an outcome of a process (aka

[6]The adjective "independent" is arguably unnecessary here because replicating RE means that it must not depend on the outcomes of any of the other replicates. If it did then it would not be a replicate.

random experiment). It follows that the data $X{:}\underline{s}$ can be viewed as an outcome of a process that depends on $SRS(n, \underline{P})$ or $RE(n)$, see Figure 3 in Appendix B. It is stressed that repeating a process, of course, would result in a different outcome (a different sample and different data).

Remark: The process-outcome notation can be used more generally. For example, $Y(\underline{s}) \leftarrow Y{:}RE(n)$ tells us that $Y(\underline{s})$ is an outcome of the process or random experiment $Y{:}RE(n)$. In this case, it is also common to refer to $Y(\underline{s})$ as a realization of random variable Y [wrt $RE(n)$].

5.2 Intuitive Population-Process Links

The "population versions" of the fundamental concepts described below make use of the following intuitive result:

$$P_{SRS(1,\underline{P})}(X \in A) = prop_{\underline{P}}(X \in A), \quad \text{for any } A, \tag{1}$$

where the LHS is the probability that $SRS(1, \underline{P})$ generates an outcome with X value in A and the RHS is the proportion of the entities in $\underline{P}$ with X values in A. To motivate this equality, picture a bag of 100 marbles, 25 of which are red and 75 are blue, and let X be the color variable. Here the population $\underline{P}$ is the collection of $N = 100$ marbles. The equality $P_{SRS(1,\underline{P})}(X = red) = prop_{\underline{P}}(X = red) = 0.25$ is the symbolic equivalent of the following: The probability that a simple random sample of size 1 from $\underline{P}$ will result in a marble that is red is 0.25, which is identical to the proportion of marbles in $\underline{P}$ that are red.

Because the process distribution $dist(X{:}RE)$ is characterized by the probabilities $P_{RE}(X \in A)$, for all A, and the population distribution $dist(X{:}\underline{P})$ is characterized by the proportions $prop_{\underline{P}}(X \in A)$, for all A, the result in (1) implies that

$$dist(X{:}SRS(1, \underline{P})) = dist(X{:}\underline{P}). \tag{2}$$

This result equates a special-case process (or probability) distribution and a population distribution. It also implies that summaries that can be computed using the distributions alone are equal; for example, $mean(X{:}SRS(1, \underline{P})) = mean(X{:}\underline{P})$ and $var(X{:}SRS(1, \underline{P})) = var(X{:}\underline{P})$.

5.3 *The* Fundamental Theorem of Statistics

Arguably one of the most fundamental results in statistics is that, when the sample $\underline{s}$ is generated or selected in a very special way, the sample distribution $dist(X{:}\underline{s})$ will resemble the corresponding process distribution $dist(X{:}RE)$ or population distribution $dist(X{:}\underline{P})$. The resemblance generally improves as the sample size grows.

We refer to this fundamental result as *The* Fundamental Theorem of Statistics (FTS). Appendix A gives a general version of the FTS and explains its vaunted "fundamental theorem" label. Here, we give a version of the FTS that can be introduced early on in an elementary course.

The **Fundamental Theorem of Statistics (FTS).**

$$\begin{cases} \text{Process Version:} & \text{If } \underline{s} \leftarrow RE(n) \text{ and } n \text{ is large, then } dist(X{:}\underline{s}) \approx dist(X{:}RE). \\ & \text{The approximation generally improves as } n \text{ grows.} \\ \text{Population Version:} & \text{If } \underline{s} \leftarrow SRS(n, \underline{P}) \text{ and } n \text{ is large, then } dist(X{:}\underline{s}) \approx dist(X{:}\underline{P}). \\ & \text{The approximations generally improve as } n \text{ grows.} \end{cases}$$

In words, if sample $\underline{s}$ comprises outcomes of n replicates of RE and n is large then the sample distribution $dist(X{:}\underline{s})$ will be approximately equal to the process distribution $dist(X{:}RE)$, see Figure 4 in Appendix B. The population version of the FTS follows from the process version because the sampling process $SRS(n, \underline{P})$ has the replicate-process form $SRS(n, \underline{P}) = RE(n)$, where $RE = SRS(1, \underline{P})$, and because $dist(X{:}SRS(1, \underline{P})) = dist(X{:}\underline{P})$ by (2). The FTS applies to any variable X, so it applies to any variable of the form $g(X)$ as well. That is, provided $\underline{s} \leftarrow RE(n)$ and n is large, we have that $dist(g(X){:}\underline{s}) \approx dist(g(X){:}RE)$.

The FTS motivates the idea behind using a sample distribution to approximate a process or population distribution. It tells us that we can improve this approximation by replicating RE more times or by taking a larger simple random sample (see Appendix A for more technical details). Importantly, the FTS can also be used to give empirical interpretations of process distributions, which can be very helpful for understanding distribution results such as those in Section 6.

The importance of the sampling process should be emphasized at this point. It should be made clear that if $\underline{s}$ is not the result of $RE(n)$ or $SRS(n, \underline{P})$, then this FTS result is generally not applicable and it could very well be that $dist(X{:}\underline{s})$ is not even close to resembling $dist(X{:}RE)$ or $dist(X{:}\underline{P})$, even when n is very large. That is, the data may tell us little about the process or population distribution, even if the sample size is very large.

5.4 Laws of Large Numbers

On an intuitive level, the FTS motivates the elementary versions of the Laws of Large Numbers (LLN) given below, see also Figures 4 and 8 in Appendix B. This first LLN states that when the sample $\underline{s}$ is generated or selected in a very special way, and the sample size is large, the sample mean $mean(X{:}\underline{s})$ will be approximately equal to the corresponding process mean $mean(X{:}RE)$ or population mean $mean(X{:}\underline{P})$. There are many good sources for a more technical treatment of LLNs (cf. Resnick, Chapter 7, 1998).

A Law of Large Numbers (LLN) for Means:

$$\begin{cases} \text{Process Version:} & \text{If } \underline{s} \leftarrow RE(n) \text{ and } n \text{ is large, then } mean(X{:}\underline{s}) \approx mean(X{:}RE). \\ & \text{The approximation generally improves as } n \text{ grows.} \\ \text{Population Version:} & \text{If } \underline{s} \leftarrow SRS(n, \underline{P}) \text{ and } n \text{ is large, then } mean(X{:}\underline{s}) \approx mean(X{:}\underline{P}). \\ & \text{The approximation generally improves as } n \text{ grows.} \end{cases}$$

This LLN applies to any variable X. Therefore, it applies to any variable of the form $g(X)$, including $g(X) = \mathbb{1}(X \in A)$. From this observation, we have the following LLN for proportions.

A Law of Large Numbers (LLN) for Proportions:

Process Version:	If $\underline{s} \leftarrow RE(n)$ and n is large, then $prop_{\underline{s}}(X \in A) \approx P_{RE}(X \in A)$. The approximation holds for any A and generally improves as n grows.
Population Version:	If $\underline{s} \leftarrow SRS(n, \underline{P})$ and n is large, then $prop_{\underline{s}}(X \in A) \approx prop_{\underline{P}}(X \in A)$. The approximation holds for any A and generally improves as n grows.

It is important that these LLNs are applied to *means* and *proportions*. The approximations are unreasonable for *sums* and *counts*. For example, given the LLN for means, a student will surely be tempted to jump to the conclusion that $n\ mean(X{:}\underline{s}) \approx n\ mean(X{:}RE)$; that is, $sum(X{:}\underline{s}) \approx n\ mean(X{:}RE)$, with the approximation improving as n grows. In fact, this approximation generally worsens as n grows! This example highlights a drawback to the use of the approximation symbol "$\approx$," which ignores the rates of convergence. It also reminds us to stress that the LLNs should be applied to means and proportions, not sums and counts.

The LLN can be combined with almost sure convergence results for continuous functions (cf. Ferguson, p.40, 1996) to produce the following law of large numbers for sample variances and standard deviations.

A Law of Large Numbers (LLN) for Variances and Standard Deviations:

Process Version:	If $\underline{s} \leftarrow RE(n)$ and n is large, then $var(X{:}\underline{s}) \approx var(X{:}RE)$ and $sd(X{:}\underline{s}) \approx sd(X{:}RE)$. The approximations generally improve as n grows.
Population Version:	If $\underline{s} \leftarrow SRS(n, \underline{P})$ and n is large, then $var(X{:}\underline{s}) \approx var(X{:}\underline{P})$ and $sd(X{:}\underline{s}) \approx sd(X{:}\underline{P})$. The approximations generally improve as n grows.

Viewed from a slightly different perspective, these LLNs give us the simple empirical interpretations of process distribution summaries that were alluded to in Section 4.2.

Empirical Interpretations of Process Distribution Summaries (Corollary to LLN):

- The LLN for Means tells us that $mean(X{:}RE) \approx mean(X{:}\underline{s})$, where $\underline{s} \leftarrow RE(M)$ and M is a very large number, e.g. $M = 10^9$, see Figure 4 in Appendix B. In other words, the process mean $mean(X{:}RE)$ is the long-run average of X values when RE is repeated over and over again.

- The LLN for Proportions tells us that $P_{RE}(X \in A) \approx prop_{\underline{s}}(X \in A)$, where $\underline{s} \leftarrow RE(M)$ and M is a very large number, e.g. $M = 10^9$. In other words, the probability $P_{RE}(X \in A)$ is the long-run proportion of X values in A, when RE is repeated over and over again.

- The LLN for Variances and Standard Deviations tells us that $var(X{:}RE) \approx var(X{:}\underline{s})$ and $sd(X{:}RE) \approx sd(X{:}\underline{s})$, where $\underline{s} \leftarrow RE(M)$ and M is a very large number, e.g. $M = 10^9$. In other words, the process variance $var(X{:}RE)$ is the long-run variance of X values and the process standard deviation $sd(X{:}RE)$ is the long-run standard deviation of X values when RE is repeated over and over again.

6 A Prelude to Statistical Inference

The LLN for Means tell us that when $\underline{s} \leftarrow RE(n)$ and n is large, the approximation $mean(X{:}\underline{s}) \approx mean(X{:}RE)$ should be reasonable. But just how reasonable is it? If we do not know $mean(X{:}RE)$, we cannot answer this question. However, it turns out that we can still address the related **Introductory Inference Questions**:

1. Just how reasonable is the approximation $mean(X{:}\underline{s}) \approx mean(X{:}RE)$ *expected to be*?
2. Given data $X{:}\underline{s}$, what are plausible estimates of $mean(X{:}RE)$?
3. Does data $X{:}\underline{s}$ give evidence against the claim that $mean(X{:}RE) = \mu_0$?

To address these questions, it will prove convenient to introduce the sample mean variable $\overline{X}$ (pronounced "X-bar") and the sample standard deviation variable S, which are defined as

$$\overline{X}(\underline{s}) \equiv mean(X{:}\underline{s}) \quad \text{and} \quad S(\underline{s}) \equiv sd(X{:}\underline{s}).$$

We say that the variables $\overline{X}$ and S measure or describe the sample $\underline{s}$ because their values depend on the entire collection of entities in $\underline{s}$; we might call these *summary variables*. We emphasize that the symbols $\overline{X}$ and S represent *variables* and $\overline{X}(\underline{s})$ and $S(\underline{s})$ represent *variable values* or *observed values of variables*.

It is critically important to note that when $\underline{s} \leftarrow RE(n)$, we can view the observed sample mean $\overline{X}(\underline{s})$ as an outcome of the process $\overline{X}{:}RE(n)$; that is, $\overline{X}(\underline{s}) \leftarrow \overline{X}{:}RE(n)$, see Figure 5 in Appendix B. As the notation suggests, the process $\overline{X}{:}RE(n)$ has action-based description, "Carry out $RE(n)$ and report the $\overline{X}$ value of the outcome." Similarly, we can view $S(\underline{s}) \leftarrow S{:}RE(n)$, where the process $S{:}RE(n)$ has an analogous action-based description. Again, we stress that repeating the process (aka random experiment), e.g. $\overline{X}{:}RE(n)$, would result in a different outcome, e.g. a different observed sample mean.

By the LLN, when $\underline{s} \leftarrow RE(n)$ and n is large, we will have that $\overline{X}(\underline{s}) \approx mean(X{:}RE)$. To assess how good this approximation is expected to be (in a probabilistic sense), we note that $RE(n)$ generated an outcome with $\overline{X}$ value equal to $\overline{X}(\underline{s})$. Now suppose that we can compute the chances that $RE(n)$ would generate an outcome with $\overline{X}$ value within k units of $mean(X{:}RE)$. If these chances are high then we have reason to be confident that our single observed value $\overline{X}(\underline{s})$ is within k units of the $mean(X{:}RE)$. To compute probabilities such as $P_{RE(n)}(-k \leqslant \overline{X} - mean(X{:}RE) \leqslant k)$, we must find or at least approximate the process distribution $dist(\overline{X}{:}RE(n))$, a distribution that is characterized by probabilities of the form $P_{RE(n)}(\overline{X} \in A)$. The next section gives a general description of process distributions of variables, especially summary variables such as $\overline{X}$ and S.

6.1 Process Distributions of Summary Variables

Besides $\overline{X}$ and S, there are infinitely many other summary variables that can be used to measure a sample. We will use the generic symbol Y to represent any one of these candidate variables. If $\underline{s} \leftarrow RE^s$, then the variable Y in the process $Y{:}RE^s$ is a random variable and $Y(\underline{s})$ is an outcome of

this process $Y{:}RE^s$; i.e. $Y(\underline{s}) \leftarrow Y{:}RE^s$. The process distribution $dist(Y{:}RE^s)$, which is characterized by probabilities of the form $P_{RE^s}(Y \in A)$, has historically been called the *sampling distribution* of variable Y [wrt the sample-generating process RE^s]. This label is appropriate because this distribution describes how the values of variable Y are expected to vary from sample to sample, in replicates of RE^s. In spite of its appropriateness, we avoid the label *sampling distribution* for three reasons: (1) it promotes conflation with a *sample distribution*, (2) it seems to restrict attention to sample-selection processes and population inference, and (3) it unnecessarily introduces a distinct name for something that has already been encountered, namely, the process distribution; e.g. there is no qualitative difference between $dist(X{:}RE)$ and $dist(\overline{X}{:}RE(n))$.

Remark: In the very special setting where $RE^s = SRS(n, \underline{P})$, it can be argued that

$$dist(Y{:}SRS(n, \underline{P})) = dist(Y{:}(\underline{s}_1, \ldots, \underline{s}_M)), \quad \text{where } (\underline{s}_1, \ldots, \underline{s}_M) \text{ are all possible samples of size } n.$$

In words, the process distribution of Y, with respect to $SRS(n, \underline{P})$, is equal to the sample distribution of Y over all possible samples of size n. Some sources use this latter sample distribution as the definition of the process distribution of Y (and call it the sampling distribution, which adds to the confusion between sampling and sample distributions!). We avoid this definition because (1) it is not generally applicable when $RE^s \neq SRS(n, \underline{P})$, (2) it masks the fact that the "sampling" distribution is actually a process (aka probability) distribution not a sample distribution, and (3) it downplays the fact that $Y(\underline{s})$ is an outcome of a process (aka random experiment), namely, $Y{:}RE^s$.

Especially in Frequentist statistical inference, a large amount of energy is devoted to finding or approximating the process distributions of variables. Toward this end, more advanced statistics courses employ tools of distribution and asymptotic theory. These tools are mostly unavailable in elementary statistics courses and, if the instructor deemed it important for students to "find" or "approximate" a process distribution, different approaches would be required. Typically, instructors of elementary courses allude to at least one of the following three approaches:

1. If RE^s can be carried out manually or via computer simulation, then we can use the FTS to approximate the process distribution $dist(Y{:}RE^s)$. Specifically, by the FTS, if $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE^s(M)$ and M is large, then $dist(Y{:}(\underline{s}_1, \ldots, \underline{s}_M)) \approx dist(Y{:}RE^s)$. See Figure 6 in Appendix B.

2. If $Y = g(\underline{X})$ and $\underline{X}{:}RE^s$ can be carried out manually or via computer simulation, then we can use the FTS to approximate the process distribution $dist(Y{:}RE^s)$. Specifically, by the FTS, if $(\underline{x}_1, \ldots, \underline{x}_M) \leftarrow (\underline{X}{:}RE^s)(M)$ and M is large, then $dist(g{:}(\underline{x}_1, \ldots, \underline{x}_M)) \approx dist(g{:}(\underline{X}{:}RE^s))$. But $g{:}(\underline{X}{:}RE^s)$ and $g(\underline{X}){:}RE^s$ are identical processes (see Figure 7 in Appendix B) and the latter is precisely $Y{:}RE^s$. Hence we have that $dist(g{:}(\underline{x}_1, \ldots, \underline{x}_M)) \approx dist(Y{:}RE^s)$.

3. If $RE^s = RE(n)$ and $Y = \overline{X}$, the Central Limit Theorem of the next section gives a very reasonable approximation to the process distribution of $\overline{X}{:}RE(n)$.

6.2 A Central Limit Theorem for the Sample Mean

We begin with two useful summaries of the process distribution $dist(\overline{X}{:}RE(n))$:

$$mean(\overline{X}{:}RE(n)) = mean(X{:}RE) \qquad \text{and} \qquad sd(\overline{X}{:}RE(n)) = \frac{sd(X{:}RE)}{\sqrt{n}}.$$

By the LLNs, these results can be given empirical interpretations. (These process mean and standard deviation results do not generally hold for $\overline{X}{:}RE^s$ when $RE^s \neq RE(n)$.)

In the population-inference setting where $RE = SRS(1, \underline{P})$, we have $mean(\overline{X}{:}SRS(n, \underline{P})) = mean(X{:}\underline{P})$. In words, under $SRS(n, \underline{P})$, "the process mean of the sample mean is equal to the population mean"! By the empirical interpretation of a process mean, this implies that the long-run average of the $\overline{X}$ values for many replications of $SRS(n, \underline{P})$ is equal to $mean(X{:}\underline{P})$.

In this special-case setting where $RE = SRS(1, \underline{P})$, we have $sd(\overline{X}{:}SRS(n, \underline{P})) = \frac{sd(X{:}\underline{P})}{\sqrt{n}}$. In words, under $SRS(n, \underline{P})$, the process standard deviation of the sample mean is equal to the population standard deviation divided by the square root of the sample size. By the empirical interpretation of a process standard deviation, we have that the long-run standard deviation of the $\overline{X}$ values for many replications of $SRS(n, \underline{P})$ is equal to $sd(X{:}\underline{P})/\sqrt{n}$. (This result tells us that the larger the sample size, the smaller the process variability of the sample mean.)

The next theorem gives us more than just the mean and standard deviation of $\overline{X}{:}RE(n)$. It gives us an approximation to $dist(\overline{X}{:}RE(n))$ that is reasonable provided n is sufficiently large. This almost magical approximation depends only on the mean and variance of $X{:}RE$, and is reasonable regardless of the form of $dist(X{:}RE)$. Because this approximation is based on a limit theorem that is considered of central importance to probability theory and statistical inference, it is called a Central Limit Theorem (see Fischer, 2010).

A Central Limit Theorem (CLT):

Process Version: If n is sufficiently large, then

$$\overline{X}{:}RE(n) \quad \overset{a}{\sim} \quad N(mean(X{:}RE), \ \frac{sd(X{:}RE)}{\sqrt{n}}).$$

This approximation holds regardless of the $dist(X{:}RE)$, and generally improves as n grows.

Population Version: If n is sufficiently large, then

$$\overline{X}{:}SRS(n, \underline{P}) \quad \overset{a}{\sim} \quad N(mean(X{:}\underline{P}), \frac{sd(X{:}\underline{P})}{\sqrt{n}}).$$

This approximation holds regardless of the $dist(X{:}\underline{P})$, and generally improves as n grows.

The FTS gives us an empirical interpretation of this CLT. For example, if $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE(n)(M)$ and M is large, then $dist(\overline{X}{:}(\underline{s}_1, \ldots, \underline{s}_M)) \overset{FTS}{\approx} dist(\overline{X}{:}RE(n)) \overset{CLT}{\approx} N(mean(X{:}RE), sd(X{:}RE)/\sqrt{n})$. It is this application of the FTS that is exploited in graphical applets that "show the CLT in action." (See for example, `http://onlinestatbook.com/stat_sim/`.) It is important to note that

$dist(\overline{X}{:}(\underline{s}_1, \ldots, \underline{s}_M)) \not\approx dist(X{:}RE)$, a common mis-perception among students (delMas et al., 2004; Thompson et al., 2007); indeed these distributions can look very different when n is large.

The condition that n be "sufficiently large" is of course an important condition in practice. For any sample size n, the reasonableness of the approximation depends on the skewness and number of support points of $dist(X{:}RE)$. Fortunately, unless $dist(X{:}RE)$ is very skewed, the CLT approximations would usually be deemed reasonable for $n \geqslant 30$ or so.

The CLT approximation is directly useful for approximating probabilities of the form $P_{RE(n)}(\overline{X} \in A)$, when $mean(X{:}RE)$ and $sd(X{:}RE)$ are known values and n is sufficiently large. For example, suppose that RE = "roll a balanced six-sided die" and X is the number of up-dots, so that $mean(X{:}RE) = 3.5$ and $sd(X{:}RE) = 1.708$. The CLT gives us a simple way to approximate the chances of seeing at least 360 up-dots on 100 rolls, i.e. a sample average of at least 3.6 up-dots on 100 rolls of the balanced die. Specifically, the CLT tells us that $\overline{X}{:}RE(100) \overset{a}{\sim} N(3.5, 1.708/\sqrt{100})$, so the chances are $P_{RE(100)}(\overline{X} \geqslant 3.6) \approx P(N(3.5, 0.1708) \geqslant 3.6) = 0.279$.

6.3 Standardized and Studentized Mean Approximations

The CLT approximation is not directly useful for the inference setting where $mean(X{:}RE)$ and $sd(X{:}RE)$ are both unknown. However, when interest lies in making inference about $mean(X{:}RE)$, there is a particularly useful "Studentized Mean Approximation" result that is motivated by the CLT and the fact that the sample standard deviation variable S is a reasonable estimator of $sd(X{:}RE)$; recall that under $RE(n)$, the LLN tells us that $S(\underline{s}) \approx sd(X{:}RE)$. To highlight the motivating role of the CLT, we first give a "Standardized Mean Approximation" result, which is an alternative, but equivalent, specification of the CLT approximation result.

Standardized Mean Approximation (CLT): For n sufficiently large,

$$\frac{\overline{X} - mean(X{:}RE)}{sd(X{:}RE)/\sqrt{n}} : RE(n) \overset{a}{\sim} N(0,1).$$

Studentized Mean Approximation (CLT variant): For n sufficiently large,

$$\frac{\overline{X} - mean(X{:}RE)}{S/\sqrt{n}} : RE(n) \overset{a}{\sim} N(0,1).$$

For the special-case $RE(n) = SRS(n, \underline{P})$, we have $mean(X{:}RE) = mean(X{:}\underline{P})$ and $sd(X{:}RE) = sd(X{:}\underline{P})$. The FTS can be used to give this Studentized approximation an empirical interpretation. Specifically, if $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE(n)(M)$ and M is large, then

$$dist\left(\frac{\overline{X} - mean(X{:}RE)}{S/\sqrt{n}}{:}(\underline{s}_1, \ldots, \underline{s}_M)\right) \approx dist\left(\frac{\overline{X} - mean(X{:}RE)}{S/\sqrt{n}}{:}RE(n)\right) \overset{a}{\sim} N(0,1).$$

Technically, the Studentized approximation follows from the Standardized approximation along with an application of Slutsky's theorem (cf. Ferguson, p. 41, 1996). The adjective "Studentized" is a nod to the historical contribution of William Gosset, who published under the pseudonym "Student." Gosset studied how the process distribution of the Standardized Mean changed when the process

standard deviation was replaced by an estimator. In Student (1908), he found the exact process distribution of what we are calling the "Studentized mean" under the restrictive assumption that $X{:}RE$ is Normal and the sample was the outcome of $RE(n)$. This process distribution is now called "Student's t distribution based on $n-1$ degrees of freedom," and denoted $t(n-1)$.

In the Studentized Approximation, the $N(0,1)$ approximation can be replaced by the Student's $t(n-1)$ approximation. Some instructors and practitioners prefer this t approximation to the Normal approximation for the following reasons: (1) For large n, the approximations are nearly identical; (2) for smaller n, the t approximation leads to more conservative inference; and (3) when $X{:}RE \sim N(mean(X{:}RE), sd(X{:}RE))$, the t approximation is exact for all n, see Student (1908).

6.4 Introductory Inference Questions: Revisited

The initial paragraph of this section listed three questions about the process mean $mean(X{:}RE)$. Here we focus on the first two of these questions. The third question would be addressed using tests of hypotheses or significance testing, an important topic that, owing to space limitations, will not be covered in the current paper.

Consider Question 1, Just how reasonable is the LLN approximation $\overline{X}(\underline{s}) \approx mean(X{:}RE)$ expected to be? Assuming that $\underline{s} \leftarrow RE(n)$, we have that $\overline{X}(\underline{s}) \leftarrow \overline{X}{:}RE(n)$ and $S(\underline{s}) \leftarrow S{:}RE(n)$.

We also know, by the Standardized Mean Approximation and a property of the $N(0,1)$ curve, that

$$0.95 \approx P_{RE(n)}(-1.96\frac{sd(X{:}RE)}{\sqrt{n}} \leqslant \overline{X} - mean(X{:}RE) \leqslant 1.96\frac{sd(X{:}RE)}{\sqrt{n}}).$$

That is, there is approximately a 95% chance that $RE(n)$ will generate a sample $\underline{s}'$ with sample mean value $\overline{X}(\underline{s}')$ that is within $1.96\frac{sd(X{:}RE)}{\sqrt{n}}$ units of $mean(X{:}RE)$. When $sd(X{:}RE)$ is unknown, as is usually the case, we can invoke the LLN and replace it by the observed sample standard deviation $S(\underline{s}) = sd(X{:}\underline{s})$. For example, suppose that $n = 35$ and $S(\underline{s}) = 12.7$, where $\underline{s} \leftarrow RE(n = 35)$. Then we can say that there is approximately a 95% chance that $RE(35)$ will generate a sample with sample mean value that is within $1.96(12.7/\sqrt{35}) = 2.147$ units of the process mean $mean(X{:}RE)$.

Consider Question 2, Given the data $X{:}\underline{s}$, what are plausible estimates of $mean(X{:}RE)$? Assuming that $\underline{s} \leftarrow RE(n)$, we have that $\overline{X}(\underline{s}) \leftarrow \overline{X}{:}RE(n)$ and $S(\underline{s}) \leftarrow S{:}RE(n)$. We also know, by the Studentized Mean Approximation and a property of the $N(0,1)$ curve, that

$$0.95 \approx P_{RE(n)}(-1.96\frac{S}{\sqrt{n}} \leqslant \overline{X} - mean(X{:}RE) \leqslant 1.96\frac{S}{\sqrt{n}}).$$

That is, there is approximately a 95% chance that $RE(n)$ will generate a sample $\underline{s}'$ with sample mean value $\overline{X}(\underline{s}')$ that is within $1.96\frac{S(\underline{s}')}{\sqrt{n}}$ units of $mean(X{:}RE)$. Equivalently, there is approximately a 95% chance that $RE(n)$ will generate a sample $\underline{s}'$ such that the interval $CI(\underline{s}') \equiv [\overline{X}(\underline{s}') - 1.96\frac{S(\underline{s}')}{\sqrt{n}},\ \overline{X}(\underline{s}') + 1.96\frac{S(\underline{s}')}{\sqrt{n}}]$ includes the value $mean(X{:}RE)$. For this reason, we say that with 95% confidence, plausible estimates of the $mean(X{:}RE)$ are those values in our single observed interval $CI(\underline{s})$; i.e. between $\overline{X}(\underline{s}) - 1.96\frac{S(\underline{s})}{\sqrt{n}}$ and $\overline{X}(\underline{s}) + 1.96\frac{S(\underline{s})}{\sqrt{n}}$.

Consider the process $CI{:}RE(n)$, where the random variable CI is aptly called a random interval. We observed $CI(\underline{s}) \leftarrow CI{:}RE(n)$ and argued that $P_{RE(n)}(CI \ni mean(X{:}RE)) \approx 0.95$. For this reason, we call $CI(\underline{s})$ an [observed] approximate 95% confidence interval for the $mean(X{:}RE)$. Most instructors believe it is important to stress that the "95% confidence" property has a probability interpretation only for the confidence interval process $CI{:}RE(n)$, not for the single observed confidence interval $CI(\underline{s})$. That is, much to the chagrin of students and instructors alike, the phrase "95% confidence" generally cannot be given the direct "post-data" probability interpretation, $0.95 \approx P(CI(\underline{s}) \ni mean(X{:}RE))$, which is to say that there is about a 95% probability that the $mean(X{:}RE)$ is contained in the observed confidence interval $CI(\underline{s})$. For an illustrative example, see DeGroot and Schervish (2002:412).

Technically, the "95%" refers to the "pre-data" probability that $RE(n)$ will generate a sample that gives a confidence interval based on the formula CI that contains $mean(X{:}RE)$. This probability is usually given an empirical interpretation via the LLN. If $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE(n)(M)$ then provided M is large, the LLN for proportions tells us that $prop_{(\underline{s}_1,\ldots,\underline{s}_M)}(CI \ni mean(X{:}RE)) \approx P_{RE(n)}(CI \ni mean(X{:}RE)) \approx 0.95$. That is, if $RE(n)$ is repeated many, say M, times then about 95% of the like-constructed confidence intervals, the $CI(\underline{s}_1), \ldots, CI(\underline{s}_M)$, will contain the $mean(X{:}RE)$. Calling $CI(\underline{s})$ a 95% confidence interval makes sense because $CI(\underline{s})$ can be thought of as one of these many generated intervals, of which 95% contain $mean(X{:}RE)$.

7 Examples

Using only the concepts discussed in this paper, we can address questions like those posed in the next two examples. See Appendix C for detailed solutions.

Example 1. Population Inference. Consider a population of potential ferry boat passengers. As part of a ferry-boat safety study, a sample of 1000 was taken from this population. For this sample, the average weight was 158 pounds and the standard deviation was 22 pounds.

(a) *What can be said about the population using the safety study data?*

(b) *If a simple random sample of* 350 *passengers will be taken, what are the chances that the mean weight will exceed* 162 *pounds; i.e. the total weight will exceed the ferry boat's weight limit of* 56700 *pounds?*

(c) *If many simple random samples of size* 350 *will be taken, about what fraction of the samples will have mean weight that exceeds* 162 *pounds; i.e. a total weight that exceeds the ferry boat's weight limit of* 56700 *pounds?*

(d) *If* 350 *passengers board the ferry, what are the chances that the mean weight exceeds* 162 *pounds; i.e. the total weight exceeds the weight limit of* 56700 *pounds?*

(e) *Suppose that the original safety study's sample was the result of a simple random sample of size* 1000*, taken with replacement, from the population. Give a range of plausible estimates of the average weight for the population.*

Example 2. Process Inference. An unbalanced six-sided die was rolled 1000 times. The number of rolls that were '1's, '2's, '3's, '4's, '5's, and '6's are 232, 229, 220, 106, 100, and 113, respectively.

(a) *What can be said about the die rolling process?*

(b) *What are plausible estimates of the mean number of up-dots for this die-rolling process?*

(c) *What are the chances that the range (= maximum minus the minimum) for the next* 10 *rolls will be at least* 4*?*

8 Discussion

Although this paper includes some material that is not meant for direct consumption by the elementary statistics student, it presents concepts and notations that are targeted at this student audience. The overarching goals of this paper are (1) to outline an approach that allows instructors to present formal links between data and the targets of inference within the first few class periods of an elementary statistics course; (2) to equip instructors with tools that facilitate a smooth transition from data to statistical inferences about populations *and* processes; and (3) to realize these first two goals earlier in a survey course so that more time can be devoted to other important statistical concepts (see for example, the American Statistical Association's GAISE report, 2012). The less conventional, more explicit notations of this paper serve to highlight the preeminent roles of entities such as variables, samples, populations, and processes. They also mitigate problems of conflating concepts such as sample, population, and process means or distributions. On a related note, we recommended avoiding the phrase "sampling distribution" for several reasons (see Section 6.1).

This paper's elementary versions of the FTS, LLN, and CLT are applicable only when the data have the special form $X{:}\underline{s}$, where the sample $\underline{s}$ is the outcome of a random experiment of the form $RE(n)$ or $SRS(n, \underline{P})$. We focused on this special setting for several reasons including: (1) It is arguably a very important and commonly-encountered setting in practice; (2) the approximation results of this paper also apply for the SRS case, where the sample is taken *without replacement*, provided the sample size n is a small fraction of the population size N, e.g. $n/N \leqslant 0.10$; (3) it allowed us to avoid introducing datum variables, the $X_1, \ldots, X_n$ of mathematical statistics; (4) it allowed us to avoid introducing the concept of independent and identically distributed random variables; and (5) it allowed us to present the fundamental linking results, FTS and LLN, early on in an elementary course.

This paper's special setting does, of course, preclude several cases encountered in more advanced statistics courses and applied work. We list a few examples of cases not covered in this paper: (1) In a more mathematical course, we encounter data of the form $X_1(\underline{s}), \ldots, X_n(\underline{s})$, where at least one of the components, say $X_k(\underline{s})$, depends on multiple s_j's. (2) In a time series course, it is usually the case that $\underline{s} \leftarrow RE^s$, where $RE^s \neq RE(n)$. For example, RE^s can often be viewed as a sequential random experiment of the form $RE^s = RE_1{:}RE_2{:}\ldots{:}RE_n$, where the components RE_j depend on the outcomes of the previous $RE_1, \ldots, RE_{j-1}$. (3) In a sampling theory course, we might consider

sub-sampling outcomes of a process of interest; or sampling that depends on X values; or sampling plans other than simple random sampling with replacement.

That the data are of the form $X{:}\underline{s}$, where $\underline{s} \leftarrow RE(n)$, can be viewed as an intuitively appealing sufficient condition for the applicability of more general versions of the FTS, LLN, and CLT that apply when data are realizations of independent and identically distributed (IID) random variables. To see why this is so, note that the data $\underline{x} = X{:}\underline{s}$ has components of the form $x_i = X_i(\underline{s})$, where $X_i(\underline{s}) = X(s_i)$. This fact along with the condition that $\underline{s} \leftarrow RE(n)$ implies that the data $x_1, \ldots, x_n$ are realizations of random variables $X_1, \ldots, X_n$, which are IID [with respect to $RE(n)$] with common distribution $dist(X{:}RE)$. Most textbooks give only the IID versions of the FTS, LLN, and CLT, which unfortunately means that discussion of these fundamental concepts must be delayed until the concept of IID random variables has been covered.

There is a rich literature on statistics education (see for example, Moore et al. 1995; delMas et al. 2004; Thompson et al. 2007; Garfield and Ben-Zvi, 2007, and references therein). This research makes it abundantly clear that there is a need to take a fresh look at the way statistics is taught and learned. The current paper was motivated by this research as well as the author's personal experience teaching elementary statistics courses to undergraduate non-statistics-majors at the University of Iowa.

References

American Statistical Association, (2012), *Guidelines for Assessment and Instruction in Statistics Education: College Report*, Alexandria, VA: Author.

Bingham, N.H. (2000), "Studies in the History of Probability and Statistics XLVI. Measure into Probability: From Lebesgue to Kolmogorov," *Biometrika*, **87**, 145-156.

Cantelli, F.P. (1933), "Sulla Determinazione Empirica della Leggi di Probabilità," *Giorn. Ist. Ital. Attauri*, **4**, 421-424.

Chandra, T.K. and Chatterjee, D. (2001), *A First Course in Probability*, Boca Raton, FL: Chapman & Hall/CRC Press.

Csörgő, M. (2002), "A Glimpse of the Impact of Pál Erdős on Probability and Statistics," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **30**, 493-556.

DeGroot, M.H. and Schervish, M.J. (2002), *Probability and Statistics*, 3rd edn, Boston, MA: Addison-Wesley.

DeHardt, J. (1971), "Generalizations of the Glivenko-Cantelli Theorem," *The Annals of Mathematical Statistics*, **42**, 2050-2055.

delMas, R.C., Garfield, J., and Chance, B.L. (2004), "Using Assessment to Study the Development of Students' Reasoning about Sampling Distributions," Paper presented at the annual meeting of the *American Educational Research Association*, San Diego, CA.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, London, UK: Chapman and Hall.

Fischer, H. (2010). *A History of the Central Limit Theorem: From Classical to Modern Probability Theory,* Sources and Studies in the History of Mathematics and Physical Sciences, New York: Springer.

Frick, R.W. (1998), "Interpreting Statistical Testing: Process and Propensity, Not Population and Random Sampling," *Behavior Research Methods, Instruments and Computers*, **30**, 3, 527-535.

Gaenssler, P. and Stute, W. (1979), "Empirical Processes: A Survey of Results for Independent and Identically Distributed Random Variables," *The Annals of Probability*, **7**, 193-243.

Garfield, J. and Ben-Zvi, D. (2007), "How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics," *International Statistical Review*, **75**, 3, 372-396.

Glivenko, V. (1933), "Sulla Determinazione Empirica della Legge di Probabilità," *Giorn. Ist. Ital. Attauri*, **4**, 92-99.

Kahneman, D. (2011), *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.

Loéve, M. (1955), *Probability Theory,* New York: Van Nostrand.

Mcnaughton, D.B. (2002), "The Introductory Statistics Course: The Entity-Property-Relationship Approach," Unpublished manuscript, last updated Jan 30, 2002, Downloaded May 12, 2008, from http://www.matstat.com/teach/eprt0130.pdf.

Moore, D.S., Cobb, G.W, Garfield, J., and Meeker, W.Q. (1995), "Statistics Education Fin de Siecle," *The Amer Statist*, **49**, 3, 250-260.

Resnick, S.I. (1998), *A Probability Path*, Boston, MA: Birkhäuser.

Student (1908), "The Probable Error of a Mean," *Biometrika*, **6**, 1-25.

Thompson, P., Liu, Y., and Saldanha, L. (2007), "The Intricacies of Statistical Inference," in M. Lovett and P. Shah (Eds.), *Thinking with Data*, pp. 207-231, Mahwah NJ: Erlbaum Thomson.

Wessels, H., and Nieuwoudt, H. (2013), "Teachers Reasoning in a Repeated Sampling Context," *Pythagoras*, **34**, No 1, 11 pages, doi: 10.4102/pythagoras.v34i1.169.

Wolfowitz, J. (1960), "Convergence of the Empiric Distribution Function on Half-spaces," In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling,* Stanford, CA: Stanford Univ. Press.

Appendix A. The Fundamental Theorem of Statistics

The Glivenko-Cantelli Theorem (Glivenko 1933, Cantelli 1933) states that the empirical distribution function based on IID real-valued random variables uniformly converges almost surely to the common distribution function. As Csörgő (2002) puts it, the Glivenko-Cantelli Theorem "guarantees that the

notion of doing statistics via taking random samples does make sense, ultimately almost surely." Csörgő (2002) goes on to note that Loéve (1955) called the Glivenko-Cantelli Theorem "the fundamental theorem of statistics." (Others agree with Loéve, cf. Bingham 2000, Chandra and Chatterjee 2001).

Below we give a simple, yet broadly applicable, generalization of the Glivenko-Cantelli Theorem and refer to it as *The* Fundamental Theorem of Statistics (FTS). For a nice discussion of generalized Glivenko-Cantelli results, see Gaenssler and Stute (1979) and references therein. For technical precision, this FTS is stated using measure-theoretic probability concepts. Obviously, this statement of the FTS would not be used in an elementary statistics course. Indeed, one of the primary goals of this paper is to develop a simplified version of this FTS that can be introduced early on in any elementary statistics course.

The Fundamental Theorem of Statistics. *Suppose that $X_1, \ldots, X_n, \ldots$ are IID random variables defined on probability space $(\Omega, \mathcal{F}, P_0)$ with values in the measurable space $(\chi, \mathcal{A})$. Consider two important cases: (1) If χ is countable (so X_i's are discrete and possibly non-Euclidean), choose $\mathcal{A} = \{$set of all subsets of $\chi\}$; (2) If $\chi \subseteq I\!R^K$ (so X_i's are K-dimensional Euclidean vectors), choose $\mathcal{A} = Borel(I\!R^K)$. Let the common distribution of the X_i's be denoted by P and defined as $P(A) \equiv P_0(X_1 \in A)$, for $A \in \mathcal{A}$. Define the sample (or empirical) distribution P_n as $P_n(A) \equiv n^{-1} \sum_{i=1}^{n} 1(X_i \in A)$, for $A \in \mathcal{A}$. It follows that*

$$\sup_{A \in \mathcal{A}_0} |P_n(A) - P(A)| \stackrel{a.s.}{\to} 0,$$

where $\mathcal{A}_0 = \{$set of all subsets of $\chi\} = \mathcal{A}$ when χ is countable and $\mathcal{A}_0 = \{\times_{k=1}^{K}(-\infty, x_k] : x_k \in I\!R\} \subset \mathcal{A}$ when $\chi \subseteq I\!R^K$. For either case, $P = Q$ on $\mathcal{A}_0$ implies that $P = Q$ on $\mathcal{A}$. In this sense, the result states that the sample distribution P_n converges almost surely to the common probability distribution P.

Glivenko (1933) showed the result for $\chi = I\!R^1$ in the continuous case. Cantelli (1933) showed the result for $\chi = I\!R^1$ in the general case. Wolfowitz (1960) showed the result for $\chi = I\!R^K$ in the general case (see also the corollary to Theorem 4 in DeHardt (1971)). A straightforward application of Scheffé's lemma (cf. Resnick, 1998:253) can be used to show the result for the case when χ is countable. See Gaenssler and Stute (1979) for a discussion of these results, as well as other generalizations.

The Fundamental Theorem of Statistics (FTS) represents a useful and direct link between data and the object of inference, namely a process or a population. The FTS motivates the utility of increased sample sizes and explains why we go through the trouble of describing a sample (data) distribution when we are really interested in inferences about a process [aka probability] or population distribution. The FTS also motivates certain laws of large numbers (LLN) and validates simulation-based approximations for process [aka sampling] distributions. Csörgő (2002) also points out that without the FTS, "the initial idea of bootstrapping would not have been possible." In sum, the FTS is truly fundamental to the study of statistics

Appendix B. Figures Corresponding to Notations and Concepts

This appendix gives graphical representations of many of the main concepts described in the body of the paper. Instructors might find these useful for explaining the concepts. In the following figures, variables are depicted as "function boxes" and processes are depicted as "clouds."

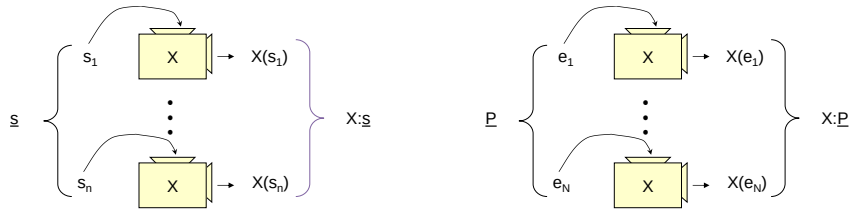

Figure 1. X values for the sample (i.e. data), $X{:}\underline{s}$, and X values for the population, $X{:}\underline{P}$.

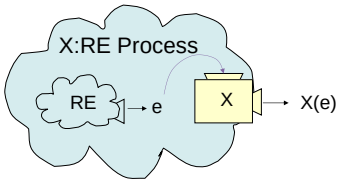

Figure 2. Process $X{:}RE$. Here $X(e) \leftarrow X{:}RE$; i.e. $X(e)$ is an outcome of the $X{:}RE$ process.

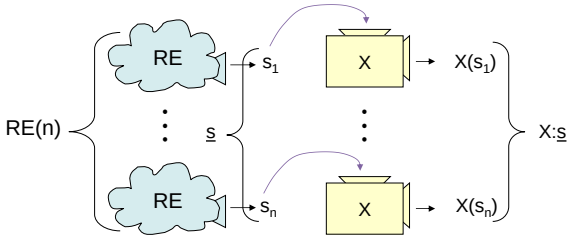

Figure 3. Data as an Outcome of a Process. Here $X{:}\underline{s} \leftarrow (X{:}RE)(n)$; i.e. the data are the outcomes of n replications of $X{:}RE$.

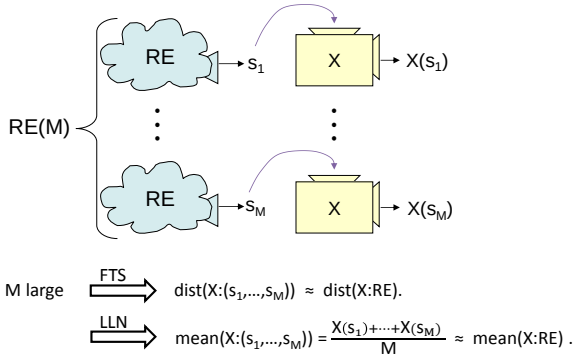

Figure 4. The Fundamental Theorem of Statistics and a Law of Large Numbers for Means. This also depicts the empirical interpretation of the process mean $mean(X{:}RE)$.

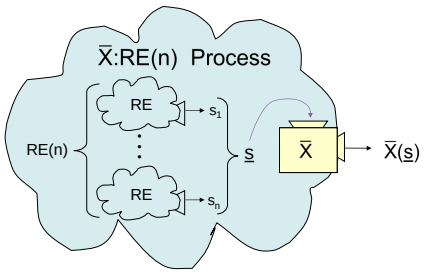

Figure 5. Process $\overline{X}$:$RE(n)$. Here $\overline{X}(\underline{s}) \leftarrow \overline{X}$:$RE(n)$; i.e. the observed sample mean is an outcome of the $\overline{X}$:$RE(n)$ process.

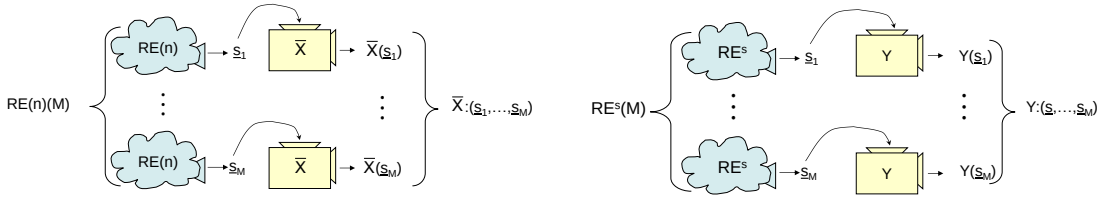

Figure 6. Approximating $dist(\overline{X}$:$RE(n))$ using the FTS. Because $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE(n)(M)$ and assuming that M is large, the FTS tells us that $dist(\overline{X}$:$(\underline{s}_1, \ldots, \underline{s}_M)) \approx dist(\overline{X}$:$RE(n))$. Approximating $dist(Y$:$RE^s)$ using the FTS. Because $(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE^s(M)$ and assuming that M is large, the FTS tells us that $dist(Y$:$(\underline{s}_1, \ldots, \underline{s}_M)) \approx dist(Y$:$RE^s)$.

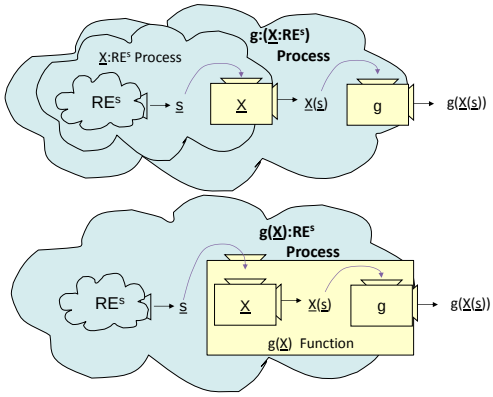

Figure 7. Equivalence of Processes g:$(\underline{X}$:$RE^s)$ and $g(\underline{X})$:RE^s.

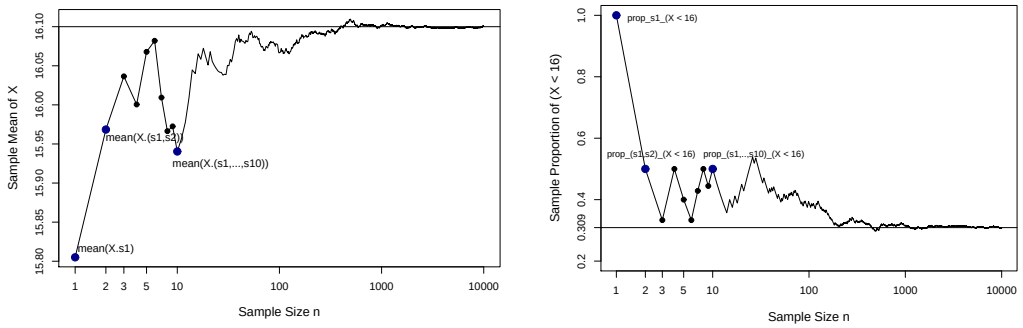

Figure 8. Laws of Large Numbers for Means and Proportions. Here $mean(X$:$RE) = 16.10$ and $P_{RE}(X < 16) = 0.309$.

Appendix C. Solutions to Examples

Example 1. Population Inference. Consider a population of potential ferry boat passengers. As part of a ferry-boat safety study, a sample of 1000 was taken from this population. For this sample, the average weight was 158 pounds and the standard deviation was 22 pounds.

(a) *What can be said about the population using the safety study data?*

Let $\underline{P}$ be the population of potential ferry boat passengers, let $W = weight$ and $\underline{s} = (s_1, \ldots, s_{1000})$ be the safety study's sample of passengers. The data are $W{:}\underline{s}$ and we are given $mean(W{:}\underline{s}) = 158$ and $sd(W{:}\underline{s}) = 22$.

If we can believe that $\underline{s} \leftarrow SRS(n = 1000, \underline{P})$, then by the FTS and LLNs,

$$dist(W{:}\underline{P}) \approx dist(W{:}\underline{s}), \quad mean(W{:}\underline{P}) \approx mean(W{:}\underline{s}) = 158, \quad sd(W{:}\underline{P}) \approx sd(W{:}\underline{s}) = 22.$$

Thus, if $\underline{s} \leftarrow SRS(1000, \underline{P})$, the population distribution of weights will look very similar to the sample distribution, the population mean will be around 158, and the population standard deviation will be around 22.

If, however, $\underline{s} \not\leftarrow SRS(1000, \underline{P})$, which is most likely the case, then the FTS and LLNs do not generally apply and we can say very little about $dist(W{:}\underline{P})$.

(b) *If a simple random sample of* 350 *passengers will be taken, what are the chances that the mean weight will exceed* 162 *pounds; i.e. the total weight will exceed the ferry boat's weight limit of* 56700 *pounds?*

The sample to be taken, say $\underline{s}_0$, can be viewed as $\underline{s}_0 \leftarrow SRS(n = 350, \underline{P})$. We use $\underline{s}_0$ because the symbol $\underline{s}$ is reserved for the safety study's sample of 1000. We define the sample mean variable $\overline{W}$ as $\overline{W}(\underline{s}_0) = mean(W{:}\underline{s}_0)$, the average weight for the sample $\underline{s}_0$. The chances that the average weight for the sample will exceed 162 is determined by the process distribution $dist(\overline{W}{:}SRS(350, \underline{P}))$. By the CLT, regardless of the shape of $dist(W{:}\underline{P})$,

$$\overline{W}{:}SRS(350, \underline{P}) \quad \overset{a}{\sim} \quad N(\ mean(W{:}\underline{P}), \frac{sd(W{:}\underline{P})}{\sqrt{350}}).$$

We are not given the population mean or standard deviation, so we will use the LLN approximations $mean(W{:}\underline{P}) \approx mean(W{:}\underline{s}) = 158$ and $sd(W{:}\underline{P}) \approx sd(W{:}\underline{s}) = 22$. That is, we will approximate the chances using $\overline{W}{:}SRS(350, \underline{P}) \quad \overset{a}{\sim} \quad N(\ 158, \frac{22}{\sqrt{350}} = 1.176)$. In particular, the chances are

$$P_{SRS(350,\underline{P})}(\overline{W} > 162) \approx P(N(158, 1.176) > 162) = 0.00034.$$

There is around a 3 in 10000 chance that the average weight of the simple random sample of 350 will exceed 162. That is, there is around a 3 in 10000 chance that the total weight for the simple random sample of 350 will exceed the weight limit of 56700.

(c) *If many simple random samples of size* 350 *will be taken, about what fraction of the samples will have mean weight that exceeds* 162 *pounds; i.e. a total weight that exceeds the ferry boat's weight limit of* 56700 *pounds?*

Let $\underline{s}_1,\ \underline{s}_2, \ldots, \underline{s}_M$ be the many simple random samples of size 350. These can be viewed as

$$(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow SRS(350, \underline{P})(M).$$

Define the sample mean variable $\overline{W}$ as in part (b). For example, $\overline{W}(\underline{s}_1)$ is the average weight for the first simple random sample $\underline{s}_1$. By the FTS and CLT, we have

$$dist(\overline{W}{:}(\underline{s}_1, \ldots, \underline{s}_M)) \overset{FTS}{\approx} dist(\overline{W}{:}SRS(350, \underline{P})) \overset{CLT}{\approx} N(\ mean(W{:}\underline{P}),\ \frac{sd(W{:}\underline{P})}{\sqrt{350}}).$$

Again we will invoke the LLNs and replace $mean(W{:}\underline{P})$ and $sd(W{:}\underline{P})$ by $mean(W{:}\underline{s}) = 158$ and $sd(W{:}\underline{s}) = 22$. We have that

$$dist(\overline{W}{:}(\underline{s}_1, \ldots, \underline{s}_M)) \overset{FTS,CLT,LLN}{\approx} N(\ 158,\ \frac{22}{\sqrt{350}} = 1.176).$$

Therefore, we have that

$$prop_{(\underline{s}_1,\ldots,\underline{s}_M)}(\overline{W} > 162) \approx P(N(158, 1.176) > 1.62) = 0.00034.$$

That is, about 3 in 10000 of the many simple random samples of size 350 will have mean weight that exceeds 162. That is, about 3 in 10000 of the many simple random samples of size 350 will have total weight exceeding 56700.

(d) *If* 350 *passengers board the ferry, what are the chances that the mean weight exceeds* 162 *pounds; i.e. the total weight exceeds the weight limit of* 56700 *pounds?*

Represent the 350 boarded passengers by $\underline{s}_b$ and view this sample as $\underline{s}_b \leftarrow RE^s$. Here RE^s is the sample-generating process. If we could assume that $RE^s = SRS(350, \underline{P})$, then we could use the result of part (b) to conclude that the chances are only about 3 in 10000. Unfortunately, passengers travel in groups (picture a traveling football team or a travelling ballet company) and the assumption that the boarding sample can be viewed as the result of a simple random sample of size 350 from $\underline{P}$ is simply not tenable. Therefore, the 3 in 10000 approximation is probably not reasonable and safety personnel should not rule out the possibility that there is a much higher chance that 350 passengers will exceed the total weight limit (again, picture that travelling football team). In practice, we could make different assumptions about RE^s and use computer simulation and the FTS to inform our guess at the process distribution $dist(\overline{W}{:}RE^s)$.

(e) *Suppose that the original safety study's sample was the result of a simple random sample of size* 1000*, taken with replacement, from the population. Give a range of plausible estimates of the average weight for the population.*

Assuming that $\underline{s} \leftarrow SRS(1000, \underline{P})$, the Studentized Mean Approximation gives

$$\frac{\overline{W} - mean(W{:}P)}{S/\sqrt{1000}} : SRS(1000, \underline{P}) \quad \overset{a}{\sim} \quad N(0, 1),$$

where the standard deviation variable S is defined as $S(\underline{s}) = sd(W{:}\underline{s})$, the standard deviation of the weight values for sample $\underline{s}$. This approximation gives

$$\begin{aligned} 0.95 &\approx P_{SRS(1000,\underline{P})}(-1.96 \leqslant \frac{\overline{W} - mean(W{:}P)}{S/\sqrt{1000}} \leqslant 1.96) \\ &= P_{SRS(1000,\underline{P})}(\overline{W} - 1.96\frac{S}{\sqrt{1000}} \leqslant mean(W{:}\underline{P}) \leqslant \overline{W} + 1.96\frac{S}{\sqrt{1000}}). \end{aligned}$$

It follows that an approximate 95% observed confidence interval is $CI(\underline{s}) = [\overline{W}(\underline{s}) - 1.96\frac{S(\underline{s})}{\sqrt{1000}},\ \overline{W}(\underline{s}) + 1.96\frac{S(\underline{s})}{\sqrt{1000}}]$. Plugging in $\overline{W}(\underline{s}) = mean(W{:}\underline{s}) = 158$ and $S(\underline{s}) = sd(W{:}\underline{s}) = 22$, we have that $CI(\underline{s}) = [158 - 1.96(22/\sqrt{1000}), 158 + 1.96(22/\sqrt{1000})] = [156.6, 159.4]$. With 95% confidence, plausible estimates of the population mean $mean(W{:}\underline{P})$ are between 156.6 and 159.4 pounds. Of course, if the sample $\underline{s}$ was not obtained via a $SRS(1000, \underline{P})$, then this interval is not valid.

Example 2. Process Inference. An unbalanced six-sided die was rolled 1000 times. The number of rolls that were '1's, '2's, '3's, '4's, '5's, and '6's are 232, 229, 220, 106, 100, and 113, respectively.

(a) *What can be said about the die rolling process?*

One possible way to define the die rolling process is RE = "roll the unbalanced six-sided die and report the number of up-dots." Let X be the number of up-dots on a die and let $\underline{s} = (s_1, \ldots, s_{1000})$ be the outcomes for the 1000 rolls. Note that with our definition of RE, we have $X(s_j) = s_j$ because outcomes of RE are in fact the number of up-dots. We are not given the data $X{:}\underline{s}$, but we are given the counts, $freq_{\underline{s}}(X = 1) = 232$, ..., $freq_{\underline{s}}(X = 6) = 113$. In this setting, these counts determine the sample distribution $dist(X{:}\underline{s})$. Because it is reasonable to assume that $\underline{s} \leftarrow RE(1000)$, the FTS is applicable and we have $dist(X{:}RE) \approx dist(X{:}\underline{s})$. That is, the data $X{:}\underline{s}$ tells us a lot about the process $X{:}RE$. For example, the FTS (or LLN) tells us

$$\begin{aligned} &P_{RE}(X = 1) \approx prop_{\underline{s}}(X = 1) = 0.232, \quad &&P_{RE}(X = 2) \approx prop_{\underline{s}}(X = 2) = 0.229, \\ &P_{RE}(X = 3) \approx prop_{\underline{s}}(X = 3) = 0.220, \quad &&P_{RE}(X = 4) \approx prop_{\underline{s}}(X = 4) = 0.106, \\ &P_{RE}(X = 5) \approx prop_{\underline{s}}(X = 5) = 0.100, \quad &&P_{RE}(X = 6) \approx prop_{\underline{s}}(X = 6) = 0.113. \end{aligned}$$

For instance, the chances that RE generates an outcome with X value equal to 1 (i.e. the chances that this unbalanced die will be rolled and land with a '1' on the up-face) is approximately equal to 0.232, the proportion in the sample $\underline{s}$ with X value equal to 1.

(b) *What are plausible estimates of the mean number of up-dots for this die-rolling process?*

Use part (a)'s definitions of process RE and variable X. The most important first step is to recognize that the target of inference is the process mean $mean(X{:}RE)$, which by the LLN can be interpreted as the long-run average of X (number of up-dots) values when RE (roll the die) is repeated over and over again.

Define $\overline{X}$ and S as $\overline{X}(\underline{s}) = mean(X{:}\underline{s})$ and $S(\underline{s}) = sd(X{:}\underline{s})$. For the observed sample $\underline{s}$, we have that $\overline{X}(\underline{s}) = 2.952$ and $S(\underline{s}) = 1.637$. Because $\underline{s} \leftarrow RE(1000)$, the Studentized Mean Approximation is applicable: we have that $\frac{\overline{X} - mean(X:RE)}{S/\sqrt{1000}} : RE(1000) \overset{a}{\sim} N(0,1)$ and

hence $CI(\underline{s}) = [\overline{X}(\underline{s}) - 1.96\frac{S(\underline{s})}{\sqrt{1000}}, \overline{X}(\underline{s}) + 1.96\frac{S(\underline{s})}{\sqrt{1000}}] = [2.851, 3.053]$ is an approximate 95% confidence interval for $mean(X{:}RE)$. That is, with 95% confidence, plausible estimates of the process mean $mean(X{:}RE)$ are between 2.851 and 3.053 .

(c) *What are the chances that the range (= maximum minus the minimum) for the next* 10 *rolls will be at least* 4*?*

Let $\underline{s}_0 = (s_1, \ldots, s_{10})$ represent the outcomes for the next 10 rolls. Let the range variable R be defined as $R(\underline{s}_0) = maximum\{s_1, \ldots, s_{10}\} - minimum\{s_1, \ldots, s_{10}\}$. It is reasonable to view $\underline{s}_0 \leftarrow RE(10)$, where RE was defined in part (a). To answer the question about the chances, we will consider the process distribution $dist(R{:}RE(10))$ and use it to compute $P_{RE(10)}(R \geqslant 4)$. Because R is not an average of the 10 outcomes, the CLT approximation to the $dist(R{:}RE(10))$ is not applicable. We will have to consider a different approximation.

In a more advanced course, we might set out to derive the distribution of $R{:}RE(10)$ using distribution theory, but this is not a simple problem. Here, we will first consider generating

$$(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE(10)(M), \quad \text{where } M \text{ is large,}$$

so that, by an application of the FTS, we have $dist(R{:}(\underline{s}_1, \ldots, \underline{s}_M)) \approx dist(R{:}RE(10))$.

Unfortunately, we cannot carry out RE because we do not have access to the unbalanced die used to generate the data in this problem. In fact, we cannot even simulate RE because we do not know the probabilities such as $P_{RE}(\{j\})$. As a fix, we will instead carry out RE^*, where $P_{RE^*}(\{j\}) = prop_{\underline{s}}(\{j\})$, so $dist(RE^*) = dist(\underline{s})$. That is, RE^* is the random experiment with action-based description, "simulate the roll of a die that has probabilities 0.232, 0.229, 0.220, 0.106, 0.100, and 0.113 of coming up '1', '2', '3', '4', '5', and '6', respectively." Now by the FTS applied to the observed sample $\underline{s} \leftarrow RE(1000)$, we have $dist(RE^*) = dist(\underline{s}) \overset{FTS}{\approx} dist(RE)$, i.e. RE and RE^* should have similar distributions. Therefore, if

$$(\underline{s}_1, \ldots, \underline{s}_M) \leftarrow RE^*(10)(M), \quad \text{where } M \text{ is large,}$$

then two applications of the FTS will give us

$$dist(R{:}(\underline{s}_1, \ldots, \underline{s}_M)) \overset{FTS}{\approx} dist(R{:}RE^*(10)) \overset{FTS}{\approx} dist(R{:}RE(10)).$$

It follows that $P_{RE(10)}(R \geqslant 4) \overset{LLN}{\approx} P_{RE^*(10)}(R \geqslant 4) \overset{LLN}{\approx} prop_{(\underline{s}_1, \ldots, \underline{s}_M)}(R \geqslant 4)$.
A computer simulation using $M = 10^6$, gave $prop_{(\underline{s}_1, \ldots, \underline{s}_M)}(R \geqslant 4) = 0.895027$. Thus, there is about a 90% chance that the range (= max minus min) for the next 10 rolls of this unbalanced die will be at least 4.