1 Introduction and Summary

The purpose of this chapter is to introduce the concept of *sampling* and to present some distribution-theoretical results that are engendered by sampling. It is a connecting chapter — it merges the distribution theory of the first five chapters into the statistical theory of the last five chapters. The intent is to present here in one location some of the laborious derivations of distributions that are associated with sampling and that will be necessary in our future study of the theory of statistics, especially estimation and testing hypotheses. Our thinking is that by deriving these results now, our later presentation of the statistical theory will not have to be interrupted by their derivations. The nature of the material to be given here is such that it is not easily motivated.

The emphasis in this book is on the *theory of statistics* as opposed to *data analysis*. As is often the case a single word has multiple meanings and such is the case with *statistics*. There are at least three meanings of statistics:

(i) a *collection of numbers*, as in the batting averages of all major league baseball players at the All Star break; the daily temperature highs and lows for major U.S. cities; the daily listing of number of shares traded, high, low, and closing price of stocks traded on the New York Stock Exchange; etc.;

(ii) The *discipline* or *subject* as in mathematics, chemistry, physics, etc. One might define *statistics* as the science and art of collecting, manipulating, analyzing, interpreting, and/or presenting information (often numerical) usually with intent of drawing inferences;

(iii) *functions of the "data"* (collection of numbers) as an arithmetic average of a set of numbers.

All three meanings will be used. In *data analysis*, one analyzes the data of meaning (i). Data analysis is often descriptive including such techniques as graphs, box-plots, etc., with no assumed "structure" of the data. In contrast, the *theory of statistics* as used here entails the assumption that the data can be viewed as the value of some random vector. *Modeling* of the data will encompass all assumptions made about the random vectors that the data is a value of.

Section 2 begins by introducing the language of the theory of statistics including such concepts as *sample, population, sample moments*. Sample moments are important and useful statistics in the sense of meaning (iii). Two types of results are presented: those associated with a *fixed sample size* (often referred to as "small" sample results) covered in Sections 3 and 4, and those associated with *increasing sample size* (often labeled "large" sample or limiting/asymptotic results) covered in Section 5. Section 3 considers general fixed sample size results. The sample or empiric distribution function is studied as well as sample moments and *sample quantiles* or *order statistics*. Order statistics, like sample moments, are important and useful statistics. Sampling from the normal distribution is considered in Section 4 where the chi-square, F, and t, distributions are introduced. Finally, asymptotic distributions for sample moments and sample quantiles are addressed in Section 5. Included is an introduction to *extreme value theory*. These results have important applications that will appear in later chapters; for example, the justification for certain confidence intervals

frequently used in the daily practice of statistics (meaning (ii)) rests on some of these limiting results.

2 Sampling

2.1 Samples and Modeling

In the study of probability theory in the earlier chapters, the link to the real world was the so-called conceptual experiment that led to the sample space which in turn led to the probability function, random variables, etc. In the coming study of statistics such link is provided by *data*, a collection of numbers as in meaning (i) of statistics in the previous section.

Let $x_1, \ldots, x_n$ be a generic notation for the "data," also called the *observed sample* (or just *sample*), where n is the generic notation for *sample size*. (n is the index needed in sample-size-increasing in the limiting results of Section 5.) Each x_i could be a vector. For example consider n individuals involved in some sport. On each individual one might observe the height, weight, amount bench pressed, 0_2 uptake, etc. Or from the academic arena, the GPA, GRE scores, class rank, etc., could be observed for n applicants for graduate admission.

The first piece of structure that takes us into the realm of *theory of statistics* is:
Assume that $x_1, \ldots, x_n$ *is a value of r.v.'s* $X_1, \ldots, X_n$; i.e., we are willing to consider our "data" as a value of a r.v. (r.v. is usually random vector here) — such structure is not needed for much of data analysis. - -

> Definition 1 $X_1, \ldots, X_n$ is called the *sample* and $x_1, \ldots, x_n$ is called the *observed sample* or *data*. If $X_1, \ldots, X_n$ are iid, then $X_1, \ldots, X_n$ is called a *random sample* from the common distribution of the X_i's. ////

Comment: We are often sloppy and fail to distinguish between a r.v. and its value; i.e., we call $X_1, \ldots, X_n$ and $x_1, \ldots, x_n$ our sample (without proper modifier). ////

The next piece of structure is to *model* $X_1, \ldots, X_n$. Here "modeling" is assuming something about the distribution of the sample $X_1, \ldots, X_n$.

Notation: Write $\underline{X}_n$ or $\underline{X}$ (depending on whether or not, respectively, sample size is important to our discussion) for $X_1, \ldots, X_n$ and $\underline{x}_n$ or $\underline{x}$ for $x_1, \ldots, x_n$.

> Definition 2 A *model* is an assumed family of distributions for $\underline{X}_n$ or $\underline{X}$, generically called $\mathcal{F}_n$ or $\mathcal{F}$. If $\mathcal{F} = \{F(\cdot;\theta)$ or $p(\cdot;\theta)$ or $f(\cdot;\theta) : \theta \in \overline{\underline{\theta}}\}$, where $\overline{\underline{\theta}}$ is some subset of Euclidean space, then we speak of a *parametric* model. $\overline{\underline{\theta}}$ is the *parameter space* and θ the *parameter*. ////

2 SAMPLING

The model consists of an assumed family of distributions for $\underline{X}_n$. If the $X_1, \ldots, X_n$ in $\underline{X}_n$ are assumed iid then one could describe the model by stating the iid assumption and then giving a family of distributions for the common distribution of the X_i's and this is what is done.

The "game" of "statistical inference" as part of the "theory of statistics" is: based on the "data" and assumed model, make some inference regarding the model; for example, estimate the parameter θ. We will study three modes of inference — point estimation (Chapter VII), interval estimation (Chapter VIII), and hypotheses testing (Chapter IX).

EXAMPLES OF MODELS:

(1)
$$\left.\begin{array}{l} X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot;\theta), \theta \in \underline{\theta} \\ \qquad\qquad\quad \text{known} \qquad\quad \text{known} \\ \qquad\qquad\qquad\quad \text{unknown} \end{array}\right\} \quad \text{model}$$

Here $x_1, \ldots, x_n$ consists of n independent observations from $f(\cdot;\theta)$ for θ fixed but unknown.

(2)
$$\underbrace{\begin{pmatrix} Y_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ Z_n \end{pmatrix}}_{=\underline{X}_n} \overset{iid}{\sim} \text{BVN}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

Here we have a joint (bivariate) parametric model. The parameter space is:

$$\underline{\theta} = \{(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) : \sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1\}$$

(3) Two samples — say
$$\begin{array}{lll} Y_1, \ldots, Y_{n_1} & \overset{iid}{\sim} & f_1(\cdot;\theta_1) \text{ and} \\ Z_1, \ldots, Z_{n_2} & \overset{iid}{\sim} & f_2(\cdot;\theta_2) \end{array}$$

where $\underline{Y}$ is independent of $\underline{Z}$.

Now $\underline{X} = (\underline{Y}, \underline{Z})$ and our generic sample size index n is not so clear. (We might have n increase as either n_1 or n_2 or both increase.)

(4) Multiple samples — e.g., a 'space' and 'time' indexed quantity, say $X_{\text{space,time}}$, as space and time range over some index set, is the sample; e.g., suppose we have k lakes numbered $1, \ldots, k$ and acidity measurements in time at each lake.

$$\begin{array}{lll} \text{lake } 1 & : & X_{1,t_1}, X_{1,t_2}, \ldots, X_{1,t_{n_1}} \\ \vdots & & \\ \text{lake } k & : & X_{k,t_1}, X_{k,t_2}, \ldots, X_{k,t_{n_k}} \end{array}$$

(time 'readings' at lake 1 may not be the same time 'readings' as lake k). Various assumptions regarding the distribution of the random vectors could be made.

(5) Identically distributed but not independent random variables; e.g.,

a) AR(1) where $\rho_{X_i,X_j} = \rho^{|i-j|}$

b) equicorrelated where $\rho_{X_i,X_j} = \rho$

for $i \neq j$, and $-1 < \rho < 1$.

(6) Independent but not identically distributed random variables.

EXAMPLE 1 Simple linear regression model $Y_z = \beta_0 + \beta_1 z + E$, where z is a real number, β_0 and β_1 (scalar) parameters and E is a r.v.

sample: $= \underline{X}_n$ is $(z_1, Y_{z_1}), \ldots, (z_n, Y_{z_n})$ with model: $Y_{z_i} = \beta_0 + \beta_1 z_i + E_i$ and $E_1, \ldots, E_n \sim$ as something. (Under normal theory $E_1, \ldots, E_n \overset{iid}{\sim} N(0, \sigma^2)$ and σ^2 is another parameter.) ////

(7) a non-parametric model:

$$X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot) \text{ and(e.g.) } f(\cdot)$$

is just assumed to be symmetric. We cannot 'parameterize' with a point in Euclidean space, hence a "non-parametric model." ////

Definition 3 The *distribution of the sample* is the distribution of $\underline{X}$ (different than "sample distribution" to be defined soon). ////

EXAMPLE 2 If $X_1, \ldots, X_n$ is random sample from $f(\cdot)$ then the distribution of the sample is given by $f_{\underline{X}}(x) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \Pi_{i=1}^n f(x_i)$. ////

2.2 Statistics

Definition 4 A *statistic* is a function of $\underline{X}$ (or $\underline{x}$) (here, again, we use the term "statistic" for r.v. or value of r.v.) generically denoted by $t(\cdot) = t(\cdot, \ldots, \cdot)$. ////

Notation: Set $\mathcal{X} = \{\underline{x} : \underline{x} \text{ is possible value of } \underline{X} \text{ under model}\}$. $\mathcal{X}$ is called the *potential data set.*

Remark For a statistic a function of $\underline{x}$, think $\mathcal{X} \xrightarrow{t(\cdot)} \mathrm{T}$, where T is usually some subset of some Euclidean space.

Remark For $X_1, \ldots, X_n$, where X_i is a r. variable (as opposed to a r. vector), $\overline{X} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, the *sample mean*, is a statistic, and $S_n^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, the *sample variance*, is a statistic. (We will return to these.) We also could write $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; 'statistic' is the function that says: $\frac{1}{n}$ times the sum of the x_i's, and T is the real line (or a subset of the real line under some models). ////

In our 'generic' statistic, we really have four different 'tees'. They are:

$t(\cdot)$, script l.c. tee, stands for the function in $\mathcal{X} \xrightarrow{t(\cdot)} \mathbf{T}$
T, bold u.c. tee, stands for the range of the function $t(\cdot)$
T, cap tee, stands for r.v. in $T = t(X_1, \ldots, X_n)$, and
t, l.c. tee, stands for value of r.v. T, or $t = t(x_1, \ldots, x_n)$, or, t is a point in T.

Statistics as defined here, and used in the sense of meaning (iii) in Section 1, will be used throughout the remainder of this text. For example, recall the three modes of statistical inference advertized in Section 2.1; a point estimator will be a statistic, an interval estimator will be two statistics, one less than the other, used to identify an interval, and a test statistic will be used to test a hypothesis.

2.3 Sample *versus* Population

A common term in statistical jargon is that of *population*, which is difficult to precisely define, possibly because it is used in so many ways. Population is linked to the model. We say that the *sample* is used (often through a statistic) to learn something about the *population*. The following "definition" illustrates various uses of the term population for one case.

Definition 5 Suppose $X_1, \ldots, X_n$ are identically distributed with range (under the model) $\mathbb{X}$; that is, $\Omega \xrightarrow{X_i(\cdot)} \mathbb{X}$; then

(i) $\mathbb{X}$ is often called the *population*; it is the possible values of any of the X_i's.

(ii) Ω itself is sometimes called the *population*.

(iii) Even the distribution of the X_i is called the *population*.

'Population' is also used as an adjective; e.g.,

(i) *population* distribution for the distribution of X_i.

(ii) *population* mean for $\mathcal{E}[X_i]$.

(iii) *population* variance for $\text{var}[X_i]$, etc. ////

Admittedly, population has not been completely defined — it is part of the 'background structure' imbedded in the model.

As mentioned, we use the 'sample' to learn something about the 'population,' which is the framework for the "population" *versus* "sample" contrast. Restrict now to a sample $X_1, \ldots, X_n$ of random variables (not vectors) with 'model' that says $X_1, \ldots, X_n$ are identically distributed; the following table lists some population and sample companions, as well as their generic notations.

All the expressions in the 'sample' column are statistics. We use these 'sample' statistics to learn something about their 'population' counterparts. And we look at "small/fixed" sample results (Sections 3 and 4) and "large/asymptotic (as $n \longrightarrow \infty$)" results (Section 5).

	population	sample
mean:	$\mu = \mathcal{E}[X_i]$	$\overline{X}_n = \overline{X} = \frac{1}{n}\sum_1^n X_i$
variance:	$\sigma^2 = \text{var}[X_i]$	$S_n^2 = S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X}_n)^2$
raw moments:	$\mu_r' = \alpha_r = \mathcal{E}[X_i^r]$	$M_r' = \frac{1}{n}\sum_1^n X_i^r$
central moments:	$\mu_r = \beta_r = \mathcal{E}[(X_i - \mu)^r]$	$M_r = \frac{1}{n}\sum_1^n (X_i - \overline{X}_n)^r$
skewness coefficient:	$\frac{\mu_3}{\sigma^3}$	$M_3/M_2^{3/2}$
kurtosis coefficient:	$\frac{\mu_4}{\sigma^4}$	M_4/M_2^2
or	$\frac{\mu_4}{\sigma^4} - 3$	$(M_4/M_2^2) - 3$
coefficient of variation:	$\frac{\sigma}{\mu}$	$\frac{S}{\overline{X}}$
quantile:	ξ_q such that $F_{X_i}(\xi_q) = q$	"qth order statistic" (defined later)
distribution:	$F_{X_i}(\cdot)$	sample cdf $F_n(\cdot)$ (defined later)
For $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim$ identically		
correlation:	$\rho = \frac{\text{cov}\,(X_i, Y_i)}{\sigma_{X_i}\sigma_{Y_i}}$	$\frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{S_X\, S_Y}$ or $\frac{\sum (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$