Tutorial

Bonnie E. Shook-Sa*, Stephen R. Cole, Paul N. Zivich, Jessie K. Edwards, Taylor J. Krajewski, Timothy Feeney, Adaora Adimora and Michael G. Hudgens

A primer on large-sample statistical inference for epidemiologists

https://doi.org/10.1515/em-2025-0036
Received August 12, 2025; accepted February 26, 2026; published online March 17, 2026

Abstract: Statistical theory forms a foundation for how epidemiologists learn about populations in public health and medical studies and is fundamental for the understanding of more advanced epidemiological methods (e.g., in causal inference and machine learning). Textbooks provide in-depth coverage of probability and statistical theory, but with such comprehensive coverage that it can be easy to miss the forest for the trees. Here, we provide a summary of fundamental concepts from large-sample statistical theory to allow for more focused understanding tailored to epidemiologists and health science researchers. This primer aims to promote appropriate understanding and application of statistical methods in epidemiologic research. We clarify several often-confused statistical topics and provide a motivation for the application of large-sample inferential methods to data from population health and medical studies. Assumptions underlying commonly used statistical methods that must be considered for valid inference are also discussed. These ideas are contextualized with an example from the Women's Interagency HIV Study.

Keywords: inference; statistics; random error; sampling; target population

Epidemiologists aim to understand and improve population health. For a target population, a specific group of people characterized in terms of person, place, and time [1], epidemiologists may wish to describe current health characteristics, quantify known contributors to risk for a disease, estimate the effects of interventions on health outcomes, or predict future health trends. As data for the entire target population are typically not available or are prohibitively expensive to collect, epidemiologists rely on samples, or subsets, of population members to make inference. Here, inference refers to the process of estimating parameters of interest (e.g., summary measures of health characteristics) and quantifying associated uncertainty.

Statistical theory forms a foundation for such inference, which is fundamental to epidemiology research. While there are many textbooks focused on probability and statistical inference, they are generally not targeted towards epidemiologists. Our aim is to complement the formal, rigorous coverage of statistical theory offered in textbooks with a more conceptual introduction that ties together key ideas and builds statistical intuition

Dr. Adimora passed away during the preparation of this manuscript.

***Corresponding author: Dr. Bonnie E. Shook-Sa**, Nuffield Department of Population Health, University of Oxford, Richard Doll Building, Old Road Campus, Oxford OX3 7LF, UK; and Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, E-mail: Bonnie.Shook-Sa@ndph.ox.ac.uk. https://orcid.org/0000-0001-9506-4047
Stephen R. Cole, Jessie K. Edwards and Timothy Feeney, Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Paul N. Zivich and Adaora Adimora, Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; and Institute of Global Health and Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Taylor J. Krajewski, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
Michael G. Hudgens, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

around topics relevant to epidemiologists. Our hope is that this tutorial will help solidify these important topics which are foundational to the understanding of both basic and more advanced epidemiologic methods.

We focus this primer on large-sample (i.e., frequentist) inference, which predominates in epidemiology. Based on our teaching experience, one common source of confusion is how large-sample inferential methods, which are commonly justified with an infinite superpopulation, can be applied to real-world finite study samples. Therefore, we start this primer by motivating the use of large-sample methods with finite populations. We then describe statistical parameters, inference, and systematic errors. We conclude the primer with a brief discussion of alternative paradigms of inference and more advanced epidemiological methods, for which this primer provides a foundation. A glossary of terms from this primer is included in Appendix Table 1.

Our motivating example throughout is an example from descriptive epidemiology. Assume we aim to estimate mean CD4 cell count among women with HIV in the United States at a given time and the proportion of women with CD4 cell counts <200 cells/mm^3. CD4 cell count is a measure of immune function relevant for measuring HIV disease progression, where a normal CD4 count is about 500–1,500 cells/mm^3 and a dangerously low CD4 count is below 200 cells/mm^3. Our study sample is comprised of n units, drawn from the target population. Here, our study sample consists of n=1,164 participants from the Women's Interagency HIV Study (WIHS) in 1995 [2]. Let the outcome, Y, represent nadir CD4 cell count, observed for all n units.

Motivation for large-sample inference

What does it mean to be iid?

With large-sample inference, we often assume the n observations of Y in the sample, i.e., $Y_1, Y_2, \ldots, Y_n$, are independent and identically distributed (iid) realizations of a random variable from the target population. *Independent* means that the variables of one unit do not depend on the distributions of other units' variables. One implication of independence for binary random variables A and B is that $P(A|B) = P(A)$. That is, with independent random variables, whether or not B occurred tells us nothing about whether or not A will occur. For example, whether one cancer patient survives 5 years following diagnosis is not expected to affect an unrelated cancer patient's status at 5 years. Violations of the independence assumption occur in situations where observed data are correlated (e.g., breast cancer statuses of mothers and daughters or repeated measures of CD4 cell counts from the same person). *Identically distributed* means that the distribution of observed variables is the same for all units [3] (p. 210). That is, the units were drawn from the same probability distribution. Violations of the identically distributed assumption occur in situations where data are sampled from two or more distinct populations. For example, the CD4 cell counts of women with HIV sampled from a single population are identically distributed. However, if *separate* samples are selected from women in and out of HIV care, and these populations have different underlying probability distributions, then CD4 cell counts from the two samples will not be identically distributed. In the motivating example, a single sample of women is selected from the target population, so it is reasonable to assume the data are identically distributed. Except where noted otherwise, the focus herein is on the standard iid setting. Settings with non-iid data require additional methodological considerations [4–6].

How can we use iid methods with finite samples?

The assumption that our sample is iid is typically at odds with how units (e.g., participants) are selected into our studies. For the purposes of demonstration, assume our target population consists of five women with advanced HIV. For each woman, whether her CD4 count is <200 cells/mm^3 is recorded, such that the outcomes for the five women are $\{1, 0, 0, 1, 1\}$. Here, the population mean is 0.6. Assume we sample two women from this population, with Y_1 and Y_2 denoting the outcomes for the first and second sampled women, respectively. If we select the women *with replacement*, meaning that all women in the population are eligible for selection at each sample draw, the two selections are independent, as the value of Y_1 does not affect the

distribution of Y_2, i.e., $P(Y_2 = 1|Y_1 = 1) = P(Y_2 = 1|Y_1 = 0) = 0.6$. Alternatively, we could sample *without replacement*, meaning that once a woman is sampled, she is ineligible to be selected again. Under sampling without replacement, if $Y_1 = 1$ (i.e., woman 1, 4, or 5 is sampled first), the mean outcome for the remaining women is 0.5, i.e., $P(Y_2 = 1|Y_1 = 1) = 0.5$. However, if $Y_1 = 0$ (i.e., woman 2 or 3 is sampled first), the mean outcome for the remaining women is 0.75, i.e., $P(Y_2 = 1|Y_1 = 0) = 0.75$. Thus, the distribution of the outcome of the second sampled woman (Y_2) depends on the value of the first sampled woman (Y_1), and thus Y_1 and Y_2 are not independent.

This simple example demonstrates that for finite populations, sampling with replacement is in agreement with the iid assumption, but without replacement sampling violates the iid assumption. In practice, most epidemiologic studies use without replacement sampling because (1) researchers often do not have enumerated lists of the target population from which to select with replacement samples and (2) with replacement designs are inefficient due to the redundancy in information provided by (possibly) selecting the same unit twice.

Although without replacement sampling in finite populations leads to dependency in observations and violates the iid assumption, the *degree* of dependence varies by the relative sizes of the target population and sample. In the above example, assume each of the five observations represents the outcome for 10,000 women, for a total population of 50,000 women. In this larger population, regardless of which woman is sampled first, the mean outcome for the remaining population members is approximately 0.6 ($P(Y_2 = 1|Y_1 = 1) = 0.599992 \approx P(Y_2 = 1|Y_1 = 0) = 0.600012$). In this setting, observations are "nearly" independent because the probability distribution of the outcome for an individual given the outcomes of previously sampled individuals is similar to its marginal distribution [3] (p. 210). A formal demonstration of near independence based on a mathematical limit is presented in the Appendix. The key takeaway from this demonstration is that, while scientists almost never have an infinite target population, for simple random sample designs inference can safely proceed using large-sample methods when the target population is large relative to the sample size such that the iid assumption holds approximately.

Evaluating samples prior to analysis

Most large-sample inferential methods assume, conceptually, that the study sample was randomly selected from the target population. One of the first steps in any analysis is to clearly specify the target population and assess the plausibility of the random-sampling assumption. In population health research, the study sample may be a convenience sample [7], which violates the random-sampling assumption. For example, epidemiologic studies can rely on patients who attended a healthcare system during a specific time period, patients who signed up for clinical trials, or population members recruited online who agreed to complete health questionnaires. When the random sampling assumption is violated, one option is to respecify the target population to align with the observed sample, i.e., to implicitly or explicitly assume that the target population is the population that, when repeatedly sampled, would align with study data in expectation [8]. However, redefining the target population might not address the original research question. Alternatively, methods to correct for biased sampling and generalize estimates from the study sample to the target population can be implemented, as further discussed in the Systematic Errors section.

Parameter of interest

Once the target population is defined, the next step is specification of the parameter of interest, i.e., the estimand. Under large-sample (frequentist) inference, parameters are viewed as fixed, but unknown, constants. That is, the parameter is a characteristic of the target population that we do not know but seek to estimate from the observed sample. Unlike the parameter of interest, the observed data result from random sampling and differ for each potential sample selected from the target population. Thus, our outcome of interest Y is viewed as a random variable defined by its cumulative distribution function [3] (p. 29): $F_Y(y) = P(Y \leq y)$. Common parameters of interest are summaries of the outcome's cumulative distribution function (e.g., risk at a particular time point

in survival analysis) or measures of central tendency such as the mean (i.e., the random variable's expected value, $E(Y) = \mu$). A formal definition of expected value is provided in the Appendix. The parameters of interest in our motivating example are the mean CD4 cell count and the proportion of women with CD4 cell counts <200 cells/mm^3.

Variance and standard deviation quantify how much spread there is in the outcome variable in the target population, with $\mathrm{Var}(Y) = E\{(Y-\mu)^2\}$, and standard deviation, $\mathrm{SD}(Y) = \sqrt{\mathrm{Var}(Y)}$. The standard deviation is typically easier to interpret than the variance, as the square root puts the measure back in the natural units of Y (rather than Y squared). Variation in Y introduces random error in estimation of the parameter of interest from the sample, which is quantified during statistical inference.

Statistical inference

A central goal of analysis is to estimate the parameter of interest and quantify uncertainty due to sampling. Estimators are formulated to provide an approximation of the parameter of interest given the sample data. Note the *estimator* is the method used to approximate the parameter with data from any given sample, while an *estimate* is a single realized numeric value of the estimator when applied to a sample. Here, we use $\hat{\theta}$ to denote estimators of an arbitrary parameter θ and other symbols to denote specific estimators.

Any function that maps the observed data to possible values of the parameter of interest is an estimator, but not all estimators are useful. For example, consider two estimators of the population mean μ: (1) the sample mean $\overline{\mu} = \mathrm{n}^{-1}\sum_{i=1}^{\mathrm{n}} Y_i$ and (2) the sample minimum $\mu^* = \min(Y_i)$,for $i = 1, \ldots,$ n. In the WIHS data, $\overline{\mu} = 394$ cells/mm^3 and $\mu^* = 0$ cells/mm^3 (Figure 2A). One might expect the sample mean to be a good approximation for the population mean (as discussed further below), but would in general not think the minimum observed value would be close to the true population mean. The properties of estimators are thus important when deciding which methods to apply for a given study.

Properties of estimators

Two properties of estimators that can be considered when gauging their utility are bias and variance. Bias is equal to the expected value of the estimator minus the true parameter value, i.e., $\mathrm{Bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$. Such bias is different than sparse data bias, which refers to bias resulting from analyzing data with few or no study participants that have particular combinations of variables [9]. Unlike sparse data bias, which is remedied as the sample size increases, bias of the form $E\left(\hat{\theta}\right) - \theta$ may persist even with large samples.

Even if an estimator is unbiased and gives us the right answer on average (i.e., across hypothetical repeated samples), we typically only observe one sample resulting in a single estimate (Figure 1B). The single estimate observed for the study is rarely exactly equal to the true parameter value, even with an unbiased estimator. The distribution of the possible values of the estimator across potential samples is often referred to as the estimator's *sampling distribution* (Figure 1C). It is important to characterize the sampling distribution by quantifying how much spread is expected in $\hat{\theta}$. The variance of the estimator is $\mathrm{Var}\left(\hat{\theta}\right) = E\left[\left\{\hat{\theta} - E\left(\hat{\theta}\right)\right\}^2\right]$. The standard error of the estimator, $\mathrm{SE}\left(\hat{\theta}\right) = \sqrt{\mathrm{Var}\left(\hat{\theta}\right)}$, is a metric that quantifies how much the estimator $\hat{\theta}$ differs from its expected value on average across potential samples of size n from the target population. Note the distinction between the standard deviation and the standard error. While the standard deviation quantifies spread in the *outcome values* (Y), the standard error quantifies spread in *an estimator* ($\hat{\theta}$) across potential repeated samples of size n from the target population. Because the standard error quantifies spread in $\hat{\theta}$, it tends to decrease as n increases. This is not true for the standard deviation, which measures the variability in the outcome for the entire population and thus remains constant regardless of sample size. Understanding distinctions between different types of variability can be difficult, even for experts [10].

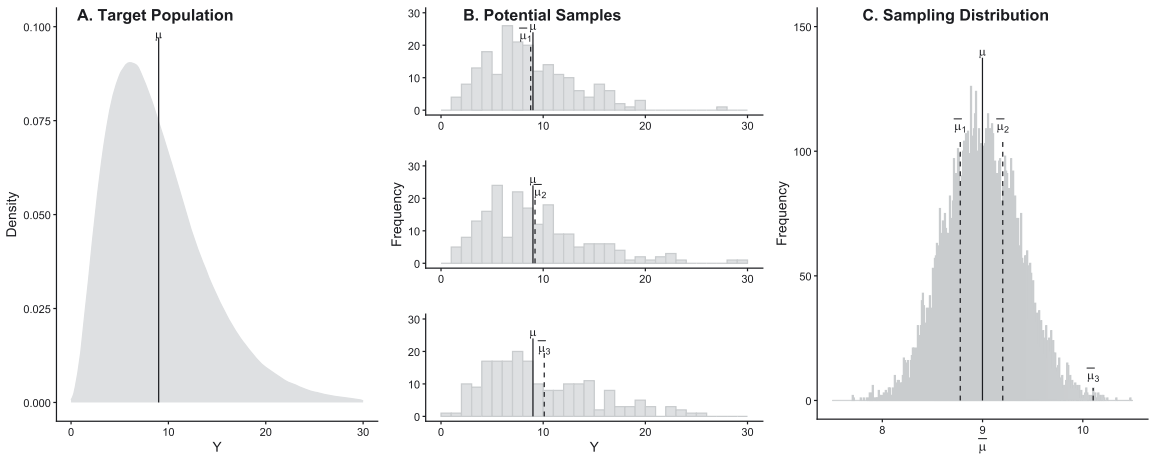

Figure 1: Sampling process. Panel A displays the distribution of the outcome of interest in the target population, panel B depicts three potential random samples of size 200 from the target population, with sample means $\overline{\mu}_1$, $\overline{\mu}_2$, and $\overline{\mu}_3$ (dotted lines), and panel C depicts the distribution of sample means across 10,000 possible random samples from the target population. In all figures, the solid line denotes the true mean in the target population (μ).

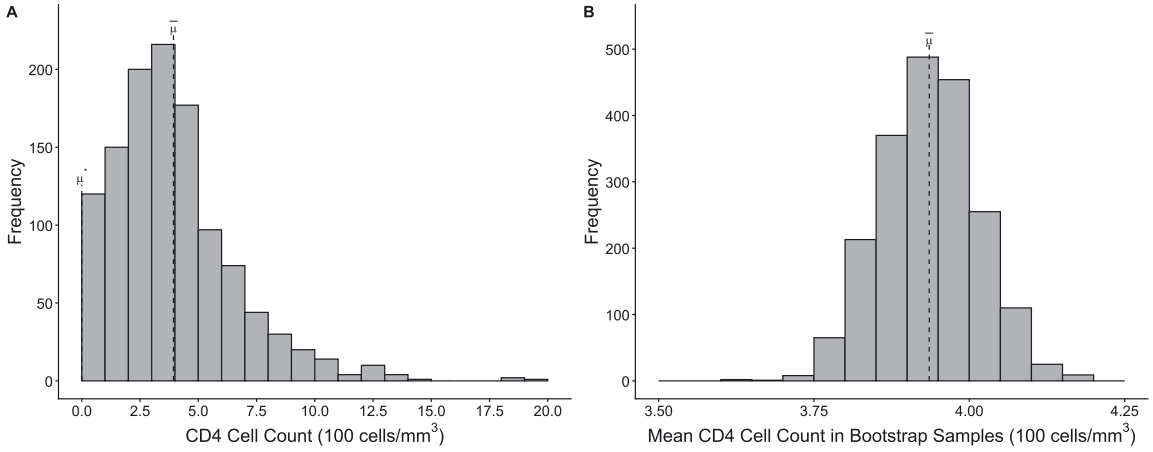

Figure 2: Distribution of (A) CD4 cell counts in WIHS sample, n=1,164 and (B) mean CD4 cell count in 2,000 bootstrap samples.

Ideally, estimators will be both unbiased and precise, such that $E\left(\hat{\theta}\right) = \theta$ and $\text{Var}(\hat{\theta})$ is small relative to the variance of other possible estimators of θ. In practical settings, epidemiologists often must make analytic decisions that have implications on both bias and variance, as some approaches that aim to reduce bias in estimators can result in more imprecision. This is often referred to as the "bias-variance trade-off" [11] (pp. 159–160). When comparing candidate estimators, researchers should consider both their bias and their variance. Mean squared error (MSE) is one way to gauge the combined performance of an estimator, as it quantifies both bias and variance [3] (p. 330), with $\text{MSE}\left(\hat{\theta}\right) = E\left\{\left(\hat{\theta} - \theta\right)^2\right\} = \text{Var}\left(\hat{\theta}\right) + \left\{\text{Bias}\left(\hat{\theta}\right)\right\}^2$. Note the MSE is equal to variance for unbiased estimators.

As an example of the above concepts, the sample mean $\overline{\mu} = \text{n}^{-1}\sum_{i=1}^{\text{n}} Y_i$ is an unbiased estimator of the population mean μ (see proof [3] (pp. 213–214)), with $\text{Var}\left(\overline{\mu}\right) = \sigma^2/\text{n}$, where $\sigma^2 = \text{Var}(Y)$. The standard deviation of Y, $\text{SD}(Y) = \sqrt{\sigma^2} = \sigma$, and the standard error of the estimator $\text{SE}\left(\overline{\mu}\right) = \sqrt{\text{Var}\left(\overline{\mu}\right)} = \sigma/\sqrt{\text{n}}$. Note the standard error and variance of $\overline{\mu}$ are scaled by $\sqrt{\text{n}}$ and n, respectively, as there is less uncertainty in estimation for larger samples than smaller samples. The standard deviation is typically estimated by $\widehat{\text{SD}}(Y) = \sqrt{\left\{\sum_{i=1}^{\text{n}}\left(Y_i - \overline{\mu}\right)^2\right\}/(\text{n}-1)}$.

The division by $n-1$ rather than n is referred to as "Bessel's correction" and is done to provide an unbiased estimator of $SD(Y)$ [12]. The standard error of $\overline{\mu}$ is typically estimated as $\widehat{SE}(\overline{\mu}) = \widehat{SD}(Y)/\sqrt{n}$. Estimated standard errors are commonly reported as a metric to gauge the precision of estimators and used to construct confidence intervals. Since $\overline{\mu}$ is an unbiased estimator of μ, $MSE(\overline{\mu}) = Var(\overline{\mu}) + \{0\}^2 = \sigma^2/n$. A proportion ($\rho$) is a special case of the mean estimator when Y is binary, with $\overline{\rho} = \overline{\mu}$ and $Var(\overline{\rho}) = \overline{\rho}(1-\overline{\rho})/n$. In the WIHS example, the estimated standard deviation and standard error for CD4 cell count are $\widehat{SD}(Y) = 264$ cells/mm^3 and $\widehat{SE}(\overline{\mu}) = 8$ cells/mm^3, respectively. The estimated proportion of women with CD4 cell counts <200 cells/mm^3 is $\overline{\rho} = 0.23$ with $\widehat{SE}(\overline{\rho}) = 0.01$.

When evaluating estimators, epidemiologists can consider their exact properties, as discussed in the previous paragraph, or their "large-sample" properties. The large-sample properties of estimators are established using asymptotic theory, which describes the limiting behavior of estimators as the sample size approaches infinity. When estimators approach a fixed value as the sample size increases, this is mathematically referred to as convergence. In practice, many useful estimators may be biased in small samples, but they have desirable large-sample properties (e.g., asymptotic unbiasedness, consistency). Asymptotically unbiased estimators have bias that tends toward zero as the sample size tends to infinity, but their variance does not necessarily tend to zero (i.e., the limit of the expected value of the estimator converges to the parameter of interest). An estimator can be asymptotically unbiased even if it sometimes produces estimates far from the true parameter value as long as the average of these estimates is close to the true parameter value as the sample size increases. Consistent estimators converge in probability to the parameter of interest, i.e., the limit of the probability that the estimator is close to the true parameter value converges to 1 (see Appendix). A consistent estimator has a high probability of being close to the parameter of interest when the sample size is large [13] (p. 51), which means that consistency is often a more desirable property than asymptotic unbiasedness. Consistent estimators may exhibit bias in small samples (see, for example [14]) but tend to perform well in large samples.

As an illustration of the above concepts, consider the following estimators. If one were to take the first observation (Y_1) as an estimator of the mean (μ), this estimator would be unbiased since $E(Y_1) = \mu$. However, it is not consistent since the estimator does not get closer to the population mean as the sample size increases. Thus, not all unbiased estimators are consistent. Alternatively, many estimators we use in practice (e.g., estimators of the odds and hazard ratios) are biased but consistent. Another example is the Hajek estimator $\tilde{\mu} = \left(\sum_{i=1}^{n} W_i Y_i\right) / \left(\sum_{i=1}^{n} W_i\right)$, where W_i represents the weight for participant i. The Hajek estimator is commonly used in inverse-probability weighted analyses. This estimator is biased but consistent [15], and often has a smaller standard error and mean squared error than the unbiased Horvitz-Thompson estimator $\breve{\mu} = \left(\sum_{i=1}^{n} W_i Y_i\right)/n$ [16, 17].

The sample mean $\overline{\mu}$ is an unbiased estimator of the population mean μ and has desirable asymptotic behavior. The law of large numbers states that, for iid samples, as n approaches infinity, the sample estimator of the mean will tend toward the population mean (formal statement in Appendix). More precisely, the weak law of large numbers is a fundamental theorem of statistics that states that the sample mean is a consistent estimator of the population mean [13] (p. 49).

Confidence intervals

Point estimation is only one piece of the inferential puzzle. As previously discussed, there is spread in $\hat{\theta}$ across the multiple potential samples that could be selected from the target population (Figure 1C). While variance and standard error quantify this spread, confidence intervals (CIs) provide additional information about the precision of an estimator that is useful in interpreting study findings. If a study were repeated an infinite number of times and our assumptions are correct, an exact 95 % CI would contain the true population parameter value in at least 95 % of these repeated studies. However, researchers rarely have the opportunity to conduct repeated sampling. In practice, the width of the CI is often interpreted as a relative measure of statistical precision. Wide CIs are indicative of high uncertainty in estimation due to random error, while narrower CIs reflect smaller amounts of sampling error and better statistical precision [18]. An alternative interpretation of CIs considers

them to be "compatibility intervals," as they provide the range of parameter values compatible with the observed data [19, 20].

Exact vs. large-sample confidence intervals

In some settings it is possible to construct an exact $(1-\alpha)$% CI for a given significance level α which has the desirable property that the interval contains the true parameter at least $(1-\alpha)$% of the time across repeated samples, even for small samples. However, exact CIs tend to be conservative, i.e., they are often wider than necessary. Additionally, they are often based on strong assumptions about the distribution of the data, or fall outside of the large-sample inferential paradigm [21]. Instead, we focus on asymptotic, or approximate, CIs.

Asymptotic CIs are common in epidemiologic studies. With such CIs, the interpretation of containing the true parameter $(1-\alpha)$% of the time across repeated samples does not hold exactly for small samples but is expected to be close for moderately sized samples. The central limit theorem (CLT) provides theoretical justification for commonly used asymptotic CIs. The CLT states that for sequences of iid random variables with finite variance, the standardized sample mean, $\sqrt{\mathrm{n}}\left(\overline{\mu}-\mu\right)/\sigma$, has a limiting standard normal distribution (formal definition in the Appendix). That is, as the sample size grows, the sampling distribution of $\overline{\mu}$ more closely resembles a normal distribution, centered at μ. Note this does not imply that the distribution of Y in the sample is or becomes normal for large samples (Figure 1A and B). Instead, the distribution of the *estimated means* across potential samples tends towards a normal distribution as the sample size grows, converging to a normal distribution in the limit (Figure 1C).

The CLT is a powerful result that holds regardless of the distribution of Y. While individual values of Y are difficult to predict when the distribution of Y is unknown, the (*mean* of Y becomes easier to estimate as sample size grows. A sample size of $\mathrm{n}=30$ or larger is often suggested as a rule of thumb for applying the CLT for the estimation of means [22] (p. 348). However, the exact sample size needed for the CLT approximation to be adequate varies depending on the distribution of Y [13] (pp. 73–74) (e.g., the outcome's distribution could be similar to a normal distribution or could be skewed or discrete). The CLT approximation may be reasonable for smaller samples when Y is known or assumed to be normally distributed.

The CLT is the basis for the construction of commonly used Wald-type asymptotic CIs which have the form $\hat{\theta}\pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\left(\hat{\theta}\right)$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. For a two-sided 95 % CI, $\alpha=0.05$ and $z_{0.975}=1.96$. In practice, Wald-type CIs have known limitations [23, 24], but because they can be used with a broad range of estimators they are widely applicable. An approximate $(1-\alpha)$% Wald-type CI for the population mean μ is $\overline{\mu}\pm z_{1-\alpha/2}\widehat{\mathrm{SD}}(Y)/\sqrt{\mathrm{n}}$.

Bootstrap confidence intervals

Another method for constructing approximate CIs is the nonparametric bootstrap [25]. As it is typically not feasible to select multiple samples from the target population, bootstrap instead relies on repeated sampling from the original sample. Typically, B repeated samples *with replacement* of size n are drawn from the sample. The estimator is applied to each resample, resulting in estimates $\hat{\theta}_b$. The distribution of the bootstrap estimates mimics the sampling distribution of $\hat{\theta}$. Therefore, the *standard deviation* of the bootstrap estimates $\hat{\theta}_b$ is an estimator of the *standard error* of $\hat{\theta}$. Then, the Wald-type CI formula can be applied to construct a Wald-type bootstrap CI. Alternatively, CIs can be constructed based on the percentiles of the bootstrap sampling distribution (e.g., the 2.5 and 97.5 percentiles form a 95 % CI). Both types of bootstrap CIs are asymptotically valid (i.e., provide coverage of $(1-\alpha)$% as the sample size tends to infinity) for the sample mean and other smooth functions of the data under regularity conditions [26, 27]. While no single value of B is appropriate for all applications, methods are available to estimate the required B for a given application [28].

There are tradeoffs for these two types of bootstrap confidence intervals. Wald-type bootstrap CIs typically require less resampling than the percentile method [25] (p. 52), making Wald-type bootstrap CIs preferred for many computationally intensive applications. However, Wald intervals are symmetric and may produce confidence bounds outside of the range of possible values for the parameter (i.e., the parameter space), while the

percentile method produces intervals within the parameter space, provided that the point estimator itself does not produce estimates outside of the parameter space. As an example, Wald intervals may be preferred for estimating the variance of the sample mean of a continuous variable in the context of a simulation study, where bootstrap resampling would need to be performed within each of thousands of simulated samples. In contrast, the percentile method may be preferred when estimating the prevalence of a rare outcome near the boundary of the parameter space.

The bootstrap method is not without limitations. Because the bootstrap is based on large-sample theory, resulting CIs may not attain the nominal level of coverage when the sample is small. The bootstrap method relies on the idea that the sample can approximate the population well enough that resampling from the original sample (creating "bootstrap samples") can mimic drawing directly from the population. However, when the original sample size is small, resamples may not faithfully represent the population distribution. There are other known limitations of bootstrap methods, so care must be taken to ensure they are appropriate for a given application [29].

Bootstrap methods have particular utility in settings where it is difficult to derive an analytical form for the variance of the estimator, including for more complex estimators that require nuisance models [30, 31]. M-estimation is another estimation approach for such complex estimators [32, 33] but is beyond the scope of this primer.

In our example, approximate 95 % CIs for the mean CD4 cell count among women living with HIV in the United States in 1995 based on the Wald interval $\overline{Y} \pm Z_{1-\alpha/2}\widehat{SD}(Y)/\sqrt{\mathrm{n}}$, bootstrap percentile method (B=2,000), and Wald-type bootstrap CI (B=2,000) are similar: 378 to 409, 379 to 409, and 378 to 409 cells/mm^3, respectively. CIs for the proportion of women with CD4 cell counts <200 cells/mm^3 based on the Wald interval, bootstrap percentile method (B=2,000), and Wald-type bootstrap CI (B=2,000) are all 0.21 to 0.25.

Systematic errors

The inferential methods discussed above are only valid when the stated assumptions hold (i.e., we have an iid random sample from a probability distribution with finite variance). When assumptions fail to hold, estimators may exhibit bias, or CIs may fail to have appropriate coverage. Systematic errors are non-random errors that lead to bias in an estimator relative to the true value in the target population. Large-sample statistical methods are often used in epidemiology to account for systematic errors. Providing a comprehensive list of the many ways assumptions can be violated and the methods used to correct for these violations is beyond the scope of this primer. Here we include a discussion of common violations, with emphasis on biased sampling.

Biased sampling

Standard large-sample inferential approaches assume that the observed sample is a random sample from the target population. However, sometimes epidemiologists intentionally over- and under-sample units with given characteristics. For example, stratified samples can be selected, where the target population is divided into two or more mutually exclusive, exhaustive groups, and independent random samples are selected from each group. This is common in survey sampling [17] and for case-control studies [34] (pp. 161–184). Both designs aim to increase representation of certain population units (e.g., based on demographic characteristics or outcome status) to improve precision of estimators. For consistent estimation, stratified sampling approaches must be considered in data analysis, e.g., through the use of weights [17].

In other settings, over- and under-sampling of population subgroups is an unintentional consequence of the sample design, e.g., when study samples are convenience samples that may differ from the target population in important ways. Bias may occur when factors associated with study participation are also associated with the outcome of interest [35]. For example, Figure 3A displays the same target population distribution as Figure 1A. However, the samples selected in Figure 3B are biased such that individuals with lower levels of the outcome were oversampled. Specifically, 115 individuals with outcomes below the median and 85 individuals

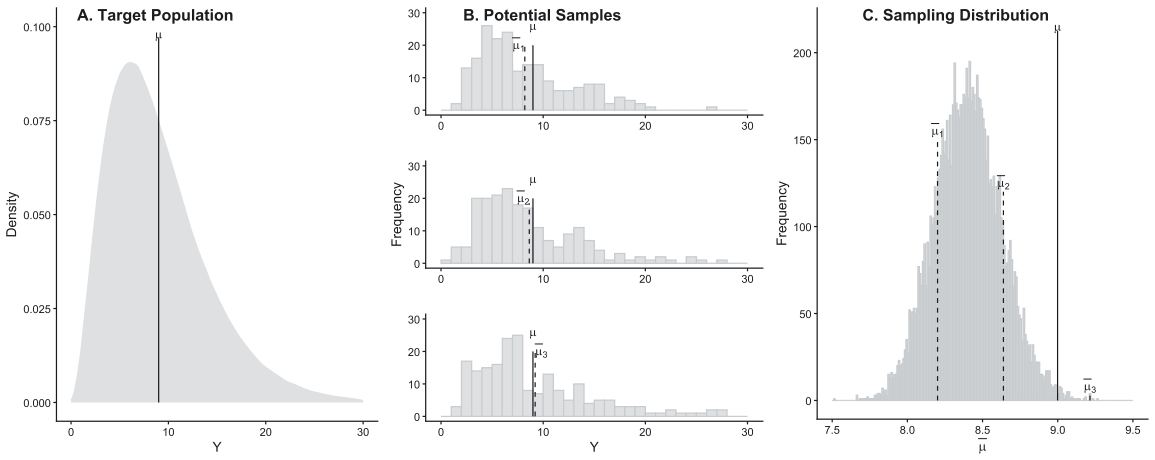

Figure 3: Sampling process with biased sampling. Panel A displays the distribution of the outcome of interest in the target population, which is unchanged from Figure 1. Panel B depicts three biased samples of size 200 from the target population, with sample means $\overline{\mu}_1$, $\overline{\mu}_2$, and $\overline{\mu}_3$ (dotted lines). Samples were selected by stratifying above and below the population median and randomly selecting 115 observations below the median and 85 observations above the median. Panel C depicts the distribution of sample means across 10,000 possible biased samples from the target population. In all figures, the solid line denotes the true mean in the target population (μ). Note the sampling distribution is centered left of the true population mean.

with outcomes above the median were sampled. The resulting sampling distribution in Figure 3C is centered below the population mean, so the estimated mean from a single observed sample is expected to be less than the population mean. Based on this biased sampling scheme, the previous methods are invalid. Estimates from the sample do not generalize to the desired target population.

With biased sampling, the target population can be respecified to better align with the study sample or analytic adjustments can be made to allow for inference to the original target population. Methods for generalizability seek to correct biases in convenience samples such that they emulate random samples from the target population, akin to how methods seek to make observational studies emulate target trials with randomized treatments [36].

Here, we demonstrate one generalizability method using inverse odds of sampling weights [37]. Alternative approaches have been proposed based on the *g*-formula [8] and augmented inverse probability weighted estimators [38]. In the observed WIHS sample, 76 % of participants were over age 30 (Table 1). Assume that the WIHS sample was older, on average, than the target population, where only 10 % of individuals were over age 30. To make inference to this hypothetical target population, we weight the WIHS data such that the distribution of age in the WIHS matches the target population distribution. We first stack the WIHS data with a random sample of 2,000 individuals from the target population (simulated for illustrative purposes). Note, when combined, the biased sample and the target population sample are not identically distributed, and thus inference relies on an extension of the standard CLT [13] (p. 571). We fit a logistic regression model, predicting the probability of being included in the WIHS sample as a function of covariates (here, an intercept term and an indicator for age >30). Inverse odds of sampling weights [37] are estimated from the fitted model, and the previously discussed Hajek estimator $\tilde{\mu}$ is applied to estimate mean CD4 cell count in the target population. The estimated CD4 cell

Table 1: Distribution of WIHS sample and hypothetical target population sample by age group.

	Age≤30	Age>30	Total
WIHS Sample	279	885	1,164
Hypothetical target population sample[a]	1,794	206	2,000

[a]Hypothetical target population sample was randomly selected from a population where 10 % of women had age >30.

count in the hypothetical target population is 427 cells/mm^3, with corresponding 95 % bootstrap CIs (B=2,000) based on the Wald-type and percentile methods of 399–454 cells/mm^3 and 400 to 454 cells/mm^3, respectively. The point estimate is notably larger than the unadjusted mean (i.e., 394 cells/mm^3) due to overrepresentation of women over age 30 in the study sample relative to the target population. CIs are over 50 % wider than the three intervals presented in the confidence interval section that assume the data are a random sample from the target population. As previously discussed, there are often tradeoffs between precision and bias.

Other systematic errors

Bias can also be introduced in the presence of measurement error [39], which occurs when the recorded values of Y do not align with the actual values of Y, due to errors in self-reported data [40, 41] or imperfect assays [42, 43], for example. Missing data due to study dropout or missed measurements on participating subjects may lead to biased estimates when the probability of missingness is associated with the outcome of interest [44]. Confounding between an observed exposure and outcome may bias estimates of causal effects, as there may be common causes of the treatment and outcome that lead to structural associations that differ from the underlying causal effect [11] (Ch. 7). Methods have been and continue to be developed for valid inference in the presence of each of these systematic errors.

Many statistical methods assume data are measured perfectly and completely on a random sample from the target population, and corresponding CIs quantify sampling error alone. In fact, large studies that yield narrow CIs can give a false sense of security in results in the presence of systematic errors [34, 45]. In the survey sampling literature, proponents of the "total survey error" framework point to these additional sources of error, which should be minimized at the study design phase and adjusted for and quantified during analyses [46]. Quantitative bias analysis methods seek to quantify potential systematic errors in terms of magnitude and direction [47]. Other research focuses on incorporating systematic error into measures of uncertainty, e.g., by proposing alternatives to conventional CIs that account for more than just sampling error [30]. Sensitivity analyses also gauge the robustness of findings to violations of assumptions [48–50].

Discussion

Large-sample inference predominates in epidemiology. Over the past several decades, advanced epidemiologic methods have been developed, building upon the foundational concepts discussed here. Yet large-sample statistical methods cannot be applied blindly, as these methods depend on assumptions. A clear understanding of these assumptions can facilitate appropriate selection and application of statistical methods. Even when clearly understood, these assumptions are often questionable, and violations may result in systematic errors. Common statistical methods quantify only random error due to sampling from the target population. But as data set size increases, random error shrinks such that, for large datasets, systematic error often becomes a larger component of total error than random error. Therefore, systematic errors become a greater driver of accuracy in the era of big data.

Large-sample inference is not the only inferential paradigm available for epidemiologists. Alternative paradigms include finite population inference and Bayesian inference. Finite population inference prevails in the design-based analysis of complex survey data [51–53] and is the basis for many exact, randomization-based statistical tests [21]. Such methods do not rely on asymptotic arguments, so are particularly useful when applied to small samples. Bayesian inference leverages Bayes' theorem to synthesize study data with prior knowledge about population parameters to inform inference [54].

Statistical inference is a broad field of study, and the theory developed to date is summarized across numerous textbooks (e.g., [3, 13, 22, 51]) and articles (e.g., [55–57]) in the peer-reviewed literature. The topics explained in this primer provide an overview of commonly used large-sample inferential methods and point towards resources with in-depth coverage of key topics. However, this primer is by no means a complete coverage of statistical inference. Several important topics (e.g., maximum likelihood estimation, imputation, modern

statistical methods) were not covered here for brevity, but are described in depth elsewhere in the epidemiology literature (e.g., [58–60]).

Notably, this primer was limited to discussions of univariate analyses. Estimating associations and causal effects are clearly of interest to epidemiologists. A wide range of statistical methods have been developed for the multivariable setting, including nonparametric and semiparametric, regression-based, inverse probability weighting, and doubly robust approaches. New methods are being developed and published every day. Importantly, the concepts in this primer are essential to understanding these and other advanced methods. Our hope is that this primer, focusing on the essentials of large-sample statistical inference, will provide a clearer picture of the forest for better comprehension of existing and newly emerging trees.

Research ethics: Not applicable, as all analyses are based on publicly available, de-identified data.
Informed consent: Not applicable, as all analyses are based on publicly available, de-identified data.
Author contributions: BES developed an initial draft of the manuscript and conducted analyses. SRC verified the analyses. All authors designed the concept and contributed to the writing of the article. Dr. Adimora passed away during the preparation of this manuscript.
Use of Large Language Models, AI and Machine Learning Tools: ChatGPT was used on the final draft of the manuscript for suggested language revisions to improve clarity and readability of the primer. The authors are solely responsible for the content.
Conflict of interest: The authors have no relevant financial or non-financial interests to disclose.
Research funding: This work was supported in part by the National Institutes of Health: R01AI157758 (Bonnie Shook-Sa, Stephen Cole, Jessie Edwards), K01AI182506 (Bonnie Shook-Sa), P30AI150410 (Bonnie Shook-Sa, Stephen Cole, Paul Zivich, Michael Hudgens), K01AI177102 (Paul Zivich), and T32AI007001 (Paul Zivich) and by Cancer Research UK (PRCRPG-Nov21\100001).
Data availability: The WIHS data analyzed in this manuscript are publicly available in Lau B, Cole SR, Gange SJ. “Competing risk regression models for epidemiologic data.” American Journal of Epidemiology 2009; 170(2): 244–56 and are available on GitHub along with analytic code at https://github.com/bonnieshook/Large_Sample_Inference_Primer.

Appendix

See Appendix Table 1.

Appendix Table 1: Glossary of terms.

Sampling and populations	
Convenience sampling	A method of sample selection where units are not selected randomly but rather from "convenient" places. An example would be a study that recruited via a flyer displayed in a clinic.
Identically distributed	The distribution of observed variables is the same for all units, i.e., the units were drawn from the same probability distribution.
IID	Independent and identically distributed
Independent	The variables of one unit do not depend on the variables of other units.
Sample	A subset of units from the target population that we use to make inference about the target population.
Sampling with replacement	A method of sampling from the target population where units are "replaced" after being selected, such that each unit is eligible for selection for each draw from the target population.
Sampling without replacement	A method of sampling from the target population where a unit is not replaced after sampling, such that each unit can be selected into the sample a maximum of one time.
Target population	The population for which we want to make inference.
Parameters and estimation	
Bias	The expected value of the estimator minus the true parameter value, i.e., $\text{Bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$. Bias tells us how far we expect our estimator to be from the true parameter value, on average.
Bias-variance trade-off	When analytic decisions that aim to reduce bias in estimators result in more imprecision, this is commonly referred to as a bias-variance trade-off.
Estimate	A single realized numeric value of the estimator when applied to a sample.
Estimator	A function of the data from any given sample used to approximate the parameter of interest.
Expected value	The average value that a variable takes across repeated sampling from the target population (see formal definition in "Definitions" section).
Inference	The process of estimating parameters of interest (e.g., summary measures of health characteristics) and quantifying associated uncertainty.
Mean	The expected value of a variable in the target population.
Mean squared error	A measure that gauges the combined performance of an estimator in terms of variance and bias, i.e., $\text{MSE}\left(\hat{\theta}\right) = E\left\{\left(\hat{\theta} - \theta\right)^2\right\} = \text{Var}\left(\hat{\theta}\right) + \left\{\text{Bias}\left(\hat{\theta}\right)\right\}^2$. Mean squared error is equal to variance for unbiased estimators.
Parameter of interest, i.e., estimand	A fixed, but unknown, characteristic of the target population that we do not know but seek to estimate from the observed sample. Examples include the mean of the outcome and risk of the outcome at a specific time point.
Random error	Variability in estimation that results from only observing a sample from the target population rather than the whole population.
Sampling distribution	The distribution of an estimator that characterizes the possible values of the estimator across potential samples.
Standard deviation (of a variable Y)	Standard deviation quantifies spread in a variable in the target population. Formally, it is equal to the square root of the variance of the variable, i.e., $\text{SD}(Y) = \sqrt{\text{Var}(Y)}$. Since standard deviation is presented in the same units as Y, it can be more intuitive than variance.
Standard error (of an estimator $\hat{\boldsymbol{\theta}}$)	Standard error quantifies spread in an estimator with a fixed sample size across possible repeated samples from the target population. Formally, it is equal to the square root of the variance of the estimator, i.e., $\text{SE}\left(\hat{\theta}\right) = \sqrt{\text{Var}\left(\hat{\theta}\right)}$.
Variance (of an estimator $\hat{\boldsymbol{\theta}}$)	The variance of an estimator quantifies how much spread is expected in the estimator with a fixed sample size across possible repeated samples from the target population. Formally, it is calculated as the expected squared deviation of the estimator from its mean, i.e., $\text{Var}\left(\hat{\theta}\right) = E\left[\left\{\hat{\theta} - E\left(\hat{\theta}\right)\right\}^2\right]$.

Appendix Table 1: (continued)

Sampling and populations	
Variance (of a variable Y)	The variance of a variable quantifies how much spread there is in the variable in the target population. Formally, it is calculated as the expected squared deviation of the variable from its mean, i.e., $\mathrm{Var}(Y) = E\{(Y-\mu)^2\}$.
Large-sample theory	
Asymptotically unbiased	An estimator is asymptotically unbiased if bias tends toward zero as the sample size tends to infinity. That is, the limit of the expected value of the estimator converges to the parameter of interest.
Central limit theorem	The central limit theorem states that for sequences of iid random variables with finite variance, the standardized sample mean, $\sqrt{\mathrm{n}}(\overline{\mu}-\mu)/\sigma$, has a limiting standard normal distribution. That is, as the sample size grows, the sampling distribution of $\overline{\mu}$ more closely resembles a normal distribution, centered at μ. A formal definition of the central limit theorem is included in the "Definitions" section.
Convergence	A mathematical concept describing the behavior of a sequence of random variables as the sample size increases. Different types of convergence include convergence in probability and convergence in distribution.
Large-sample inference	The predominant statistical paradigm of inference in epidemiology, which is primarily based on asymptotic theory, describing the limiting behavior of estimators as the sample size approaches infinity.
Law of large numbers	The law of large numbers states that, for iid samples, as n approaches infinity, the sample estimator of the mean will tend toward the population mean. More precisely, the weak law of large numbers is a fundamental theorem of statistics that states that the sample mean is a consistent estimator of the population mean. A formal definition of the weak law of large numbers is included in the "Definitions" section.
Statistical consistency	Estimators that are (statistically) consistent converge in probability to the parameter of interest. A consistent estimator has a high probability of being close to the parameter of interest when the sample size is large (see formal definition in "Definitions" section).
Confidence interval construction	
Asymptotic confidence intervals	If a study were repeated an infinite number of times, an asymptotic 95 % confidence interval is expected to contain the true parameter value in approximately 95 % of these repeated studies. This property does not hold exactly for small samples but is expected to be close for moderately sized samples.
Bootstrap confidence intervals	Confidence intervals constructed by repeatedly sampling (with replacement) from the original sample to mimic the estimator's sampling distribution.
Exact confidence intervals	If a study were repeated an infinite number of times, an exact 95 % confidence interval would contain the true parameter value in at least 95 % of these repeated studies, even for small samples.
Wald-type confidence intervals (asymptotic)	Type of confidence interval of the form $\hat{\theta} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\left(\hat{\theta}\right)$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. Wald confidence intervals are based on the central limit theorem.
Systematic errors	
Biased Sampling	A systematic error resulting from intentionally or unintentionally over- or under-sampling population subgroups relative to their representation in the target population.
Confounding	A systematic error resulting from common causes of the treatment and outcome that lead to a structural association that differs from the underlying causal effect.
Generalizability methods	Methods that seek to correct biases in sampling to enable valid inference to the target population.
Measurement error	An error which occurs when the recorded values of a variable do not align with the actual values of that variable.
Systematic errors	Systematic errors are non-random errors that lead to bias in an estimator relative to the true value in the target population. Bias from systematic errors is not remedied by increasing the sample size.

Definitions

Expected Value: The expected value of a random variable is the mean value the random variable takes across repeated samples from the target population, i.e., the weighted average of the random variable's values, where the random variable's probability distribution serves as the weight. The expected value can be thought of as the central location of the distribution of a random variable. Formally [3], if Y is continuous, $E(Y) = \int_{-\infty}^{\infty} y f_Y(y) \mathrm{d}y$, where $f_Y(y)$ is the probability density function for Y. If Y is discrete, $E(Y) = \sum_y yP(Y = y)$, where the summation is taken over all values of y that Y can equal.

Statistical Consistency: An estimator $\hat{\theta}_\mathrm{n}$ is consistent for a population parameter θ if $\hat{\theta}_\mathrm{n}$ *converges in probability* to θ. That is, for a sequence of random variables $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_\mathrm{n}$, $\lim_{\mathrm{n}\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1$ for every $\varepsilon > 0$ [3] (p. 232). This means that the limit of the probability that the estimator is close to the true parameter value converges to 1. For this reason, a consistent estimator has a high probability of being close to the parameter of interest when the sample size is large.

The Weak Law of Large Numbers: Consider the independent and identically distributed (iid) random sample $Y_1, Y_2, \ldots, Y_\mathrm{n}$ with $E(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$, where σ^2 is finite. Let $\overline{\mu}_\mathrm{n} = \mathrm{n}^{-1}\sum_{i=1}^{\mathrm{n}} Y_i$. Then $\overline{\mu}_\mathrm{n}$ is a consistent estimator for μ, i.e., $\lim_{\mathrm{n}\to\infty} P(|\overline{\mu}_\mathrm{n} - \mu| < \varepsilon) = 1$ for every $\varepsilon > 0$. A proof of this result can be found in Casella and Berger [3] (pp. 232–233).

The Central Limit Theorem: Before stating the central limit theorem, we first introduce convergence in distribution, which relies on the concept of a limit. A sequence of random variables $Y_1, Y_2, \ldots, Y_n$ is said to converge in distribution to a random variable Y if.

$\lim_{\mathrm{n}\to\infty} F_{Y_\mathrm{n}}(y) = F_Y(y)$ for all y where $F_Y(y)$ is continuous, where $F_{Y_\mathrm{n}}(y) = P(Y_\mathrm{n} \le y)$ and $F_Y(y) = P(Y \le y)$ are the cumulative distribution functions for Y_n and Y, respectively [3] (p. 235). That is, the cumulative distribution function for Y_n approaches the cumulative distribution function for Y as n becomes large.

The central limit theorem relies on the concept of convergence in distribution. Consider the iid random sample $Y_1, Y_2, \ldots, Y_\mathrm{n}$ with $E(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$, where σ^2 is finite. Then, $Z_\mathrm{n} = \sqrt{\mathrm{n}}(\overline{\mu} - \mu)/\sigma$ converges in distribution to Z as $\mathrm{n} \to \infty$, where Z has a standard normal distribution [61] (pp. 238–239).

Near independence

The idea of observations being nearly independent can be seen more formally with the introduction of a mathematical limit. If a function $f(x)$ approaches the value L as x goes to c, we say that L is the limit of $f(x)$ as x approaches c [62] (p. 51), denoted as $\lim_{x\to c} f(x) = L$. In the example in the "Motivation for Large-Sample Inference" section, let the population size be denoted by N and the outcome for each member of the population by y_i, with $i = 1, \ldots, \mathrm{N}$. Then, the population mean is $\mu = N^{-1}\sum_{i=1}^{N} y_i$. Under simple random sampling without replacement, consider what happens to the mean outcome in the population for units eligible to be selected second (i.e., not selected first) as the population size gets larger. Denote this mean by $\mu\prime = (N-1)^{-1}\left\{\left(\sum_{i=1}^{N} y_i\right) - Y_1\right\}$, where Y_1 is the outcome for the first sampled unit, i.e., $Y_1 = y_i$ for a randomly selected $i \in \{1, 2, \ldots, N\}$. Note that $\lim_{N\to\infty} \mu\prime = \lim_{\mathrm{N}\to\infty} (\mathrm{N}-1)^{-1}\sum_{i=1}^{\mathrm{N}} y_i - \lim_{N\to\infty} (N-1)^{-1} Y_1 = \lim_{N\to\infty} \mu\frac{N}{\mathrm{N}-1} - 0$. Then, by L'Hopital's rule, $\lim_{N\to\infty} \mu\frac{\mathrm{N}}{N-1} = \mu$. Thus, $\lim_{N\to\infty} \mu\prime = \mu$. As the population size approaches infinity, the limit of the mean outcome for units eligible to be selected second is the overall population mean.

References

1. Lesko CR, Fox MP, Edwards JK. A framework for descriptive epidemiology. Am J Epidemiol 2022;191:2063—70.
2. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol 2009;170:244—56.
3. Casella G, Berger RL. Statistical inference, 2nd ed. Pacific Grove, CA: Duxbury Press; 2002.
4. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis. New York: CRC Press; 2008.
5. Shumway RH, Stoffer DS, Stoffer DS. Time series analysis and its applications. New York: Springer; 2000.
6. Kosorok MR. Bootstraps of sums of independent but not identically distributed stochastic processes. J Multivariate Anal 2003;84:299—318.
7. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. BJPsych Bull 2016;40:57—60.
8. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology 2017;28:553—61.
9. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ 2016;352:i1981.
10. Zhang S, Heck PR, Meyer MN, Chabris CF, Goldstein DG, Hofman JM. An illusion of predictability in scientific results: even experts confuse inferential uncertainty and outcome variability. Proc Natl Acad Sci 2023;120:e2302491120.
11. Hernán M, Robins J. Causal inference: what If. Boca Raton: FLChapman & Hall/CRC; 2020.
12. Upton G, Cook I. Variance. In: A dictionary of statistics. Oxford: Oxford University Press; 2008.
13. Lehmann EL. Elements of large-sample theory. New York: Springer; 1999.
14. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. Am J Epidemiol 2018;187:864—70.
15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846—66.
16. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937—60.
17. Lohr SL. Sampling: design and analysis. New York: Chapman and Hall/CRC; 2021.
18. Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12:291—4.
19. Gelman A, Greenland S. Are confidence intervals better termed “uncertainty intervals”. BMJ 2019;366:l5381.
20. Cole SR, Edwards JK, Greenland S. Surprise! Am J Epidemiol 2021;190:191—3.
21. Weerahandi S. Exact statistical methods for data analysis. New York: Springer Science & Business Media; 2003.
22. Wackerly D, Mendenhall W, Scheaffer RL. Mathematical statistics with applications, 6th ed. Belmont: Cengage Learning; 2002.
23. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat 2000;54:280—8.
24. Vollset SE. Confidence intervals for a binomial proportion. Stat Med 1993;12:809—24.
25. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton: CRC Press; 1994.
26. Gill RD, Wellner JA, Præstgaard J. Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. Scand J Stat 1989:97—128.
27. Van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 2000.
28. Andrews DW, Buchinsky M. A three-step method for choosing the number of bootstrap repetitions. Econometrica 2000;68:23—51.
29. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141—64.
30. Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. Int J Epidemiol 2004;33:1389—97.
31. Shook-Sa BE, Zivich PN, Lee C, Xue K, Ross RK, Edwards JK, et al. Double robust variance estimation with parametric working models. Biometrics 2025;81:ujaf054.
32. Stefanski LA, Boos DD. The calculus of M-estimation. Am Stat 2002;56:29—38.
33. Ross RK, Zivich PN, Stringer JS, Cole SR. M-estimation for common epidemiological measures: introduction and applied examples. Int J Epidemiol 2024;53:dyae030.
34. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
35. Shook-Sa BE, Boyce RM, Aiello AE. Estimation without representation: early severe acute respiratory syndrome coronavirus 2 seroprevalence studies and the path forward. J Infect Dis 2020;222:1086—9.
36. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;79:70—5.
37. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017;186:1010—4.
38. Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 2019;75:685—94.
39. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. Int J Epidemiol 2015;44:1452—9.

40. Lim S, Wyker B, Bartley K, Eisenhower D. Measurement error of self-reported physical activity levels in New York city: assessment and correction. Am J Epidemiol 2015;181:648—55.
41. Wang H, Heitjan DF. Modeling heaping in self-reported cigarette counts. Stat Med 2008;27:3789—804.
42. Edwards JK, Cole SR, Shook-Sa BE, Zivich PN, Zhang N, Lesko CR. When does differential outcome misclassification matter for estimating prevalence? Epidemiology 2023;10:1097.
43. Meyer MJ, Yan S, Schlageter S, Kraemer JD, Rosenberg ES, Stoto MA. Adjusting COVID-19 seroprevalence survey results to account for test sensitivity and specificity. Am J Epidemiol 2022;191:681—8.
44. Cole SR, Zivich PN, Edwards JK, Ross RK, Shook-Sa BE, Price JT, et al. Missing outcome data in epidemiologic studies. Am J Epidemiol 2022;192:6—10.
45. Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. J R Stat Soc 2016;179:319—76.
46. Groves RM, Lyberg L. Total survey error: past, present, and future. Publ Opin Q 2010;74:849—79.
47. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. Am J Publ Health 2016;106:1227—30.
48. Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996;25:1107—16.
49. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 2011;22:42—52.
50. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 2003;14:451—8.
51. Fuller WA. Sampling statistics. Hoboken: John Wiley & Sons; 2011.
52. Wu C, Thompson ME. Sampling theory and practice. Gewerbestrasse: Springer; 2020.
53. Cochran WG. Sampling techniques. New York: John Wiley & Sons; 1977.
54. Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. Am J Epidemiol 2001;153:1222—6.
55. Fisher RA. On the mathematical foundations of theoretical statistics. Philos Trans R Soc London Ser A 1922;222:309—68.
56. Henderson CR. Estimation of variance and covariance components. Biometrics 1953;9:226—52.
57. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457—81.
58. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. Am J Epidemiol 2014;179:252—60.
59. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol 2008;168:355—7.
60. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017:157—66. https://doi.org/10.2147/clep.s129785.
61. Bain LJ, Engelhardt M. Introduction to probability and mathematical statistics. Belmont, CA: Duxbury Press; 1992.
62. Larson R, Hostetler R, Edwards B. Calculus of a single variable, 6 ed. Boston, MA: Houghton Mifflin; 1997.