Hung T. Nguyen
Tonghui Wang

概率统计高级教程

Ⅱ统计学基础

A GRADUATE COURSE IN PROBABILITY AND STATISTICS

Volume Ⅱ
Essentials of Statistics

Tsinghua University Press

Hung T. Nguyen
Tonghui Wang

概率统计高级教程

Ⅱ统计学基础

A GRADUATE COURSE IN OBABILITY AND STATISTICS

Volume Ⅱ
Essentials of Statistics

Tsinghua University Press

Bei Jing

图书在版编目 (CIP) 数据

概率统计高级教程. 第 2 卷, 统计学基础 /(美) 源亨 (Hung T. Nguyen), (美) 王通惠 (Tonghui Wang) 编著. —北京：清华大学出版社，2009.4

书名原文：A Graduate Course in Probability and Statistics. Volume II: Essentials of Statistics

ISBN 978-7-302-19501-6

I. 概 ··· II. ① 源 ···② 王 ··· III.① 概率论—教材 ② 数理统计—教材 ③ 统计学—教材 IV.O21 C8

中国版本图书馆 CIP 数据核字 (2009) 第 019319 号

责任编辑：刘　颖
责任校对：刘玉霞
责任印制：王秀菊
出版发行：清华大学出版社　　地　址：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　邮　编：100084
社 总 机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：三河市金元印装有限公司
经　销：全国新华书店
开　本：165×230　印　张：26.5　字　数：603 千字
版　次：2009 年 4 月第 1 版　　印　次：2009 年 4 月第1 次印刷
印　数：1～2500
定　价：49.00 元

Preface

This Volume II is the second half of a text for a course in statistics at the beginning graduate level. Statistics is a man-made science aiming at assisting humans in making decisions in the face of uncertainty. This science is built upon the rigorous theory of probability as described in Volume I. Thus, in studying this text, students should consult Volume I whenever needed.

As stated in the preface of Volume I, there are various reasons to write another text in statistics at the introductory level. An obvious reason is to make the topic of statistics pleasant for students!

In an introductory course in statistics such as this one, one can only include basic ideas, concepts, procedures and applications at a very standard level. By this we mean that only the topics of estimation, hypothesis testing and prediction are included. Also, all inference procedures are developed for the standard type of data, namely precise observations which are numerical or vector-valued. The students should easily recognize that it is the data which dictate the developed statistical procedures in this text. Thus, other types of data, such as censored data in survival analysis, missing data in questionnaires, coarse data in biostatistics, imprecise data (or partially observed data, such as those occurring in the problem of identification of DNA sequences in bioinformatics, using hidden Markov models), and perception-based data (which are expressed linguistically) will not be discussed. However, the methodology for precise data clearly indicates the general framework for analyzing other types of data. After all, statistics is a science of data analysis.

With the rapid advances of technology, the use of statistics has been extended to many new emerging applications, both in physical and social sciences. The text does not cover these new statistical techniques. The text is written as a pedagogical source for instruction at universities. A solid understanding of statistics, at the simplest level, will open the door for embarking on any new problems which call for statistical assistance.

We thank our families for their love and support during the preparation of this text. Our Department of Mathematical Sciences at New Mexico State University provided us with a constraint-free environment for carrying out this project. We thank Dr. Ying Liu of Tsinghua University Press for asking us to write this two-volume text for Tsinghua University Press. We would like to thank our Ph. D. student Yanhong Tong who created all statistical tables for Volume II and Dr. Baokun Li who gave us many valuable suggestions for both volumes. Finally, we thank graduate students who took our statistics courses (Volumes I and II) at New Mexico State University in 2005-2008 for their comments during the preparation our textbooks.

Hung T. Nguyen and **Tonghui Wang**
Las Cruces, New Mexico, USA
August 2008

Contents

Chapter 1

An Invitation to Statistics

This introductory chapter aims at answering three basic questions concerning the topic of statistics, namely "WHAT is statistics?", "WHY do we need statistics?", and "HOW to carry out statistical analysis?".

This text is about the foundation of the science of statistics. Statistics is a body of concepts and techniques to carry out *inductive logic* in almost all activities of our daily lives. Although the applied concepts of the theory, such as *experiment designs, sampling methods, and data analysis*, will not be discussed in a text such as this, we feel obligated to introduce the students to the field of statistics from what statistics is created for.

1.1 A Motivating Example

Suppose that we are interested in the annual income of individuals in the population of Las Cruces, say, in 2004. Suppose that, for some reasons (such as cost and time), we are unable to conduct a *census* (i.e. a complete enumeration) throughout the whole population, and hence we could rely only on the information about the income from a part of that population. Of course, before going out to do that, we need to prepare the ground carefully. Specifically, first we need to decide who to be included in the population. Since the *variable* of interest is the annual income, we should exclude, for example, children who do not work from the population. Next, we should worry about whether or not when asking (by phone or by sending out questionnaires) selected individuals, their answers are with or without errors. Then, in going out to select a *sample*, a part of the population, we might want to conduct the *survey* in some beneficial way, e.g. by dividing the geography of the city into appropriate zones. All that is part of what we call the *design of experiments.* For this applied topic, see a text like Dean and Voss (1999).

Suppose that the *physical* population of individuals is identified as a finite set $U = \{u_1, u_2, \cdots, u_N\}$, where N is the *population size*. Our *variable* of interest is θ, the annual income. We will use $\theta(u_k)$ to denote the annual income of the individual u_k. Thus θ is a map from U to $\mathbb{R}$, i.e. $\theta : U \to \mathbb{R}$. The map (or function) θ is *unknown* but fixed at the outset. We are going to obtain partial knowledge about θ by conducting a *sampling survey*, i.e. select a *sample* A from U and discover the value of θ from A. In this obvious situation, A is a *subset* of U (in statistical parlance, we select a sample by "drawing" without replacement, and the order of drawings does not matter). From the knowledge of the restriction of θ to A, we wish to "guess" or estimate θ, or some functions of it, e.g. the *population total*

$$\tau(\theta) = \sum_{u \in U} \theta(u) = \sum_{k=1}^{N} \theta(u_k).$$

This is *inductive logic*: making statements about the whole population U from the knowledge of a part of it. Then the basic question is: *How to make this inductive logic valid?* For example, how do we know that, say,

$$\sum_{u \in A} \theta(u) \text{ is a } \textbf{good} \text{ estimate of } \tau(\theta)?$$

Can we specify the error in our estimation process? Obviously, questions such as these are related to the *quality of the data* we collected, e.g. does our data (i.e. the values of θ in our selected sample A) *representative* or *typical* for the whole population? Thus it all boils down to "how to select a *good* sample?". It seems that to eliminate bias in the selection of samples, and to gain public acceptance (with regard to objectivity), we could select samples at *random*. For example, if we decide to select a sample of size n, then any subset of size n of U should have the *same chance* to be selected, which is $1/\binom{N}{n}$. While the population and our variable of interest θ have nothing to do with randomness, we introduce a *man-made randomization* into our process of the sample selection in the hope of making our intended inductive reasoning valid. In other words, we create a *chance model*. As we will see, by doing so, we will obtain more than just getting a "good" data set, namely we will be able to assess the qualities of our estimation procedures.

Now observe that when we select samples according to a *probability sampling scheme* (or plan), we actually perform a random experiment (with known structure, like a game of chance) whose outcomes are samples which are subsets of the population (say, in the case of sampling without replacement and the order does not matter). In Volume I, we have that a random element whose values are subsets of some set is called a *random set*. Thus, formally, a probability sampling design is a random set since samples are obtained at random. The distribution of the random set S is given as a bona fide probability density function (or density) on the space of all subsets of the finite population U, the power set of U, which is denoted as $\mathcal{P}(U)$. It is this given density function which allows us to select samples in some random fashion. The choice of such a density depends on practical problems at hand and is left to applied statisticians!

Before making further statistical models, we see that, at a very primitive level of induction, randomness, and hence probability theory, enters the picture. It provides us with a good framework to carry out the inductive logic for applications. Specifically, let

$$f : \mathcal{P}(U) \longrightarrow [0,1], \qquad f(A) = P(S = A), \qquad A \subseteq U$$

be the density function of the random set S. In our example, $\theta : U \to \mathbb{R}$ is unknown, and is referred to as a *population parameter.* The population total $\tau(\theta)$ is referred to as a *parametric function.* Various aspects of *statistical inference* (i.e. inductive logic using probability theory) can be then properly formulated. We can consider an abstract probability space $(\Omega, \mathcal{A}, P)$, or just $(\mathcal{P}(U), \mathcal{P}(\mathcal{P}(U)), P_f)$, where $\mathcal{P}(\mathcal{P}(U))$ is the σ-field of all subsets of $\mathcal{P}(U)$, and P_f is the probability measure on $\mathcal{P}(\mathcal{P}(U))$, induced by f, i.e.

$$P_f(\mathbb{A}) = \sum_{A \in \mathbb{A}} f(A), \qquad \text{for} \quad \mathbb{A} \in \mathcal{P}(\mathcal{P}(U)).$$

For example, the expected sample size is

$$E(\#(S)) = \sum_{A \subseteq U} \#(A) f(A).$$

To illustrate an estimation problem, consider the target $\tau(\theta)$. Let S be a random sample selected according to the random mechanism generated by f. Then we could propose a "good" estimator for $\tau(\theta)$ as some function of S, T_S, such that

$$E(T_S) = \tau(\theta), \qquad \text{for all} \quad \theta : U \to \mathbb{R}.$$

Note that the requirement "for all $\theta : U \to \mathbb{R}$" is necessary since our actual θ is unknown. For example,

$$T_S = \sum_{u \in S} \frac{\theta(u)}{\pi(u)},$$

where

$$\pi(u) = P(\omega : u \in S(\omega)) = \sum_{\substack{A \subseteq U \\ u \in A}} f(A),$$

provided, of course, that $\pi(u) > 0$ for all $u \in U$.

1.2 Generalities on Survey Sampling

As we will see, the theory of statistical inference developed in this section is traditional (or standard) in the sense that the statistical data are assumed to be a collection of *independent and identically distributed* (i.i.d.) random variables. However, students should be aware of "classical" or "practical" aspects of statistical applications. For this reason, we intend to mention here the area of survey sampling and its statistical inference.

The framework for survey sampling is very simple. Let U be a finite population, say, $U = \{1, 2, \cdots, N\}$. As stated in the previous section, a *probability sampling design* is a density f on $\mathcal{P}(U)$, i.e.

$$f : \mathcal{P}(U) \longrightarrow [0, 1] \qquad \text{such that} \quad \sum_{A \subseteq U} f(A) = 1.$$

Let S be a random set with density f, defined on $(\Omega, \mathcal{A}, P)$, or just the identical map defined on the probability space $(\mathcal{P}(U), \mathcal{P}(\mathcal{P}(U)),\ P_f)$. The density f induces *covering functions* for subsets of U. For $j \in U$, let $\pi(j)$ denote the probability that j will be included in a sample "drawn" according to f, i.e.

$$\pi(j) = P(j \in S) = \sum_{\substack{A \subseteq U \\ j \in A}} f(A).$$

We can write $\pi(j) = \pi(\{j\})$ and call $\pi(\cdot)$ the *one-point coverage function* (or *first order probabilities of inclusion*) of S (or of f). By abuse of notation, we write

$$\pi(i, j) = \pi(\{i, j\}) = \sum_{\substack{A \subseteq U \\ \{i,j\} \subseteq A}} f(A)$$

to be the *two-point coverage function* (or *second order probabilities of inclusion*), and more generally, $\pi(A)$ for $A \subseteq U$. Of course, if $\pi(A)$ is known for *any* $A \subseteq U$, then f can be recovered (exercise). These covering functions are similar to moments of random variables.

In applications, it is desirable to specify the one-point coverage function $\pi(\cdot)$ and look for f having precisely $\pi(\cdot)$ as its one-point coverage function. We will discuss shortly the role played by probabilities of inclusion in statistical inference in survey sampling.

The key point of analysis is the introduction of *Bernoulli random vectors*. For each $j \in U$, let $I_j : \mathcal{P}(U) \longrightarrow \{0, 1\}$ be a Bernoulli random variable with parameter

$$P(I_j = 1) = P_f\{A \subseteq U : I_j(A) = 1\} = \pi(j),$$

where

$$I_j(A) = \begin{cases} 1 & \text{if} \quad j \in A \\ 0 & \text{if} \quad j \notin A. \end{cases}$$

For $\#(U) = N$, we consider the random vector $(I_1, I_2, \cdots, I_N)$. We note that $\pi(\cdot)$ is simply a function from U to $[0,1]$. Now the density f on $\mathcal{P}(U)$ is "equivalent" to the *joint distribution* of the Bernoulli random vector $(I_1, I_2, \cdots, I_N)$. This can be seen as follows. Making the bijection between $\mathcal{P}(U)$ and $\{0,1\}^N$:

$$A \longleftrightarrow (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N) = \boldsymbol{\varepsilon}$$

with $A_{\boldsymbol{\varepsilon}} = \{j \in U : \varepsilon_j = 1\}$, we have

$$f(A_{\boldsymbol{\varepsilon}}) = P(I_1 = \varepsilon_1, I_2 = \varepsilon_2, \cdots, I_N = \varepsilon_N), \quad \boldsymbol{\varepsilon} \in \{0,1\}^N.$$

Thus, if we specify a function $\pi : U \to [0,1]$, then the Bernoulli random variables I_j with parameters $\pi(j)$ have *fixed marginal distributions*.

As such, their *joint distributions* are determined by *N-copulas* according to *Sklar's theorem* (Volume I). Specifically, let F_j be the distribution of I_j, namely,

$$F_j(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \pi(j) & \text{if } 0 \leqslant x < 1 \\ 1 & \text{if } x \geqslant 1. \end{cases}$$

Let C be an N-copula, then the joint distribution function of $(I_1, I_2, \cdots, I_N)$ could be

$$F(x_1, x_2, \cdots, x_N) = C\left[F_1(x_1), F_2(x_2), \cdots, F_N(x_N)\right].$$

For example, by choosing

$$C(y_1, y_2, \cdots, y_N) = \prod_{j=1}^{N} y_j,$$

we obtain the well-known *Poisson sampling design*:

$$f(A) = \prod_{j \in A} \pi(i) \prod_{j \in A^c} (1 - \pi(j)),$$

where $A^c = U \setminus A$ is the complement of A.

Remark. The above simple analysis provides a general way to obtain various sampling designs from the specification of $\pi : U \to [0,1]$. For example, if we choose the N-copula $C(y_1, y_2, \cdots, y_N)$ to be the minimum of the $y_j \in [0,1]$, i.e.

$$C(y_1, y_2, \cdots, y_N) = \bigwedge_{j=1}^{N} y_j,$$

then

$$f(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \left[1 - \max_{j \in B^c} \pi(j)\right],$$

is a probability sampling design having $\pi(\cdot)$ as its one-point coverage function (where $|A|$ denotes $\#(A)$). See exercise. For additional reading on copulas and the problem of the joint distribution with given marginal distributions(Frechet's problem), see Nelson (1999) and Dall'Aglio (1991). □

Some aspects of statistical inference in survey sampling are given here. Let $\theta : U \to \mathbb{R}$ be a quantity of interest. The *parameter space* is the function space $\mathbb{R}^U = \{g : U \to \mathbb{R}\}$. First, under a sampling design $f : \mathcal{P}(U) \to [0,1]$, the size of the sample S is

$$\#(S) = \sum_{j=1}^{N} I_j(S),$$

so that

$$E\left(\#(S)\right) = \sum_{j=1}^{N} E(I_j(S)) = \sum_{j=1}^{N} P(j \in S) = \sum_{j=1}^{N} \pi(j).$$

This can be viewed also as a special case of *Robbin's formula* (Volume I) for counting measure (see exercise).

Consider the population total

$$\tau(\theta) = \sum_{j \in U} \theta(j) = \sum_{j=1}^{N} \theta(j).$$

The well-known unbiased *Horvitz-Thompson* estimator of $\tau(\theta)$ is

$$\hat{\tau}(S) = \sum_{j \in S} \frac{\theta(j)}{\pi(j)}.$$

Indeed,

$$\hat{\tau}(S) = \sum_{j=1}^{N} \frac{\theta(j)}{\pi(j)} I_j(S)$$

so that

$$E\left(\hat{\tau}(S)\right) = \sum_{j=1}^{N} \frac{\theta(j)}{\pi(j)} E(I_j(S)) = \sum_{j=1}^{N} \frac{\theta(j)}{\pi(j)} \pi(j) = \sum_{j=1}^{N} \theta(j) = \tau(\theta),$$

for all $\theta \in \mathbb{R}^U$.

For additional reading on the state-of-the-art of the theory of statistical inference in survey sampling, see Cassel et al (1977), Hajek (1981), Foreman (1991), Sarndal et al (1992), and Knottnerus (2003). For a classical text on *Sampling Techniques*, see e.g. Cochran (1977).

1.3 Statistical Data

The motivating example in Section 1.1 provides a typical situation in which a statistical science is needed. Statistics is a science of making inference from samples to populations. Starting with providing useful information for states (hence the name statistics), the framework and methodology of statistics spread out to almost all fields of our society. These include engineering, science, economics, medicine, agriculture, and business. This is due to the common features of these fields with respect to estimation, testing of theories and prediction, all based upon observations of parts of the whole population. In a broader sense, statistics is a part of a general theory of decision-making under uncertainty. A *statistic* is a function of the observations from phenomena or systems. The science of statistics consists of using probability theory to arrive at valid inductive logic. From this perspective, it is easy to list the applications of statistics in almost all human activities. Statistical science provides a framework and methodology for solving problems which, otherwise, should be left to fortune tellers! The need to use statistics to reach conclusions is thus apparent since, after all, we live in a world full of uncertainties, and the quest for knowledge discoveries is inherent in human nature.

In order to carry out valid inductive logic, we need data. One way that randomness enters the picture is through man-made randomization such as sample survey. Since inference cannot be absolutely certain, we need to use the *language of probability theory* to formulate results of statistical inference. Students interested in logical aspects of statistical reasoning can read, e.g., Hacking (1976).

Another situation where *sets* appear as outcomes of a natural (not man-made) random experiment is the following.

Let X be a random variable of interest, say,

$$X : (\Omega, \mathcal{A}, P) \longrightarrow (U, \mathcal{B}, P_X),$$

where the probability law P_X of X is unknown. To discover P_X, we perform repeatedly experiments on X to obtain observations $X_1, X_2, \cdots, X_n$. In the case where the observations X_i cannot be observed (directly or precisely), due to various reasons, such as precisions of measurement instruments, observations are corrupted by noise. We might need to find ways to extract some information from our experiments. A mechanism for achieving this is called a *coarsening*. As in the problem of selecting samples from a population, coarsening mechanism can be deterministic or random. Here is a coarsening example. Suppose that, while an outcome $X(\omega)$ from X cannot be observed, it can be located in one of the elements of a finite $\mathcal{B}$-partition of U, say, $\{A_1, A_2, \cdots, A_k\}$, where $A_i \in \mathcal{B}, i = 1, 2, \cdots, k$,

$$A_i \cap A_j = \varnothing \quad \text{for} \quad i \neq j, \quad \text{and} \quad \bigcup_{i=1}^{k} A_i = U.$$

Specifically, we observe the set A_i which contains $X(\omega)$. Thus, what we will observe are sets in $\{A_1, A_2, \cdots, A_k\}$. As in the random sampling of samples of fixed size, the A_i's will be obtained at random (not from a man-made random mechanism, but from the natural randomness coming from the unknown P_X). Note that a chosen partition $\{A_1, A_2, \cdots, A_k\}$ is similar to selecting samples of some given size in survey sampling. Each A_i represents the information about the value of X which falls into it. Of course, the sizes of the A_i's represent the precision of the coarsening scheme. When performing an experiment on X, the chance for observing A_i (i.e. $X \in A_i$) is precisely

$$P(X \in A_i) = P_X(A_i), \qquad i = 1, 2, \cdots, k.$$

Thus, the coarsening is in fact random. Specifically, if we let

$$S : (\Omega, \mathcal{A}, P) \longrightarrow \{A_1, A_2, \cdots, A_k\}$$

be a (finite) *random set* with probability density

$$f_S(A_i) = P(S = A_i) = P_X(A_i), \qquad i = 1, 2, \cdots, k,$$

then $P(X \in S) = 1$, i.e. X is an almost sure selector of S, or S is a *coarsening* of X. Thus, formally, a random set is a mathematical model for coarsening. The "outcomes" on X turn out to be values of S, i.e. an outcome of our experiment in this situation is a set.

In this coarsening scheme, clearly we have that

$$P(S = A_i | X = x) = 1 \qquad \text{as long as} \quad x \in A_i,$$

so that $P(S = A_i | X = x)$, as a function of x, is *constant* on A_i. It is this fact that suggests a general model for coarsening known at the *coarsening at random* (CAR models), see Chapter 2 of Volume I. If we set

$$\pi(A) = \begin{cases} 1 & \text{for} \quad A \in \{A_1, A_2, \cdots, A_k\} \\ 0 & \text{for other} \quad A \in \mathcal{B}, \end{cases}$$

then

$$f_S(A) = \pi(A) P_X(A) \qquad \text{for any} \quad A \in \mathcal{B}.$$

Here we set $f_S(A) = 0$ when $A \notin \{A_1, A_2, \cdots, A_k\}$. In particular, when U is finite ($\mathcal{B} = \mathcal{P}(U)$, the power set of U), we have that

$$\sum_{A \ni x} \pi(A) = 1 \qquad \text{for each} \quad x \in U.$$