Vietnam Journal of Mathematics **37**:2&3(2009) 315-337

Vietnam Journal
of
MATHEMATICS

A Survey of Some Current Statistical Research Problems

Nguyen Trung Hung

Department of Mathematical Sciences
New Mexico State University, Las Cruces, NM 88003, USA

Received January 21, 2009
Revised May 14, 2009

Abstract. The goal of this survey paper is to point out where mathematical research problems in statistics come from! As statistics is a science of uncertainty analysis, it is a methodology and a tool for all real world problems in which uncertainty is a fact of life. These include both physical and social domains.

In a general sense, statistics is used for prediction purposes, either through time series models or regression models. While a standard theory of statistics is available, such as Bayesian statistics, U-statistics and logistic regression, research arises when we embark on a specific domain of applications. In each domain of applications, e.g. biology, financial economics, mathematical finance, control engineering,... at least two distinguished features show up, namely the structured equations due to the theory itself, and more importantly, the type of data available. The important field of econometrics is a typical example in which structured equations come from economic principles, and data (time series, cross section or panel data) are coarse in various possible ways, such as missing, censored, hidden, partially observed...

We will discuss research problems in theoretical statistics which arise mainly *from various types of coarse data*, especially in economics. These include statistical decision theory (based upon Von Neumann's utility theory) in the context of financial investments and portfolio selection via the theory of stochastic dominance and risk management, qualitative choice problems with latent dependent variables (generalized linear models, probit, logit and tobit regression models), James Heckman's sample selection models, hidden Markov models in biology, indirect observed data in auction theory (via Nash's equilibrium concepts in games with incomplete information). The exposition is *tutorial* in nature.

2000 Mathematics Subject Classification: 62-01, 62G05

Key words: Censored, hidden Markov, measurement-error, missing, random set, unobserved data.

1. Measurement-error Data

1.1. Random Samples and Statistics

Let start out by fixing some basic terminologies and notation. A population is denoted by a random variable (or more generally a random vector) X with distribution function F. A random sample, of size n, drawn from X is a collection of random variables $X_1, X_2, ..., X_n$ which are mutually independent and having the same distribution as X, in notation I.I.D. In all generalities, we are interested in discovering F. We have at our disposal a random sample $X_1, X_2, ..., X_n$ and wish to use information from it to estimate F . The information in the random sample is expressed as various functions of it, say, $T_n(X_1, X_2, ..., X_n)$, called statistics. Here, we have a *complete sample*, in the sense we can observe all the values of the X_i and with accuracy. Before collecting the values of $X_1, X_2, ..., X_n$, these are random variables, and so are the statistics T_n.

Since there are many possible ways to, say, estimate F, we wish to find "good" or even the "best" *estimators*. Techniques for finding good estimators depend on how much we know about F itself. We can classify the prior knowledge about F into two categories:

(i) *Nonparametric models:* We only know that the "true" F belongs to some class of distributions, such as F has a density f (i.e. F is absolutely continuous).

(ii) *Parametric models:* We know more: We known the form of F, up to some (finitely dimensional) *parameters*, say θ, belonging to some known set Θ , called the parameter space.

Recently, especially in econometrics, we run into a combination of the two above models, namely *semi-parametric models* in which F depends on both a finitely dimensional parameter and an infinitely dimensional parameter, i.e. to identify F, we need to estimate its two different types of components, one is parametric and the other is nonparametric.

1.2. Regression

One of the main objective of statistical analysis is to provide "educated" ways for making *predictions*.

There are two ways to do so:

(i) *Time-series models:* This is the case where we know nothing about the causality that affects our variable of interest Y (the one we try to forecast). However, we can observe its past behavior (over time), say, $Y(t_1), Y(t_2), ..., Y(t_n)$. In this case, prediction of future $Y(t)$ is a problem of *extrapolation*, i.e. propose models which relate $X(t)$ to its past observations, from which predictions could be made.

(ii) *Regression models:* We can find out that the variable to be predicted Y (response variable) is explained by a number of other variables X_1, $X_2, ..., X_k$ (explanatory variables or covariates). Taking this and other relevant factors, an adequate representation of this relationship could take a form such as

$$Y = \varphi(X_1, X_2, ..., X_k) + \varepsilon,$$

where ε is a random variable representing the error. In particular, when appropriate, φ is taken as a linear function of the $X_1, X_2, ..., X_k$ leading to a linear regression model:

$$Y = \sum_{j=1}^{k} \beta_j X_j + \varepsilon.$$

Given the observed data $(X_i, Y_i), i = 1, 2, ..., n$,

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ji} + \varepsilon$$

the β_j can be estimated *consistently* by the well-known method of least squares.

It is convenient to write the above linear model in vector form. Let β denote the column vector with components β_j, and $\mathbf{X'}$ denote the tranpose (row vector) of the X_j,

$$\mathbf{Y}_i = \mathbf{X}_i'\beta + \varepsilon_i.$$

Under suitable conditions, the least squares estimator of β is $(\mathbf{X'X})^{-1}\mathbf{X'Y}$. At least, for an estimator β_n to be reasonable, it has to be consistent, i.e. as we get more data ($n \to \infty$), β_n should get closer and closer to β. More specifically, $\lim_{n\to\infty} \beta_n = \beta$ in some probabilistic senses, such as convergence in probability or convergence almost surely.

1.3. Measurement-error Data in Linear Models

On the list of typical and important coarse data, the classical and well-studied type is the measurement-error data. This is referred to the fact that, in statistical studies, it might happen that some of variables involved cannot be observed or measured with accuracy. Instead, they can be only estimated with random errors.

Consider the classical normal linear model

$$Y = \beta X + \varepsilon. \tag{1}$$

If the data, say, (X_i, Y_i), $i = 1, 2, ..., n$ are available, then β can be estimated consistently by least squares. Now suppose that the observed data X_i are imperfectly measured, say, as

$$Z = X + u, \tag{2}$$

where u is $N(0, \sigma_u^2)$. Let see what happens if we simply use the approximate values of Z_i to estimate β. Insert (2) into (1), we get

$$Y_i = \beta(Z_i - u_i) + \varepsilon_i = \beta Z_i + w_i, \tag{3}$$

where the new error term is $w = \varepsilon - \beta u$. Under standard assumption of uncorrelatedness of all errors (Z, ε, u are mutually independent),

$$\mathrm{Cov}(Z, w) = \mathrm{Cov}(X + u, \varepsilon - \beta u) = -\beta\sigma_u^2 \neq 0,$$

i.e. w is correlated with Z, violating a basic assumption in standard linear model, and as such the least squares estimated of β from (3), namely

$$\beta_n = \left[\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right]\left[\frac{1}{n}\sum_{i=1}^{n} Z_i^2\right]^{-1}$$

is not consistent. In fact, β_n converges in probability to $\beta/[1+\sigma_u^2/a] \neq \beta$, where $a = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Z_i^2$, in probability.

1.4. Statistical Analysis and Research Issues

Statistics with measurement-error data is well documented mainly in regression analysis, for both linear and nonlinear models (see e.g. [3, 8, 10]). General statistics with this type of imprecise data remains to be investigated, for example, how to consistently estimate probability distributions (or densities) of random vectors from a random sample of measurement-error data?

2. Censored Data

2.1. Data in Survival Analysis

Survival analysis is an area concerning lifetimes of subjects (humans, electronic components, etc...). Thus the unknown distribution function F of the variable of interest X (lifetime) or its hazard rate $\lambda(t) = f(t)/[1 - F(t)]$, where $f(.)$ is the derivative of F, is to be estimated from data. Now if the data is a *complete* random sample $X_1, X_2, ..., X_n$, then F can be estimated consistently, pointwise, by the *empirical distribution function*

$$F_n(x) = (1/n)\sum_{i=1}^{n} I(X_i \leq x).$$

However, in medical statistics, we do not have, in general, complete data. Specifically some of the X_i are observed, while the others are only *partially observed*. This is due to the nature of lifetimes! For example, with n subjects in a study, we need to conduct our analysis, say at a fixed time T. Thus the observed X_i are those which are smaller than T, whereas the others are only known to be greater than T. Specifically, we observe the indicator function $I(X_i \leq T)$ and the X_i for which the indicator function is 1. Thus, the available data is subject

to a *censoring mechanism,* and the so obtained data is referred to as *censored data.* More generally, a random censoring mechanism is this.

Let $Y_1, Y_2, ..., Y_n$ be I.I.D. as the censoring times (Y). What we are able to observe is the values of

$$Z_i = X_i \wedge Y_i$$

and

$$\delta_i = I(X_i \leq Y_i).$$

Without having the values of $X_1, X_2, ..., X_n$, we cannot even carry out the so-called "Fundamental Theorem of Statistics" (for statistical inference), namely estimating the distribution function F of X, by above usual procedure. Thus, the problem is how to estimate F from the observed data $(Z_i, \delta_i, i = 1, 2, ..., n)$?

For each $G \in \mathcal{F}$ (a class of distributions), the likelihood of getting the observed data under G is

$$L(G|Z_i, \delta_i, i = 1, 2, ..., n) = \prod_{i=1}^{n} p^{\delta_{(i)}} \left(\sum_{j=i+1}^{n+1} p_j \right)^{1-\delta_{(i)}},$$

where $p_i = dG(\{x_{(i)}\})$, $p_{n+1} = 1 - G(x_{(n)})$, with $x_{(i)}$ denoting order statistics.

Maximizing this likelihood over $\mathcal{F}$, subject to $p_i \geq 0, \sum_{i=1}^{n+1} p_i = 1$, we obtain its MLE

$$F_n^*(x) = 1 - \prod_{Z(i) \leq t} (1 - \delta_{(i)}/[n - i + 1])$$

the famous *Kaplan -Meier product-limit estimator* (1958).

2.2. Censored Data in Linear Models

In many economic data, the censored data are in fact missing data of some special type. We illustrate this situation in regression models in which data on the response variables are censored in some specific way and cause selection bias. A popular linear regression model with censored data is the *Tobit model* (Tobin, 1958). The kind of censored data in Tobit model is a special case of sample selection models of James Heckman which we will discuss later in the context of missing data. The treatment of Tobit model will illustrate how we should modify existing statistical procedures to cope with new types of data. Consider a standard linear model

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

In contrast to measurement-error data in which the explanatory variable X is coarsed in the sense that its values are obtained with errors, we consider now the case where X_i are all observed with accuracy, for all individual i, in a sample of size N, but the response variable Y is partially latent (unobserved) in that some of the values Y_i are missing. The basic question is: can we just use the available (observed) data (i.e. ignoring the missing ones) to estimate parameters by ordinary least squares (OLS) method? In fact, the precise question is: Are

OLS estimators of the parameters, based only on the subsample of observed data, *consistent*? This clearly depends on "how data are missing?". We consider a common situation in economics where missing data are in fact censored of a special type. In other words, data are not missing at random, but rather in a "self selected" manner. In such cases, the observed subsample is biased, and hence standard statistical procedures in linear models will lead to inconsistent estimators of model parameters. In order to use only the observed subsample to obtain consistent estimators of parameters, we need to find ways to "correct the sample bias", and that will lead to new statistical procedures.

Now, since we are in a linear model setting, with, say, standard assumptions about the random error term ε, we would like to use the ordinary least squares (OLS) method to estimate the parameters. It is well known that with full sample data, OLS estimators are consistent (since $E(\varepsilon_i) = 0$), a property of estimators that we consider as "good". For example, suppose the least expensive car costs c. For a houshold i whose desired level of car expenditure is less than c will not be able to buy a new car, and hence Y_i is not available. In such a situation, we know the source of missing data. The observed data in the sample is obtained by the *sample selection rule*: Y_i is selected (observed) from the sample if $Y_i \geq c$. In terms of censoring, the data Y_i are censored by itself, i.e. the value of Y is observed only when some criterion in term of Y itself is met. Specifically, in the above model, the Y_i are observed when $Y_i \geq c$. Suppose we just consider the model with the data available in the observed subsample (Y_i, X_i), say, $i = 1, 2, ..., n < N$. Doing that, we are in fact considering the variable Y *truncated* by the sample selection rule $Y \geq c$. In other words, using only the observed subsample is using a sample from the truncated Y.

We have

$$E(Y_i|X_i, Y_i \geq c) = \alpha + \beta X_i + E(\varepsilon_i|X_i, Y_i \geq c).$$

Now

$$E(\varepsilon_i|X_i, Y_i \geq c) = E(\varepsilon_i|X_i, \varepsilon_i \geq c - (\alpha + \beta X_i)).$$

The error term $\varepsilon_i|X_i, \varepsilon_i \geq c - (\alpha + \beta X_i)$ has a non-zero mean. Indeed, since $\varepsilon_i \geq c - (\alpha + \beta X_i)$, we see that, for any particular value of X_i, its mean can be negative, positive or zero. As a result, OLS estimators will be inconsistent.

In view of the above, we ask "Is it possible to estimate consistently the model parameters using only the observed subsample?", because, after all, that's all available data we got! Since OLS fails in this linear model with censored data, let turn to another estimation method! Tobin suggested to estimate the parameters by MLE by assuming normal distribution $N(0, \sigma^2)$ for the error ε_i', where the likelihood is computed as follows

$$L_N(\alpha, \beta|(X_i, Y_i), i = 1, 2, ..., N) = L_n L_*,$$

where L_n denotes the part with observed Y_i , and L_* the part correspongding to unobserved Y_i. Specifically, denoting by φ and Φ the density and distribution of the standard normal random variable, respectively, the density of the uncensored

observations Y_i is

$$f(y_i) = (1/\sigma)\varphi[(y_i - (\alpha + \beta x_i)/\sigma]$$

and, for censored Y_i , L_* is a product of

$$P(Y_i < c|X_i) = P(\varepsilon_i < [c - (\alpha + \beta X_i)]/\sigma) = \Phi([c - (\alpha + \beta X_i)]/\sigma).$$

First, observe that $E(\varepsilon_i|X_i, Y_i \geq c) \neq 0$ is the source of inconsistency. So let remove it!

Let $E(\varepsilon_i|X_i, Y_i \geq c) = \zeta_i$ (an unknown non-random term, unknown since it involves α, β), and $u_i = \varepsilon_i - \zeta_i$. We have clearly

$$E(u_i|X_i, Y_i \geq c) = 0.$$

So consider

$$Y_i = \alpha + \beta X_i + \zeta_i + u_i. \tag{4}$$

Suppose we view ζ_i as a second regressor for Y_i, let examine the correlation between the random error term u_i and the regressors X_i and ζ_i.

We have

$$\begin{aligned} E(X_i u_i|X_i, Y_i \geq c) &= EE(X_i u_i|X_i, Y_i \geq c)|X_i, Y_i \geq c) \\ &= E(X_i E(u_i|X_i, Y_i \geq c)) \\ &= 0 \end{aligned}$$

also

$$E(\zeta_i u_i|X_i, Y_i \geq c) = 0.$$

Thus, u_i is uncorrelated with the regressors X_i and ζ_i.

Just like in measurement-error models, ordinary least squares estimators of parameters are biased and inconsistent since $E(\varepsilon_i) \neq 0$. Indeed, since $Y_i \geq 0$, it follows that $\varepsilon_i \geq -(\alpha + \beta X_i)$ and hence for any particular value of X_i, $E(\varepsilon_i)$ can be negative, positive or zero. Let turn to maximum likelihood estimation!

Suppose ε_i is distributed as $N(0, 1)$ and let φ, Φ denote its density and distribution functions, respectively.

Since $\varepsilon_i \geq -(\alpha + \beta X_i)$, ε_i^* is a truncated random variable of ε_i, i.e. the $\varepsilon_i^* = \varepsilon_i|\varepsilon_i \geq -(\alpha + \beta X_i)$ so that its density is

$$f(\varepsilon_i^*) = \varphi(\varepsilon_i^*)/P(\varepsilon_i \geq -(\alpha + \beta X_i)) = \varphi(\varepsilon_i^*)/[1 - \Phi(-(\alpha + \beta X_i)].$$

Thus

$$E(\varepsilon_i|\varepsilon_i \geq -(\alpha + \beta X_i)) = \sigma\varphi(\alpha + \beta X_i)/\Phi(\alpha + \beta X_i),$$

where σ is the standard deviation of ε_i^*. The quantity $\lambda_i = \varphi(\alpha + \beta X_i)/\Phi(\alpha + \beta X_i)$ is called the inverse Mill ratio. If we have estimates of the λ_i, then we can use them to normalize $E(\varepsilon_i)$ to zero (considering the error $\varepsilon_i - \sigma\lambda_i$) and

hence obtaining consistent estimators for the parameters. The famous two-stage estimation procedure of James Heckman [11] is this.

(i) We estimate the λ_i by estimating α, β and plug in $\lambda_i = \varphi(\alpha+\beta X_i)/\Phi(\alpha+\beta X_i)$.

Consider the probit model with $Z_i = 1$ if $Y_i^* > 0$ and $= 0$ if $Y_i^* \leq 0$

$$p_i = \Phi(\alpha + \beta X_i).$$

The parameters α, β are estimated by MLE, using the whole sample.

(ii) Since the estimates $\hat{\lambda_i}$ of λ_i converge to λ_i as $n \to \infty$, we consider

$$Y_i = \alpha + \beta X_i + \sigma\hat{\lambda_i} + \varepsilon_i - \sigma\hat{\lambda_i} = \alpha + \beta X_i + \sigma\hat{\lambda_i} + u_i.$$

This is a linear model with two regressors X and $\hat{\lambda}$ with the error term u_i having

$$E(u_i|\varepsilon_i \geq -(\alpha + \beta X_i)) = 0.$$

The final estimates of α, β are now obtained by OLS method in this linear model using only the observed data on Y_i.

We pause now to illustrate a research problem in survival analysis where work on censored data remains to be done.

2.3. An Example of a Research Problem

We present now a research problem in survival analysis to suggest research problems with censored data. This is the so-called *change-point problems.*

The lifetime X of subjects is exponential with a change-point: it is exponentially distributed with rate α up to an unknown time point τ, and then exponentially distributed with another rate $\beta \neq \alpha$, from that time on. The probability density function $f(x;\theta)$ of X, where the population parameter $\theta = (\alpha, \beta, \tau) \in (0, \infty)^3$, is

$$f(x;\theta) = \alpha e^{-\alpha x} I(0 \leq x \leq \tau) + \beta e^{-\alpha\tau - \beta(x-\tau)} I(x > \tau),\ x \in \mathbb{R}.$$

Given a conventional (i.e. non censoring) random sample $X_1, X_2, ..., X_n$ drawn from X, we wish to estimate θ, at least, consistently. The problem is in the component τ (the change-point) which is in the support of the distribution. Suppose we intend to use maximum likelihood method for estimation θ.

The log-likelihood is

$$\log L_n = \sum_{i=1}^{n} I(0 \leq X_i \leq \tau) \log\alpha - \alpha \sum_{i=1}^{n} X_i I(0 \leq X_i \leq \tau) + \sum_{i=1}^{n} [1 - I(0 \leq X_i \leq \tau)][\log\beta - (\alpha - \beta)\tau] - \beta \sum_{i=1}^{n} X_i [1 - I(0 \leq X_i \leq \tau)]$$

or, if we let $W(\tau)$ denote the (random) number of $X_i \leq \tau$, then in terms of W and the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$, $\log L_n$ is written as

$$W \log \alpha - \alpha \sum_{i=1}^{W} X_{(i)} + (n - W)[\log \beta - (\alpha - \beta)\tau] - \beta \sum_{i=W+1}^{n} X_{(i)}/n!.$$

Now, for $X_{(n-1)} \leq \tau < X_{(n)}$, the right hand side of the above equation is effectively

$$(n-1) \log \alpha - \alpha \sum_{i=1}^{n-1} X_{(i)} - \alpha\tau + \log \beta - \beta(X_{(n)} - \tau).$$

Thus, for $\beta^* = 1/([X_{(n)} - \tau^*] > 0$ and let τ^* get close to $X_{(n)}$, $\log L_n$ will tend to infinity.

Thus, in this survival model, the likelihood is unbounded and hence MLE method cannot be used. In view of the failure of the popular MLE method, how to estimate the change-point τ of this *non-regular model*, consistently and also derive its large sample sampling distribution? Note that, once τ is estimated as such, the other parameters α, β are estimated consistently by the plug-in method (for given τ, $\alpha(\tau)$ and $\beta(\tau)$ are obtained by maximizing the likelihood in which τ is replaced by its estimator). Another situation is the 3-parameter Weibull density

$$(\beta/\alpha)[(t-\tau)/\alpha]^{\beta-1} \exp[-((t-\tau)/\alpha)^{\beta}]I(t \geq \tau)$$

for which the likelihood is unbounded. The density

$$f(x;\theta) = \alpha e^{-\alpha x} I(0 \leq x \leq \tau) + \beta e^{-\alpha\tau - \beta(x-\tau)} I(x > \tau) \ , x \in \mathbb{R}$$

can be written as a mixture of two densities but with *unknown weights* (depending upon the unknown parameters). Indeed, let

$$f_1(t) = \alpha \exp(-\alpha t) I(0 \leq t \leq \tau)/[1 - \exp(-\alpha t)]$$

and

$$f_2(t) = \beta \exp[-\beta(t-\tau)] I(t > \tau).$$

Then f is a mixture of f_1 and f_2 with weights $1 - \exp(-\alpha\tau)$ and $\exp(-\alpha\tau)$.

Solutions to the problem of estimating the change-point τ:

(i) *A strong consistent estimator based upon stochastic processes*

After "discovering" that the likelihood function is unbounded, Nguyen, Rogers and Walker [17] took a closer look at the problem "How to estimate the change-point τ?" The answer is in the construction of a consistent estimator of τ using a geometric approach.

(ii) *A modified form of MLE*

A simpler consistent estimator of τ is obtained (by Pham and Nguyen, [19]) by using Maximum Likelihood Method but restricted to a sequence on random sets (to avoid unboundedness) in such a way that, as in U. Grenander's methods of sieves, as the sample size increases, these random sets (as random ranges of the likelihood) spread out to the whole $\mathbb{R}^+$.

(iii) *Boostrapping*

The above work in fact covers a much larger class of statistical models, namely *non-regular models* (where MLE cannot be directly applied), and more general, it's about the problem of estimating the location of a discontinuity in a density function (studied in Chernoff and Rubin, 1956). In 1993, Pham and Nguyen [20] obtained the sampling distribution for our restricted MLE of τ using the *bootstrap method*. Research issues: How to modify the above statistical procedures when data are censored?

3. Missing Data

Missing data occur frequently in studies involving humans. Not accounting for missing data properly could lead to incorrect inferences.

3.1. The Nature of Missing Data

The most common occurrence of missing data is in survey data, in which some respondents fail to answer questions. As James Heckman [11] has pointed out to us, when facing missing data, we need to ask the basic question: Why data are missing? The answer to this question is a must before we use the available data (i.e. the non-missing observations) for statistical inference. This is so because, as we will see, the source of missing data will dictate how we should design our statistical procedures appropriately to yield meaningful results. Failure for recognizing this issue could lead to erroneous statistical conclusions from our analysis. Basically, there are two cases:

(i) Some are simply unavailable for reasons unknown to the analyst, so we can just "ignore it" (Griliches, 1986). The observed subsample is an usable data set. Perhaps we might wonder whether we can extract some useful information from the "incomplete data".

(ii) The missing data are systematically related to the phenomenon being modeled. For example, in surveys when the data are "self-selected" or "self-reported" (such as high-income individuals tend to withold information about their income). In this case, the gaps in the data set would represent more than just missing information, and hence the observed subsample would be qualitatively different. Let elaborate a little bit on this case. Recall the censored data in Tobit model. We know that response variable Y_i, for individual i, is missing when it is below some threshold c. Thus, the observed data are from a truncated distribution of Y. Two things come out of this: in one hand, we can describe the "sample selection rule", namely, the sample "selects" the data in the sample

according to a rule to reveal to us the observed ones, and on the other hand, this sample selection rule provides us with the correct distributional information to analyze our observed data, leading to correct inferences.

This situation is general. Figuring out the source of the missing data, we will describe it as a sample selection rule. Once we have this rule, we proceed as in the Tobit special type of missing data. As we will see in Heckman's sample selection models, this will lead to two equations to represent the situation:

(i) *An outcome equation*, which is our main equation of interest, describing the relation between our response variable and its determinants (covariates),

(ii) *A selection equation*, which describes the sample selection rule, leading to the observed data.

Missing data occur frequently in statistical analysis of economic data for forecasting. We will elaborate on it in the context of regression. But since we will introduce Heckman sample selection models and his two-step estimation procedure in linear regression models, we need to review some special *generalized linear models*, namely, the *probit* and *logit models*.

3.2. Regression for Discrete Response Variables

Consider the so-called "linear probability model", or models for qualitative choices. This is the case where our response variable Y is *discrete* , but the regressor X could be continuous. For example, consider the labor force participation problem, in which $Y_i = 1$ or 0 according to the respondent either works or does not work. We believe that the decision to work or not can be explained by a set of covariates such as age, martial status, education,... gathered in a vector X, i.e. we can write

$$P(Y = 1|X) = \Psi(X, \beta) = 1 - P(Y = 0|X),$$

where Ψ is some function, and β is a vector of parameters . A simplest model (i.e. specifying the function Ψ) is the linear regression model, where we take Ψ as a linear function in X, i.e.

$$\Psi(X, \beta) = X'\beta.$$

In this model, the mean of Y is predicted by the linear predictor $X'\beta$, since $E(Y|X) = P(Y = 1|X) = X'\beta$.

From that, we can write

$$Y = E(Y|X) + Y - E(Y|X) = X'\beta + \varepsilon.$$

However this linear model for Y has a number of shortcomings. First, the variance of ε will depend on β (ε is heteroscedastic). Indeed, as $\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y)$ and since Y is Bernouilli, we have

$$Var(\varepsilon|X) = X'\beta(1 - X'\beta).$$

More importantly, recalling that $E(Y|X) = P(Y = 1|X) = X'\beta$, there is no guarantee that $X'\beta \in [0,1]$ for all possible values of X, since we cannot constraint $X'\beta$ to the interval $[0,1]$. This failure could lead to both nonsense probabilities and negative variance!

Thus we should replace the linear probability model by another model which makes sense! One way is to "scale" the value $X'\beta$ so that it will be always in $[0,1]$, in a suitable fashion. This can be achieved by using an appropriate function F, like a distribution function of a random variable, and write

$$P(Y = 1|X) = F(X'\beta).$$

Each choice of F leads to a (non linear) model. For example, if we choose $F(x)$ to be the distribution function of the standard normal random variable, then we get the so-called *probit model*. How to estimate β from our observed data $(X_i, Y_i), i = 1, 2, ..., n$?

Well, conditional on X_i, Y_i is Bernouilli with parameter $p_i = F(X_i'\beta)$, and hence the likelihood of β given the observations is

$$\prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n}(F(X_i'\beta))^{Y_i}(1-F(X_i'\beta))^{1-Y_i}$$

so that, we are using maximum likelihood method for estimation. Now the use of F as the normal distribution function in the above probit model presents difficulties in computation since there is no closed form for F^{-1}. Another much simple F which is quite "close" to the normal distribution function is the *logistic* distribution function, namely

$$F(x) = 1/(1 + e^{-x})$$

for which $F^{-1}(t) = \log(t/(1-t))$. Using the logistic F, the associated model is called the *logit model*, and we are using *logistic regression models.*

There are special features of economic data that distinguish econometrics from other branches of statistics. Economic data are generally observational, rather than being derived from controlled experiments. Because the individual units in an economy interact with each other, the observed data tend to reflect complex econometric equilibrium conditions rather than simple behavioral relationships based on preferences or technology. Note that, early work in econometrics focused on time-series data, but now econometrics also fully covers cross-sectional and panel data (see e.g. [7]).

In statistics, *logistic regression* is a model used for prediction of the probability of occurrence of an event. It is used extensively in the medical and social sciences as well as marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription. In *neural networks,* the logistic function is called the *sigmoid function.* A random variable is said to be *logistically distributed* if its distribution function is of the form

$$F(x;\mu,s) = [1 + e^{-(x-\mu)/s}]^{-1}$$

for the location parameter $\mu \in \mathbb{R}$ and the scale parameter $s > 0$.

Its density is

$$f(x;\mu,s) = e^{-(x-\mu)/s}[s(1 + e^{-(x-\mu)/s})^2]^{-1}.$$

The standard logistic distribution (also called the sigmoid function, a name due to its sigmoid shape of its graph) is

$$f(x) = 1/[1 + e^{-x}].$$

In the context of decision theory, based upon Von Neumann's utility theory, it was D. Luce who seems to first propose the *logistic discrete choice model* in his monograph *Individual Choice Behavior*, J. Wiley, 1959. He offered a theoretical *justification* based upon "independence of irrelevant alternatives". As cited by D. Luce and P. Suppes in 1965, Marley showed that if the random error term on utilities are i.i.d. with Gumbel distribution, i.e. with CDF $\exp(e^{-x})$, then probabilities of selecting alternatives follow the logit model. In 1974, Daniel McFadden showed the converse: If the errors are i.i.d. and the choice probabilities are described by the logit model, then necessarily these errors are distributed according to the Gumbel distribution.

General linear models are linear models in their most general setting, whereas generalized linear models are extensions of linear models, and thus contain non-linear models as well. Probit and Logit models are examples of non linear models which are generalized linear models. The set-up of generalized linear models is this.

(i) The random component: the variables Y_i are independent responses that follow a distribution belonging to the exponential family, with mean $E(Y_i) = \mu_i$.

(ii) The linear predictor $X'\beta$, from covariates X, is used.

(iii) There is a link function Ψ which relates the linear predictor to the mean response, i.e. $X_i'\beta = \Psi(\mu_i)$.

3.3. Sample Selection Models

As stated before, we are concerned with statistical inference, especially in regression models, with missing data where we can describe the missing data mechanism. Facing missing observations in our survey data, we seek to discover "why our data are missing?". A full understanding of the situation under study will help in answering this question! With that qualitative information, we specify the way data are missing. Each such description of the sample selection rule gives rise to a *model*, and hence the term "sample selection models". More general than Tobit censored regression model, in Heckman's sample selection models, the response variable is observed only when some criterion based on some other random variable is met. Note that we could classify samples (in regression with response variable Y , and regressor X) as follows:

(i) *Censored samples*: Y is observed only if some criterion defined in terms of Y is met, X is observable in the whole sample (regardless of whether Y is observed or not),

(ii) *Truncated samples:* Y is observed if some criterion defined in terms of Y is met, X is observed only if Y is observed,

(iii) *Selected samples:* Y is observed only if some criterion defined in terms of some other random variable (Z) is met, X and another variable W (which forms the determinants of whether $Z = 1$) are observable for the whole sample (regardless of whether Y is observed or not).

Let describe now the setting of *linear regression in a model of selected samples.* To understand the set-up we are going to put down, it is helpful to have a simple example. Consider the classic example of female labor supply. We wish to study the determinants of market wages of female workers. The market wage level for each female worker (Y_i) could depend on observable characteristics (determinants) of the worker (X_i) such as age, education, experience, martial status,...). Each woman i sets a reservartion wage level to accept to work, so let Z_i denote the difference between the market wage offered to i and her reservation wage. when $Z_i > 0$, the woman i is employed and her wage is observed. Suppose linear models are appropriate for the study!

The model of interest is:

$$Y_i = \alpha X_i + \varepsilon_i. \tag{5}$$

Suppose that Z can also be regressed linearly by some covariate W, i.e.

$$Z_i = \beta W_i + u_i. \tag{6}$$

This is the *selection equation* of the model. Again, in fact Z is a latent variable (representing the "propensity" to work). The *sample selection rule* is this: Y_i is observed only when $Z_i > 0$. But Z_i is not observable, what is observed is whether it is positive or negative, i.e. only the sign of Z_i is observed. If we let the dummy variable δ_i taking value 1 or 0 according to Z_i is positive or negative, then, given a sample of n individuals, we actually "observe" $(X_i, W_i, \delta_i), i = 1, 2, ..., n$, whereas we observe Y_i only for i for which $\delta_i = 1$.

3.4. Statistical Analysis and Research Problems

Using only the subsample of available data (i.e. only for the individuals i for which Y_i is observed, equivalently, for which $\delta_i = 1$, by sample selection rule) to estimate the primary parameter α in equation (5) will result in biased estimates (i.e. the OLS fails like in Tobit model), since $E(\varepsilon_i|\text{sample selection rule}) \neq 0$.

Indeed,

$$E(\varepsilon_i|X_i, W_i, \delta_i = 1) = \sigma\lambda(\beta W_i) \neq 0,$$

where $\sigma = \text{Cov}(\varepsilon_i, u_i)$, assuming that (ε_i, u_i) is bivariate normal with $\text{Var}(\varepsilon_i) = \text{Var}(u_i) = 1$ (for simplicity), and $\lambda(t) = \varphi(t)/[1 - \Phi(t)]$ which is the hazard rate of the standard normal random variable, which is also called the reciprocal of the

Mills ratio. We note as before by φ and Φ the density function and distribution function of the standard normal random variable. Now, with the specification of error distribution, in principle, the parameters α and β can be estimated at the same time by maximum likelihood method which is complicated! James Heckman proposed a *two-step estimation procedure* to avoid complications of maximum likelihood method. There are two steps since his method is sequential: Estimate β first, then use it to estimate α.

Step 1. Estimate β by probit analysis, applying to the selection equation alone.

We have

$$Z_i = \beta W i + u_i,$$

where u_i is $N(0,1)$, and $\delta_i = 1$ or 0 according to $Z_i > 0$ or $Z_i \leq 0$.

We then have

$$\begin{aligned}\pi_i = P(\delta_i = 1) = P(Z_i > 0) = P(\beta W_i + u_i > 0)\\ = P(u_i > -\beta W_i) = 1 - \Phi(-\beta W_i) = \Phi(\beta W_i).\end{aligned}$$

Thus, we are in the setting of probit analysis. The likelihood of β given the observations (W_i, δ_i), $i = 1, 2, ..., n$ (i.e. the whole sample) is

$$\begin{aligned}L_n(\beta|(W_i,\delta_i), i = 1,2,...,n) &= \prod_{i=1}^{n} \pi^{\delta_i}(1-\pi)^{1-\delta_i}\\ &= \prod_{i=1}^{n} (\Phi(\beta W_i))^{\delta_i}(1-\Phi(\beta W_i))^{1-\delta_i}.\end{aligned}$$

Maximizing this marginal likelihood, we obtain the estimator $\hat{\beta}$ of β.

Step 2. Estimate α as follows.

Since $E(\varepsilon_i|X_i, W_i, \delta_i = 1) = \sigma\lambda(\beta W_i) = \sigma\lambda_i$ where we set $\lambda_i = \lambda(\beta W_i)$, we could normalize the error under the sample selection rule, i.e. $(\varepsilon_i|X_i, W_i, \delta_i = 1)$ to have zero mean by changing it to $u_i - \sigma\lambda_i$. For that, we rewrite (5) as

$$Y_i = \alpha X_i + \sigma\lambda_i + (u_i - \sigma\lambda_i).$$

Now if we estimate λ_i by the plug-in estimator $\hat{\lambda_i} = \lambda(\hat{\beta} W_i)$, for all $i = 1, 2, ..., n$, and consider

$$Y_i = \alpha X_i + \sigma\hat{\lambda_i} + (u_i - \sigma\hat{\lambda_i}) \tag{7}$$

then, asymptotically, the error $(u_i - \sigma\hat{\lambda_i})$ under the sample selection rule has zero mean and uncorrelated with X_i and $\hat{\lambda_i}$. Thus, we could consider (7) as a linear model with two regressors, X_i and $\hat{\lambda_i}$. Applying only the observed data to (7) and using OLS estimation procedure, we get a consistent estimator for α.

4. Hiddent Markov Data

4.1. Markov Chains

In many physical domains, the observed data are not necessarily independent. A type of dependence which is not "too far" from independence is "conditional independence". A sequence of random variables $(X_n, n \geq 0)$, where each X_n takes values in a common discrete (state) space S, is called a (discrete-time) *Markov chain* when it satisfies the following *Markov property:*

For any $i_o,\ i_{n-1}, i, j \in S$, and $n \geq 0$,

$$P(X_{n+1} = j | X_0 = i_o, ..., X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i).$$

Markov chains are *stochastic processes*. Its random evolution is characterized by the transition probabilities $P(X_{n+1} = j | X_n = i) = P_{ij}$. The main purpose of statistical inference is to estimate these transition probabilities from an observed realization of the chain. Statistics with Markov data has been investigated to a matured degree. What we are going to present is the situation where the data come from a Markov chain observed in noise, i.e. they are *hidden Markov data.* This presents a different type of imprecise data. *Statistics with hidden Markov data* is currently a "hot" topic in *bioinformatics* (see e.g. [2, 6]).

4.2. Hidden Markov Models

For an illustration, consider a simple situation. Let $(X_n, n \geq 0)$ be a discrete-time stochastic process with state space S (finite). The objective of studies is to answer questions about this process. If the X_n are observable, then the problem is simply statistical inference with Markov chains where a research literature exists. Suppose we cannot directly observe the X_n , and instead, we can observe another process Y_n which is "linked" to X_n in some probabilistic way. How to carry out the studies about the process X_n in such a situation? Motivated by speech recognition and biology, let consider the following.

(i) The process $(X_n, n \geq 0)$ is a stationary Markov chain, so that it is characterized by an initial distribution π on S, and a stationary one- step transition matrix $\mathbb{P} = [P_{ij}]$.

(ii) Each X_n, when in a state, emits symbols according to some probability law. Let Y_n be the emitted symbol of X_n. Then $(Y_n, n \geq 0)$ is the observed process, with state space A, called an alphabet (assuming, for simplicity, finite).

(iii) Conditional upon the realizations of X_n (unobservable), the distributions of Y_n are completely specified (i.e. the distribution of Y_n depends only on X_n). Specifically, each distribution $P(Y_n = a | X_n = i)$, $a \in A$, is known, for each $i \in S$.

Example 4.1. Let $S = \{1, 2\}$ and $A = \{a, b\}$. $\pi(1) = \pi(2) = 1/2$.

$$P = \begin{matrix} .7 & .3 \\ .6 & .4 \end{matrix}$$

$$P(Y_n = a|X_n = 1) = 1/4 = 1 - P(Y_n = b|X_n = 1),$$
$$P(Y_n = a|X_n = 2) = 1/6 = 1 - P(Y_n = b|X_n = 2).$$

One question about $(X_n, n \geq 0)$ could be: Giving the observation $Y_n = a$, $Y_{n+1} = a,\ Y_{n+2} = b$, what is the true underlying states at X_n, X_{n+1}, X_{n+2}?

This question reminds us of the maximum likelihood principle in classical statistics!

The most plausible triple of states which produces (emits) the observed (a, a, b) is the one which maximizes the above conditional probability function of (i, j, k). How to compute $P(X_n = i,\ X_{n+1} = j,\ X_{n+2=k}|Y_n = a,\ Y_{n+1} = a,\ Y_{n+2} = b\ )$ from the specified structures?

5. Unobserved Data

5.1. Auction Theory

There are various auction formats, i.e. different ways to conduct an auction, such as first-price, sealed-bid auction; oral, ascending-price; second -price , sealed-bid auction, etc... Of course there are compelling reasons for economists to study auctions. Since there is a variety of different selling mechanisms, how a seller should choose to sell products? The way in which buyers form their valuations of objects remains an open question in economics. Clearly, both the sellers and the buyers in an auction are making decisions to maximize their profits. To understand their behavior, we need to place ourselves in a decision framework. Now, when apply statistics to decision-making in economics, we need to know on what people base their decisions? The buzz word is *utility.* Game theory provides mathematical models for economic behavior. The task of statisticians is to use empirical data to assess the validity of economic principles, leading to economic forecasts.

5.2. Game Theory

Game theory is a name given to the important problem of understanding how people make decisions in societies (see e.g. [9]). We call the persons involved the players since they participate in a game, where we interprete games in a general sense. The structure of a game is this. There are n players. Each player i has her set of strategies S_i, and a payoff function

$$u_i : S_i \times S_2 \times ... \times S_n \to \mathbb{R}.$$

The main problem in game theory is how to predict the players' behavior, i.e. to make an educated guess at how the player will likely to play?! Each player will act in such a way to maximize her payfoff, but her payoff depends essentially on other players' actions! The solution is in the concept of Nash's equilibrium (the proof of its existence is based on fixed point theorems). The equilibrium of a game is the expected behavior of that game (of course, assuming that all

players are "rational"!), i.e. we could expect that, in their own interests, players will likely play strategies forming the equilibrium, as predicted by theory. In auctions, each bidder (player)'s willingness to pay for the object is unknown to the other bidders. Thus, auctions are static games of incomplete information.

The strategies $(s_1^*, ..., s_n^*)$ of the players form a (Nash) equilibrium of the game if, for each player i, s_i^* is player $i's$ best response to the strategies specified for the $n-1$ other players, $(s_1^*, ..., s_{i-1}^*, s_{i+1}^*, ..., s_n^*)$, i.e. if the other players choose $(s_1^*, ..., s_{i-1}^*, s_{i+1}^*, ..., s_n^*)$, then player i will make less "profit" (utility) if she chooses any other strategy s_i different than s_i^*.

Specifically, for any i, and $s_i \in S_i$,

$$u_i(s_1^*, ..., s_{i-1}^*, s_i, s_{i+1}^*, ..., s_n^*) \leq u_i(s_1^*, ..., s_{i-1}^*, s_i^*, s_{i+1}^*, ..., s_n^*).$$

In a static game of incomplete information, each player cannot solve the above optimization problem since $u_i(s_1^*, ..., s_{i-1}^*, s_i, s_{i+1}^*, ..., s_n^*)$ is not available to player i. There is a subtle point in games of incomplete information as compared to games of complete information. In games of complete information, we specify the sets of players' strategies S_i. In games of incomplete information, we only specify the sets of players' *actions,* A_i. Their strategies are "rules" leading to their actions from their "private knowledge" (about their own payoff functions) and "subjective beliefs" about other players' payoff functions.

Let specify the uncertainty of players about the payoff functions of other players. The idea that a player i know her own payoff function but may be uncertain about other's payoff functions is described as follows. Each player i knows the set of possible payoff functions of another player j, say, T_j, while not specifically which one. Thus each player i has set of possible payoff functions indexed by her "type T_i", i.e. the payoff function of player i, when players choose their actions $a_1, ..., a_n$, is of the form

$$u_i(a_1, ..., a_n; t_i),$$

where $t_i \in T_i$.Player i knows her t_i, but does not know the t_j of other players. In such a situation, player i has to guess the types of other players in order to choose her best course of action. Such a guess is possible if player i has some personal (subjective) belief on other players possible payoff functions.

The following specific model of static games of incomplete information is due to Harsanyi (1967). First, the uncertainty about players' actual payoff functions is viewed as follows: each $(t_1, ..., t_n)$ is "drawn" by nature with some prior distribution.

Let $t_{-i} = (t_1, ..., t_{i-1}, t_{i+1}, ..., t_n) \in T_{-i}$. The player i's belief about other players' types t_{-i} is the conditional probability $p_i(t_{-i}|t_i)$. Thus, a Bayesian game is specified as

$$G = \{A_i, T_i, p_i, u_i, i = 1, 2, ..., n\}.$$

Now, a strategy for player i is a function $s_i : T_i \to A_i$.The strategies $(s_1^*, ..., s_n^*)$ is a Bayes Nash equilibrium if, for each player i, and for each of i's types $t_i \in T_i$, $s_i^*(t_i)$ solves

$$\max_{a_i \in A_i} \sum_{t_{-i} \in T_{-i}} u_i(s_1^*(t_1), ..., s_{i-1}^*(t_{i-1}), a_i, s_{i+1}(t_{i+1}), ..., s_n^*(t_n); t_i) p_i(t_{-i}|t_i).$$

The interpretation is simple: each player's strategy must be a best response to the other players' strategies. At equilibrium, no player wants to change her strategy. Of course, existence of such equilibrium and how to get it are main issues in game theory! We will apply the concept of Bayes Nash equilibrium to auctions as providing the link between observed and unobserved auction data.

5.3. Statistics with Auction Data and Related Research Problems with Approximate Random Samples

Consider the simplest auction format, namely the "first-price, sealed-bid auction". There is one seller and n bidders. For each given object, the seller likes to know how the bidders evaluate the object, for obvious reason! However, bidders' valuations are only known to the bidders themselves, and not to the seller. Each bidder submits a bid in a sealed envelop. At a given time, the seller opens these envelops and the bidder who tendered the highest bid wins. So what are observables are the bids.

Let V denote the random variable representing the bidders' valuation of the object, and B the bid random variable, with distributions and densities (assuming existence) F_V, $F_B(f_V, f_B)$, respectively. Assume also that the valuations V_i and the bids B_i, $i = 1, 2, ..., n$, are I.I.D. V, B, respectively. To estimate F_V, we need the data V_i, but unfortunately they are not observable. Instead, we observe a realization of another variable B. Fortunately, there should have a "link" between V and B! This situation reminds us of the popular "hidden Markov models" in biology. Rather than postulate some form of links, i.e. propose a model, we take advantage of the fact that auctions are special "static games of incomplete information", i.e. can be viewed through the lens of Harsanyi's theory of noncooperative games of incomplete information. At (Bayes-Nash) equilibrium, we have

$$V_i = B_i + [F_B(B_i)/(n-1)f_B(B_i)], i = 1, 2, ..., n. \tag{8}$$

We need data from our variable of interest, namely V, to estimate its distribution function F_V, but we do not have it! Instead, we observe another variable B which is linked to V by the above equation. Unfortunately the functions F_B and f_B in that equation are unknown. However, they can be estimated by the random sample $B_1,\ B_2, ...,\ B_n$. If we do so and then compute the corresponding $V_i,\ i = 1, 2, ..., n$, then we only obtain an "approximate" random sample rather than the exact random sample. However, the type of approximate data is different. The "errors" in our available data come from different sources. Let pursue our quest for obtaining the approximate sample.

A two-stage estimation is needed.

(i) *First*, we need to "estimate" the unobserved V_i from the observed B_i via (8). But both F_B and f_B are unknown! We need to estimate both F_B and f_B from the $B_i's$. This is a *nonparametric estimation* problem.

Of course $F_B(x)$ can be consistently estimated by the empirical distribution function

$$F_{B,n}(x) = (1/n)\sum_{i=1}^{n} I(B_i \leq x),$$

but we also need an estimate for its density function f_B. Such an estimate cannot be derived from the estimator of F_B since $F_{B,n}(x)$ is a step function, resulting in identically zero derivative! Thus, we need *nonparametric estimation method for probability density functions*. For now, the kernel (functional) estimator of $f_B(.)$ is of the form

$$f_{B,n}(x) = (1/n\lambda_n)\sum_{i=1}^{n} k((x - B_i)/\lambda_n)).$$

(ii) *Step two*: With the estimates of both F_B and f_B in the link equation (8), we can at least recover approximatively the unobserved values on the variable V.

These are "pseudo-values" of the $V_i's$. They can be viewed as coarse data, in the sense that they are the V_i but recorded with "error". The source of the "measurement error" comes from the Equation (8) and the estimation error in our nonparametric estimation procedure. With the approximate values, say, $V_i^{'}$, we estimate F_V as usual.

6. Random Set Data

6.1. Random Sets

Roughly speaking, random sets are measurable functions with respect to some topology on the space of subsets, say, of $\mathbb{R}^d$, such as closed subsets. As such, random sets fall entirely with the general framework of probability theory, where, in all generalities, the probabilistic law of a random element is simply its probability measure on the range space. The case of finite random sets is much easier to digest. Let U be a finite set. Let S be a random set taking values as subsets of U (i.e. in the power set of U). Then the "distribution" function of S is $F(A) = P(S \subseteq A)$. Here, in the context of statistics, we look at a random set S as observed data containing the unobserved values of our random variable X of interest.

6.2. Random Sets as Coarse Data

Coarse data refer to observations with "low quality" such as interval valued statistics in which the true observations are only known to lie within bounds.

Coarse data is a special type of imprecise data. Specifically, let $X_1, X_2, ..., X_n$ be a random sample from X with unknown distribution function F. The $X_i's$ are not observable, and instead we observed an I.I.D. sample $S_1, S_2, ..., S_n$ of *random sets* from a random set population S. The "link" is this, we know $X_i \in S_i$ with probability one, $i = 1, 2, ..., n$. We say that the values of the X_i are coarsened, and S is a coarsening of X.

Let's look at a standard situation where coarse data are *set-valued observations*. While set-valued observations, i.e. outcomes of random experiments or records of natural phenomena, have different interpretations, depending on the goals of the analysis, such as tumor growth patterns in medical statistics, shape analysis, the specific situation related to coarse data is this. Let X be a random vector of interest. Either by performing a random experiment or observing X in a sample data $X_1, X_2, ..., X_n$, to discover, say, the distribution of X, we are unable to observe or measure this sample with accuracy. Instead, what we observe is a collection of sets $S_1, S_2, ..., S_n$ which contain the sample, i.e. $X_i \in S_i$, $i = 1, 2, ...n$. The statistical problem is the same, but instead of using $X_1, X_2, ..., X_n$, we only have at our disposal the coarse sample $S_1, S_2, ..., S_n$. This clearly is a generalization of multivariate statistical analysis. In order to analyse the set-valued observations, we need to model the observation process. Since probability theory provides us with a fairly general setting, namely random elements in general measurable spaces of arbitrary nature, such as metric spaces , we can simply view $S_1, S_2, ..., S_n$ as a random sample from a *random set* S which contains X almost surely, i.e. X is an almost sure selector of S, or the other way around, S is a coarsening of X. Random set models for coarse data turn out to be useful in exhibiting various uncertainty measures in artificial intelligence.

Consider the case where S is a coarsening of X on a finite set U, to avoid topological details. Let $F : 2^U$ (power set of U)$\rightarrow [0, 1]$, $F(A) = P(S \subseteq A)$ which is the distribution function of the random set S, in the sense that F determines the probability law of S . Indeed, the set-function F satisfies the following basic properties:

(i) $F(\varnothing) = 0, F(U) = 1$.

(ii) For any $n \geq 1$, and $A_1, A_2, ..., A_n$,

$$F(\cup_{i=1}^{n} A_i) \geq \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} F(\cap_{i \in I} A_i).$$

Proof. (i) follows from the fact that S is a non-empty random set with values in 2^U.

(ii) This is a weakening of Poincare's equality for probability measures. Observe that since U is finite,

$$F(A) = P(S \subseteq A) = \sum_{B \subseteq A} f(B),$$

where we set $f(B) = P(S = B)$

Let $J(B) = \{i \in \{1, 2, ..., n\} : B \subseteq A_i\}$. Clearly, $\{B : J(B) \neq \varnothing\} \subseteq \{B : B \subseteq \cup_{i=1}^{n} A_i\}$. Now,

$$\begin{aligned} P(\{B : J(B) \neq \varnothing\}) &= \sum_{B \subseteq U, J(B) \neq \varnothing} f(B) \\ &= \sum_{B \subseteq U, J(B) \neq \varnothing} f(B) \left[\sum_{\varnothing \neq I \subseteq J(B)} (-1)^{|I|+1} \right] \\ &= \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} \sum_{B \subseteq \cap_{i \in I} A_i} f(B) \\ &= \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} F(\cap_{i \in I} A_i). \end{aligned}$$

To see that the axiomatic theory of belief functions is precisely the axiomatization of distributions of random sets, exactly like the case of random variables, it suffices to show the converse. Let $F : 2^U \to [0, 1]$ such that:

(i) $F(\varnothing) = 0, F(U) = 1$.

(ii) For any $n \geq 1$, and $A_1, A_2, ..., A_n$,

$$F(\cup_{i=1}^{n} A_i) \geq \sum_{\varnothing \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} F(\cap_{i \in I} A_i).$$

Then there exist a probability space $(\Omega, \mathcal{A}, P)$ and a non-empty random set $S : \Omega \to 2^U$ such that $F(A) = P(S \subseteq A)$.

Indeed, by Mobius inversion, we have

$$f(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} F(B)$$

which is a bona fide probability density function on 2^U. ■

For general topological spaces U, see e.g. Nguyen [16].

References

1. T. Amemiya, *Advanced Econometrics,* Cambridge University Press, 1985.
2. O. Cappe', E. Moulines, and T. Ryden, *Inference in Hidden markov Models*, Springer-Verlag, 2005.
3. R. Carroll, et al., *Measurement Error in Nonlinear Models,* Chapman and Hall/CRC, 2006.
4. R. L. Cheu, H. T. Nguyen, T. Magoc, and V. Kreinovich, Logit discrete choice model: a new distribution-free justification, *Soft Computing* **13** (2009), 133–137.
5. D. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, 1985.

6. W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics,* Springer-verlag, 2005.
7. E. Frees, *Longitudinal and Panel Data,* Cambridge university Press, 2004.
8. W. A. Fuller, *Measurement Error Models,* J. Wiley, 1987.
9. R. Gibbons, *Game Theory for Applied Economists,* Princeton University Press, 1992.
10. P. Gustafson, *Measurement Error and Misclassification in Statistics and Epidemiology,* Chapman and Hall/CRC, 2004.
11. J. J. Heckman, Sample selection bias as a specification error, *Econometrica* **47** (1) (1979), 153–161.
12. J. L. Horowitz, *Semiparametric Methods in Economics,* Springer-Verlag, 1998.
13. V. Krishna, *Auction Theory,* Academic Press, 2002.
14. P. McCullagh and J. A. Nelder), *Generalized Linear Models,* Chapman and Hall, 1989.
15. D. McFadden, Conditional logit analysis of quantitative choice behavior, in: *Frontiers in Econometrics* (P. Zarembka, Ed.), Academic Press, 1973.
16. H. T. Nguyen, *An Introduction to Random Sets,* Chapman and Hall, 2006.
17. H. T. Nguyen, G. S. Rogers, and E. A. Walker, Estimation in change-point hazard rate models, *Biometrika* **71** (1984), 299–304.
18. H. J. Paarsch and H. Hong, *An Introduction to The Structural Econometrics of Auction Data,* The MIT Press, 2006.
19. T. D. Pham and H. T. Nguyen, Strong consistency of MLE in the change-point hazard tate model, *Statistics* **21** (1990), 299–304.
20. T. D. Pham and H. T. Nguyen, Boostrapping the change-point of a hazard rate, *Ann. Inst. Statist. Math.* **45** (2) (1993), 331–340.
21. R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, McGraw -Hill, 1998.
22. A. N. Shiryaev, *Essentials of Stochastic Finance,* World scientific, 2008.
23. J. Tobin, Estimation of relationships for limited dependent variables, *Econometrica* **26** (1958), 24–36.