1

Motivation and Basic Tools

1.1 Introduction

Statistics is a body of mathematically based methodology designed to organize, describe, model, and analyze data. In this context, *statistical inference* relates to the process of drawing conclusions about the unknown frequency distribution (or some summary measure therefrom) of some characteristic of a population based on a known subset thereof (the *sample data*, or, for short, the *sample*). Drawing statistical conclusions involves the choice of suitable models that allow for random errors, and this, in turn, calls for convenient probability laws. It also involves the ascertainment of how appropriate a postulated probability model is for the genesis of a given dataset, and of how adequate the sample size is to maintain incorrect conclusions within acceptable limits.

Finite statistical inference tools, in use for the last decades, are appealing because, in general, they provide "exact" statistical results. As such, finite methodology has experienced continuous upgrading with annexation of novel concepts and approaches. Bayesian methods are especially noteworthy in this respect. Nevertheless, it has been thoroughly assessed that the scope of exact statistical inference in an optimal or, at least, desirable way, is rather confined to some special classes of probability laws (such as the exponential family of densities). In real-life applications, such optimal statistical inference tools often stumble into impasses, ranging from validity to efficacy and thus, have practical drawbacks. This is particularly the case with large datasets, which are encountered in diverse (and often interdisciplinary) studies, more so now than in the past. Such studies, which include biostatistical, environmetrical, socioeconometrical, and more notably bioinformatical (genomic) problems, in general, cater to statistical reasoning beyond that required by conventional finite-sample laboratory studies. Although this methodology may yet have an appeal in broader interdisciplinary fields, some extensions are needed to capture the underlying stochastics in large datasets and thereby draw statistical conclusions in a valid and efficient manner. The genesis of asymptotic methods lies in this infrastructure of finite-sample statistical inference. Therefore it is convenient to organize our presentation of asymptotic statistical methods with due emphasis on this finite- to large-sample bridge.

To follow this logically integrated approach to statistical inference, it is essential to encompass basic tools in probability theory and stochastic processes. In addition, the scope of applications to a broad domain of biomedical, clinical, and public health disciplines may dictate additional features of the underlying statistical methodology. We keep this dual objective of developing methodology with an application-oriented spirit in mind and

thereby provide an overview of finite-sample (exact or small) statistical methods, appraising their scope and integration to asymptotic (approximate or large-sample) inference.

In Section 1.2, we motivate our approach through a set of illustrative examples that range from very simple to more complex application problems. We propose some simple statistical models for such problems, describe their finite-sample analysis perspectives, and, with this in mind, stress the need for asymptotic methods. In Section 1.3, we go further along this line, specifying more realistic models for the problems described previously as well as for those arising from additional practical examples and fortifying the transit from the limited scope of finite sample inference to omnibus asymptotic methods. In Section 1.4, we present a brief description of the basic coverage of the book. We conclude with a summary of some basic tools needed throughout the text.

1.2 Illustrative Examples and Motivation

In general, the strategy employed in statistical inference involves (i) the selection of an appropriate family of *stochastic models* to describe the characteristics under investigation, (ii) an evaluation of the compatibility of chosen model(s) with the data (*goodness of fit*), and (iii) subsequent *estimation* of, or *tests of hypotheses* about, the underlying *parameters*. In this process, we are faced with models that may have different degrees of complexity and, depending on regularity assumptions, different degrees of restrictiveness. They dictate the complexity of the statistical tools required for inference which, in many instances, call for asymptotic perspectives because conventional finite-sample methods may be inappropriate. Let us first appraise this scenario of statistical models through some illustrative examples.

Example* 1.2.1 *(Inspection sampling). Suppose that we are interested in selecting a random sample to estimate the proportion of defective items manufactured in a textile plant. The data consist of the number x of defective items in a random sample of size n. For a random variable X representing the number of defective items in the sample, four possible stochastic models follow.

(a) We assume that the items are packed in lots of N (known) units of which D (unknown) are defective, so that our interest lies in estimating $\pi = D/N$. The probability of obtaining x defective items in the sample of size n may be computed from the probability function of the hypergeometric distribution

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, \quad x = \max[0, n - (N - D)], \ldots, \min(n, D).$$

(b) If, on the other hand, we assume that the items are manufactured continuously and that the proportion of defective items is π, $0 < \pi < 1$, the probability of obtaining x defective items in the random sample of size n may be computed from the probability function of the binomial distribution,

$$P(X = x) = \binom{n}{x}\pi^x(1 - \pi)^{n-x}, \quad x = 0, \ldots, n.$$

(c) Alternatively, if under the same assumptions as in item (b), we decide to sample until r defective items are selected, the probability of sampling $x + r$ items may be

obtained from the probability function of the negative binomial distribution, that is,

$$P(X = x) = \binom{r + x - 1}{x} \pi^r (1 - \pi)^x, \quad x = 0, 1, 2, \ldots.$$

(d) In some cases, it is assumed that X has a Poisson distribution, namely,

$$P(X = x) = \frac{e^{-\lambda} \lambda^r}{r!}, \quad r \geq 0, \ \lambda > 0,$$

where $\lambda > 0$. □

Example* 1.2.2** ***(One sample location model). We consider the problem of estimating the average height μ of a (conceptually infinite) population based on a random sample of n individuals. Let Y denote the height of a randomly selected individual and F denote the underlying distribution function. In such a context, three alternative (stochastic) models include the following:

(a) F is assumed to be symmetric and continuous with mean μ and finite variance σ^2, and the observations are the heights $Y_1, \ldots, Y_n$ of the n individuals in the sample.

(b) F is assumed to be normal with mean μ and known variance σ^2, and the observations are like those in item (b).

(c) The assumptions on F are as in either (a) or (b), but the observations correspond to the numbers of individuals falling within each of m height intervals (grouped data). This is, perhaps, a more realistic model, because, in practice, we are only capable of coding the height measurements to a certain degree of accuracy (e.g., to the nearest millimeter). In this case, using a multinomial distribution with ordered categories would be appropriate. □

Example* 1.2.3** ***(Paired two-sample location model). Let $X(Y)$ denote the blood pressure of a randomly selected hypertense individual before (after) the administration of an antihypertensive drug and μ_X (μ_Y) represent the corresponding (population) mean. Our objective is to estimate the average reduction in blood pressure, $\Delta = \mu_X - \mu_Y$, based on a sample of n hypertense individuals randomly selected from a conceptually infinite or very large population for which

(i) both X and Y are observed for all n subjects in the sample;

(ii) there are missing observations, that is, X or Y or both are not recorded for all subjects. A common stochastic model for this problem assumes that the vector $(\boldsymbol{X}, \boldsymbol{Y})$ follows a bivariate normal distribution with $\mathbb{V}\text{ar}(X) = \sigma_X^2$, $\mathbb{V}\text{ar}(Y) = \sigma_Y^2$, and $\mathbb{C}\text{ov}(X, Y) = \sigma_{XY}$. □

Example* 1.2.4** ***(Multisample location model). A manufacturing company produces electric lamps wherein the cross section of the coil (say, X) may be one of s (≥ 2) possible choices, designated $x_1, \ldots, x_s$, respectively. Our objective is to verify whether the level x_i has some influence on the expected life of the lamps. For each level x_i, we consider a set of n_i lamps taken at random from a (very large) production lot, and let Y_{ij} denote the life length (in hours, say) corresponding to the jth lamp in the ith lot ($i = 1, \ldots, s$; $j = 1, \ldots, n_i$). It may be assumed that the Y_{ij} are independent for different i ($= 1, \ldots, s$) and j ($= 1, \ldots, n_i$) and follow distributions characterized by distribution functions F_i,

defined on $\mathbb{R}^+ = [0, \infty)$, $i = 1, \ldots, s$. As in Example 1.2.2, we may consider a variety of models for the F_i, among which we pose the following:

(a) F_i is normal with mean μ_i and variance σ_i^2, $i = 1, \ldots, s$, where the σ_i^2 may or may not be the same.
(b) F_i is continuous (and symmetric) with median ν_i and scale parameter τ_i, $i = 1, \ldots, s$, but its form is not specified.
(c) Although F_i may satisfy (a) or (b), because the Y_{ij} are recorded in class intervals (of width one hour, say), we must replace it by some appropriate (ordered) categorical data distribution, such as the multinomial distribution postulated in Example 1.2.2.(c).

Under model (a), if we assume further that $\sigma_1^2 = \cdots = \sigma_s^2 = \sigma^2$, we have the *classical (normal theory) multisample shift in location model* (or simply the *location model*); in this context, if $n_i = n$, $i = 1, \ldots, s$, the data are said to be balanced; in (b), if we let $F_i(y) = F(y - \nu_i)$, $i = 1, \ldots, s$, we have the so-called *nonparametric multisample location model*. Either (a) or (b) may be made more complex when we drop the assumption of homogeneity of the variances σ_i^2 or of the scale parameters τ_i or allow for possible scale perturbations in the location model, that is, if we let $F_i(y) = F[(y - \mu_i)/\sigma_i]$ or $F_i(y) = F[(y - \nu_i)/\tau_i]$, $i = 1, \ldots, s$, where the σ_i^2 (or the τ_i) are not necessarily the same. □

Example* 1.2.5** ***(Simple linear regression model). We consider a study designed to evaluate the association between the level of fat ingestion (X) on weight (Y) of children in a certain age group. The data consist of the pairs of measurements of X and Y for n randomly selected subjects (from a large conceptual population) in the appropriate age group, namely, (X_i, Y_i), $i = 1, \ldots, n$. A possible relation between X and Y may be expressed by the simple linear regression model

$$Y_i = \alpha + \beta X_i + e_i, \quad i = 1, \ldots, n,$$

where α and β are unknown parameters and the e_i are random errors with distribution function F. Under this model, X is, in general, assumed fixed and known without error. Again, many assumptions about the stochastic features of the model may be considered.

(a) The error terms e_i are assumed to be independent and to have mean 0 and constant variance σ^2, but the form of F is not specified. This is the *Gauss–Markov setup*.
(b) Additionally to the assumptions in (a), F is required to be normal.
(c) The assumptions in (a) are relaxed to allow for heteroskedasticity, that is, the variances may not be the same for all observations.
(d) Under (a), (b), or (c), X is assumed to be measured with error, that is, $X = W + u$, where W is a fixed (unknown) constant and u is a random term with mean 0 and variance τ^2. This is the *error in variables simple linear regression model*.
(e) In some cases, W in (d) is also a random variable with mean ν and variance σ_W^2. In that case, the regression of Y on W has slope $\gamma = \beta\sigma_W^2/(\sigma_W^2 + \tau^2)$. This is referred to as a *measurement-error model* [Fuller (1986)]. □

Example* 1.2.6** ***(Repeated measures). We suppose that in a study similar to that described in Example 1.2.5, the weight of each child (Y) is observed under p fixed, randomly assigned levels (X) of physical activity (e.g., light, medium, or intense). The data consist of $p > 1$ pairs of measurements of X and Y for each of n randomly selected subjects in the appropriate

age group, namely, (X_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, p$. Because the same response (Y) is observed two or more times on the same subject, this type of data is generally referred to as *repeated measures*. Under this setup, a commonly used model to express the relationship between Y and X is

$$Y_{ij} = \mu + \alpha_j + e_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p,$$

where μ and α_j are fixed constants and e_{ij} are random errors with distribution function F. Some possible assumptions for the stochastic components of the model follow:

(a) The error terms e_{ij} are normally distributed with mean 0 and covariance structure given by $\mathbb{V}\text{ar}(e_{ij}) = \sigma^2$, $\mathbb{C}\text{ov}(e_{ij}, e_{ij'}) = \sigma_a^2$, $j \neq j'$, and $\mathbb{C}\text{ov}(e_{i'j}, e_{ij'}) = 0$, $i \neq i'$.
(b) The assumptions are similar to those in (a), but heteroskedasticity is allowed, that is, $\mathbb{V}\text{ar}(e_{ij}) = \sigma_i^2$.
(c) The assumptions are similar to those in (a), but F is allowed to be a member of the exponential family.
(d) The assumptions are as in (a), (b), or (c) but some observations are missing. □

Example* 1.2.7** ***(Longitudinal data). In the study described in Example 1.2.5, we suppose that measurements of fat ingestion (X) and weight (Y) are taken on each child at different instants in a one-year period. The data consist of triplets of the form (X_{ij}, Y_{ij}, T_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, p_i$, where T_{ij} denotes the instants at which the jth measurement on the ith child was taken. This is a special case of repeated measures data where the repeated observations are taken sequentially along an ordered dimension (time). To express the variation of Y with X, taking time into consideration, a useful model is the *random coefficients model*:

$$Y_{ij} = (\alpha + a_i) + (\beta + b_i)X_{ij} + \gamma T_{ij} + e_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p_i,$$

where α, β and γ are unknown parameters, and a_i, b_i, and e_{ij} are random terms. Again, we may consider different assumptions for the stochastic components of the model.

(a) The pairs (a_i, b_i) follow independent bivariate normal distributions with mean vector $\mathbf{0}$ and positive definite covariance matrix $\mathbf{\Sigma}$ and the e_{ij} follow independent normal distributions with means 0 and variances σ^2. Furthermore, e_{ij} is independent of (a_i, b_i). Because given (a_i, b_i), the Y_{ij} are independent, this model is usually referred to as the *conditional independence model*.
(b) The assumptions are like in (a), but a uniform structure is imposed upon $\mathbf{\Sigma}$ by letting the a_i (b_i) follow independent normal distributions with means 0 and variances $\sigma_a^2 = \sigma_b^2$ and $\mathbb{C}\text{ov}(a_i, b_i) = \sigma_{ab}$. □

Example* 1.2.8** ***(Generalized linear model). In a typical ecological study, we are interested in assessing the association between the concentration of a certain atmospheric pollutant like SO_2 (X) and the death count (Y) in a certain region, controlling for temperature (T) and relative humidity (H). The data consist of vectors (Y_i, X_i, T_i, H_i), $i = 1, \ldots, n$, containing the measurements of all variables along a period of n days.

The relation of the response variable (Y) and the explanatory variables (X, T, H) may be postulated as

$$g[\mathbb{E}(Y_i)] = \beta_0 + \beta_1 X_i + \beta_2 T_i + \beta_3 H_i, \quad i = 1, \ldots, n,$$

where β_j, $j = 0, 1, 2, 3$ are unknown regression coefficients and g is a convenient link function, for example, the logarithmic function. To specify the stochastic components of the model we may assume consider the following assumptions:

(a) The Y_i follow independent Poisson distributions, or more generally any distribution in the exponential class, in which case, the model may be classified as a *generalized linear model*.

(b) We do not specify the distribution of Y_i but assume that $\mathbb{V}\text{ar}(Y_i) = v[\mathbb{E}(Y_i)]$, where v is a given function known as the *variance function*. When v is such that $\mathbb{V}\text{ar}(Y_i) > \mathbb{E}(Y_i)$ we say that there is *overdispersion*. □

Example* 1.2.9 *(Dilution bioassay). We consider a bioassay experiment designed to compare the effects of a standard preparation (SP) and an experimental preparation (EP) on the occurrence of some response (Y), such as death or tumor onset. The data consist of pairs (X_{ij}, Y_{ij}), $i = 1, 2$, $j = 1, \ldots, n_i$ where X_{ij} denotes the dose applied to the jth subject submitted to the ith preparation ($1 = SP, 2 = EP$) and Y_{ij} is equal to 1 if the jth subject submitted to the ith preparation presents a positive response (e.g., death) and to 0, otherwise.

Let X_S denote the *dose* of SP above, which the response for a given subject is positive, and let X_T be defined similarly for the EP. The random variables X_S and X_T are called *tolerances*. Assume that X_S follows the (tolerance) distribution F_S defined on $\mathbb{R}^+$ and that X_T follows the distribution F_T, also defined on $\mathbb{R}^+$. Thus, $F_S(0) = 0 = F_T(0)$ and $F_S(\infty) = 1 = F_T(\infty)$. In many assays, it is conceived that the experimental preparation behaves as if it were a *dilution* (or *concentration*) of the standard preparation. In such a case, for $x \in \mathbb{R}^+$ we may set $F_T(x) = F_S(\rho x)$, where ρ (> 0) is called the *relative potency* of the test preparation with respect to the standard one. Note that for $\rho = 1$, the two distribution functions are the same, so that the two preparations are equipotent; for $\rho > 1$, the test preparation produces the same frequency of response with a smaller dose than the standard one and hence is more potent, whereas for $\rho < 1$, the opposite conclusion holds. Possible stochastic models follow:

(a) If we assume that F_S corresponds to a normal distribution with mean μ_S and variance σ_S^2, then F_T is normal with mean $\mu_T = \rho^{-1}\mu_S$ and variance $\sigma_T^2 = \sigma_S^2/\rho^2$. Thus, $\mu_S/\mu_T = \rho$ and $\sigma_S/\sigma_T = \rho$.

(b) Because of the positivity of the dose, a normal tolerance distribution may not be very appropriate, unless μ_S/σ_S and μ_T/σ_T are large. As such, often, it is advocated that instead of the dose, one should work with dose transformations called *dose metameters* or *dosages*. For example, if we take $X_T^* = \log X_T$ and $X_S^* = \log X_S$, the response distributions, denoted by F_T^* and F_S^*, respectively, satisfy

$$F_S^*(x) = F_T^*(x - \log \rho), \quad -\infty < x < \infty, \tag{1.2.1}$$

which brings in the relevance of linear models in such a context. □

Example* 1.2.10 *(Quantal bioassay). We consider a study designed to investigate the mutagenic effect of a certain drug (or a toxic substance). Suppose that the drug is administered at different dose levels, say $0 \leq d_1 < \cdots < d_s$, $s \geq 2$ so that at each level, n subjects are tried. We let m_j denote the numbers of subjects having a positive response under dose j. Here the response, Y, assumes the value 1 if there is a mutagenic effect and 0 otherwise.

Thus, the data consist of the pairs (d_j, m_j), $j = 1, \ldots, s$. Possible stochastic models are given below.

(a) At dose level d_j, we denote the probability of a positive response by $\pi_j = \pi(d_j)$. Then $\pi_j = P\{Y = 1|d_j\} = 1 - P\{Y = 0|d_j\}$, $1 \le j \le s$, and the joint distribution of $m_1, \ldots, m_s$ is

$$\prod_{j=1}^{s} \binom{n}{m_j} \pi_j^{m_j} (1 - \pi_j)^{n - m_j}, \quad 0 \le m_j \le n, \ 1 \le j \le s.$$

(b) As in the previous example, it is quite conceivable that there is an underlying tolerance distribution, say F, defined on $[0, \infty)$ and a threshold value (the tolerance), say T_0, such that whenever the actual dose level exceeds T_0, one has $Y = 1$ (i.e., a positive response); otherwise, $Y = 0$. Moreover, we may also quantify the effect of the dose level d_j by means of a suitable function, say $\beta(d_j)$, $1 \le j \le s$, so that

$$\pi_j = \pi(d_j) = 1 - F[T_0 - \beta(d_j)], \quad 1 \le j \le s.$$

With this formulation, we are now in a more flexible situation wherein we may assume suitable regularity conditions on F and $\beta(d_j)$, leading to appropriate statistical models that can be more convenient for analysis. For example, taking $x_j = \log d_j$, we may put $\beta(d_j) = \beta^*(x_j) = \beta_0^* + \beta_1^* x_j$, $1 \le j \le s$, where β_0^* and β_1^* are unknown and, as such,

$$1 - \pi_j = F(T_0 - \beta_0^* - \beta_1^* x_j), \quad 1 \le j \le s.$$

It is common to assume that the tolerance follows a logistic or a normal distribution. In the first case, we have $F(y) = [1 + \exp(-y/\sigma)]^{-1}$, $-\infty < y < \infty$, with σ (> 0) denoting a scale factor. Then, $F(y)/[1 - F(y)] = \exp(y/\sigma)$ or $y = \sigma \log\{F(y)/[1 - F(y)]\}$, implying that for $1 \le j \le s$,

$$\log \frac{\pi_j}{1 - \pi_j} = \log \frac{1 - F(T_0 - \beta_0^* - \beta_1^* x_j)}{F(T_0 - \beta_0^* - \beta_1^* x_j)} = \alpha + \beta x_j, \tag{1.2.2}$$

where $\alpha = (\beta_0^* - T_0)/\sigma$ and $\beta = \beta_1^*/\sigma$. The quantity $\log[\pi_j/(1 - \pi_j)]$ is called a *logit* (or *log-odds*) at dose level d_j. Similarly, when F is normal, we have

$$\pi_j = 1 - F(T_0 - \beta_0^* - \beta_1^* x_j) = \Phi(\alpha + \beta x_j), \quad 1 \le j \le s,$$

where $\Phi(y) = (\sqrt{2\pi})^{-1} \int_{-\infty}^{y} \exp(-x^2/2)\mathrm{d}x$. Therefore,

$$\Phi^{-1}(\pi_j) = \alpha + \beta x_j, \quad 1 \le j \le s, \tag{1.2.3}$$

and the quantity $\Phi^{-1}(\pi_j)$ is called a *probit* or a *normit* at the dose level d_j. Both logit and probit models may be employed in more general situations where we intend to study the relationship between a dichotomous response and a set of explanatory variables along the lines of linear models. For the logit case, such extensions are known as *logistic regression models*. In a broader sense, the models discussed above may be classified as *generalized linear models*. □

Example* 1.2.11** ***(Hardy–Weinberg equilibrium). We consider the OAB blood classification system (where the O allele is recessive and the A and B alleles are codominant). Let p_O, p_A, p_B and p_{AB}, respectively, denote the probabilities of occurrence of the phenotypes

OO, (AA, AO), (BB, BO), and AB in a given (conceptually infinite) population; also, we let q_O, q_A, and q_B, respectively, denote the probabilities of occurrence of the O, A, and B alleles in that population. This genetic system is said to be in *Hardy–Weinberg equilibrium* if the following relations hold: $p_O = q_O^2$, $p_A = q_A^2 + 2q_Oq_A$, $p_B = q_B^2 + 2q_Oq_B$, and $p_{AB} = 2q_Aq_B$. A problem of general concern to geneticists is to test whether a given population satisfies the Hardy–Weinberg conditions based on the evidence provided by a sample of n observational units for which the observed phenotype frequencies are n_O, n_A, n_B, and n_{AB}.

Under the assumption of random sampling, an appropriate stochastic model for such a setup corresponds to the multinomial model specified by the probability function

$$p(n_O, n_A, n_B, n_{AB}) = \frac{n!}{n_O!n_A!n_B!n_{AB}!} p_O^{n_O} p_A^{n_A} p_B^{n_B} p_{AB}^{n_{AB}},$$

with $p_O + p_A + p_B + p_{AB} = 1$. □

1.3 Synthesis of Finite to Asymptotic Statistical Methods

The illustrative examples in the preceding section are useful to provide motivation for some statistical models required for inference; even though exact inferential methods are available under many of such models, alternative approximate solutions are required when more realistic assumptions are incorporated. In this section, we build on the examples discussed earlier, as well as on additional ones, by considering more general statistical models that seem more acceptable in view of our limited knowledge about the data generation process. In this context, we outline the difficulties associated with the development of exact inferential procedures and provide an overview of the genesis of approximate large-sample methods along with the transit from their small-sample counterparts.

We start with Example 1.2.2(b). When F, the distribution of Y is assumed to be normal, with mean μ and variance σ^2, the sample mean $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ has also a normal distribution with mean μ and variance $n^{-1}\sigma^2$, so that the *pivotal quantity*

$$Z_n = \sqrt{n}(\overline{Y}_n - \mu)/\sigma \tag{1.3.1}$$

has the standard normal distribution, that is, with null mean and variance equal to 1. If μ, the unknown parameter, is of prime interest and the variance σ^2 is known, then Z_n in (1.3.1) can be used for drawing statistical conclusions on μ. In the more likely case of σ^2 being unknown too, for $n \geq 2$, we may take the sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2,$$

and consider

$$t_n = \sqrt{n}(\overline{Y}_n - \mu)/s_n, \tag{1.3.2}$$

which follows the Student t-distribution with $n-1$ degrees of freedom, for drawing statistical conclusions on μ. Both Z_n and t_n have distributions that are symmetric about 0, but t_n is more spread out than Z_n in the sense that

$$P(|t_n| > c) \geq P(|Z_n| > c), \quad \forall c > 0, \tag{1.3.3}$$

more dominantly when n is small. Thus, assuming σ^2 to be unknown, may result in less precise statistical conclusions when n is not large, the basic assumption of normality of F is crucial in this respect.

We consider now the dilution bioassay problem in Example 1.2.9, and suppose that we are interested in estimating the mean response Y of one of the preparations. Being a nonnegative random variable, Y has, typically, a positively skewed distribution on $\mathbb{R}^+ = [0, \infty)$. Thus, often, a logarithmic transformation ($Y^* = \log Y$) is advocated to achieve more symmetry for the distribution F^* (of Y^*), albeit there is still no assurance that F^* be normal. Indeed, in many applications, F^* is assumed to be logistic instead of normal. For a logistic distribution, the scale parameter and the standard deviation are not the same as in the case of the normal distribution, and the pivotal quantity Z_n (or t_n) may not be (even approximately) normal, more noticeably when n is small. Thus, assuming normality for F, when truly it is not normal, may induce a loss of precision of the statistical conclusions. For some simple specifications of the distribution function F, the *exact* sample distribution of $\overline{Y}_n$ may be obtained in closed form and that can be used to draw conclusions on $\mu = \mathbb{E}(\overline{Y}_n)$. This occurs, for example, when F belongs to the regular exponential family where optimal estimators [often characterized as *maximum likelihood estimators* (MLE)] may have a mean-like structure. However, for a general nonnormal F, the MLE is not $\overline{Y}_n$, and thereby the use of $\overline{Y}_n$ may entail further loss of statistical information about the parameter of interest. As an example, we consider the case of F as a Laplace distribution with location parameter μ and scale parameter $\lambda > 0$. The MLE of μ is the sample median, not the sample mean, and hence, inference based on $\overline{Y}_n$ may be less efficient. Recall that the logistic and the Laplace distributions are not members of the exponential family.

Two issues transpire from the above discussion. First, if some specific distributional assumption is made, how easy is it to extract full statistical information from the sample observations to draw optimal statistical conclusions? Computational complexity or mathematical intractability generally dominate the scenario, specially for nonnormal F. Second, how sensitive is the derived statistical conclusion to plausible departures from the assumed model? In many instances, they may be severely affected even by a single *outlier*. The first issue dominated the statistical literature from the 1930s to 1960s with some hope that advances in computer technology would make the situation more tractable. Nevertheless, the second issue has raised some basic questions regarding *robustness* of classical (finite-sample) parametric inference procedures to plausible model departures, either in a local sense, or more typically in a global sense; developments in nonparametric and robust statistical procedures have their genesis in this complex. We illustrate these points with some additional examples.

***Example* 1.3.1.** In Example 1.2.2.(a), we assumed that F was symmetric and continuous with mean μ and variance $\sigma^2 < \infty$. We consider several alternative models that incorporate different levels of knowledge about the nature of the data:

(a) Let $F(x) = F_o[(x - \mu)/\lambda], x \in \mathbb{R}$, with μ and $\lambda > 0$ being, respectively, the *location* and *scale parameter* and let the distribution F_o be free from (μ, λ). The normal, logistic, Laplace, and Cauchy distributions are all members of this *location-scale family*, denoted by $\mathcal{F}_{LS}$.

(b) Let $\mathcal{F}$ be the class of all distributions defined on $\mathbb{R}$ with finite first- and second-order moments. Note that $\mathcal{F}$ includes $\mathcal{F}_o$, the class of all symmetric distributions

having finite second-order moments, as a subclass. In this setup, the parameter μ (or σ^2) may be viewed as a *functional* $\theta(F)$ of the distribution $F \in \mathcal{F}$. Many other parameters may be expressed in this form.

(c) Consider the subclass $\mathcal{F}_{oS}$ of all symmetric distributions defined on $\mathbb{R}$ with the origin of symmetry regarded as the location parameter. In this setup, the location parameter is the *median* of F, and will be its mean whenever F admits a finite first-order moment.

In (a) exact statistical inference pertaining to (μ, λ) may be obtained in some cases when the functional form of F_o (in a parametric setup) is specified; however, when F_o is treated nonparametrically (i.e., without specification of its functional form), approximate methods are called for. In (b) or (c), the generality of the assumptions clearly point toward the need for nonparametric methods. In this regard, the *nonparametric estimation* of statistical functionals developed by Halmos (1946) and Hoeffding (1948) constitutes a paradigm. □

Example* 1.3.2** ***(Quantile function). For a continuous distribution F defined on $\mathbb{R}$, and for every p $(0 < p < 1)$, we let

$$Q_F(p) = \inf\{x : F(x) \geq p\}, \tag{1.3.4}$$

be the p-quantile of F. The function $Q_F = \{Q_F(p) : 0 < p < 1\}$ is called the quantile function. Nonparametric measures of location and dispersion are often based on Q_F. For example, $Q_F(0.5)$, the *median* is a measure of location, while the *interquartile range* $Q_F(0.75)$–$Q_F(0.25)$ is a measure of *dispersion*.

Even in a parametric setup [e.g., (c) of Example 1.3.1)], statistical inference under this model must rely on linear functionals of Q_F (called linear combinations of *order statistics*) for which only approximate (large-sample) methods are available. □

Example* 1.3.3** ***(Contamination model). Instead of letting $F \in \mathcal{F}$, we consider a model that allows for some local departures from a stipulated distribution $F_o \in \mathcal{F}$ (usually normal). For example, we take for some (small) $\varepsilon > 0$ and let

$$F(x) = (1 - \varepsilon)F_o(x) + \varepsilon H(x), \quad x \in \mathbb{R}, \tag{1.3.5}$$

where H has a heavier tail than F_o in the sense of (1.3.3).

Depending on the divergence between F_o and H, even a small $\varepsilon(> 0)$ can create a great damage to the optimality properties of standard (exact) statistical inference procedures and, again, we must rely on approximate (large-sample) statistical methods. This is the genesis of *robustness* in a local sense as discussed in Huber (1981), for example. □

Example* 1.3.4** ***(Nonparametric regression). In Example 1.2.8 we consider a generalized linear model where the regression function is specified by means of a finite dimensional vector $\boldsymbol{\beta}$ of regression parameters and known explanatory variables. In a more general setup, corresponding to a (possibly vector valued) regressor $\boldsymbol{Z}$ $(\in \mathbb{R}^q$, for some $q \geq 1)$, we assume that the dependence of the response Y on $\boldsymbol{Z}$ is described by

$$m(\boldsymbol{z}) = \mathbb{E}(Y|\boldsymbol{Z} = \boldsymbol{z}) = \int y f(y|\boldsymbol{z}) \mathrm{d}y, \quad \boldsymbol{Z} \in \mathbb{R}^q, \tag{1.3.6}$$

where f, the conditional density (or probability function) of Y given $\boldsymbol{Z}$ does not have a specified functional form. The objective is to draw statistical conclusions about $m(\boldsymbol{z})$,