# Introduction to Empirical Processes and Semiparametric Inference Lecture 01: Introduction and Overview

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

# What Is an Empirical Process?

An *empirical process* is a function or other quantity computed from a data sample.

*Empirical process methods* are a set of specialized techniques and ways of thinking that enable statistical inference and problem solving for empirical processes.

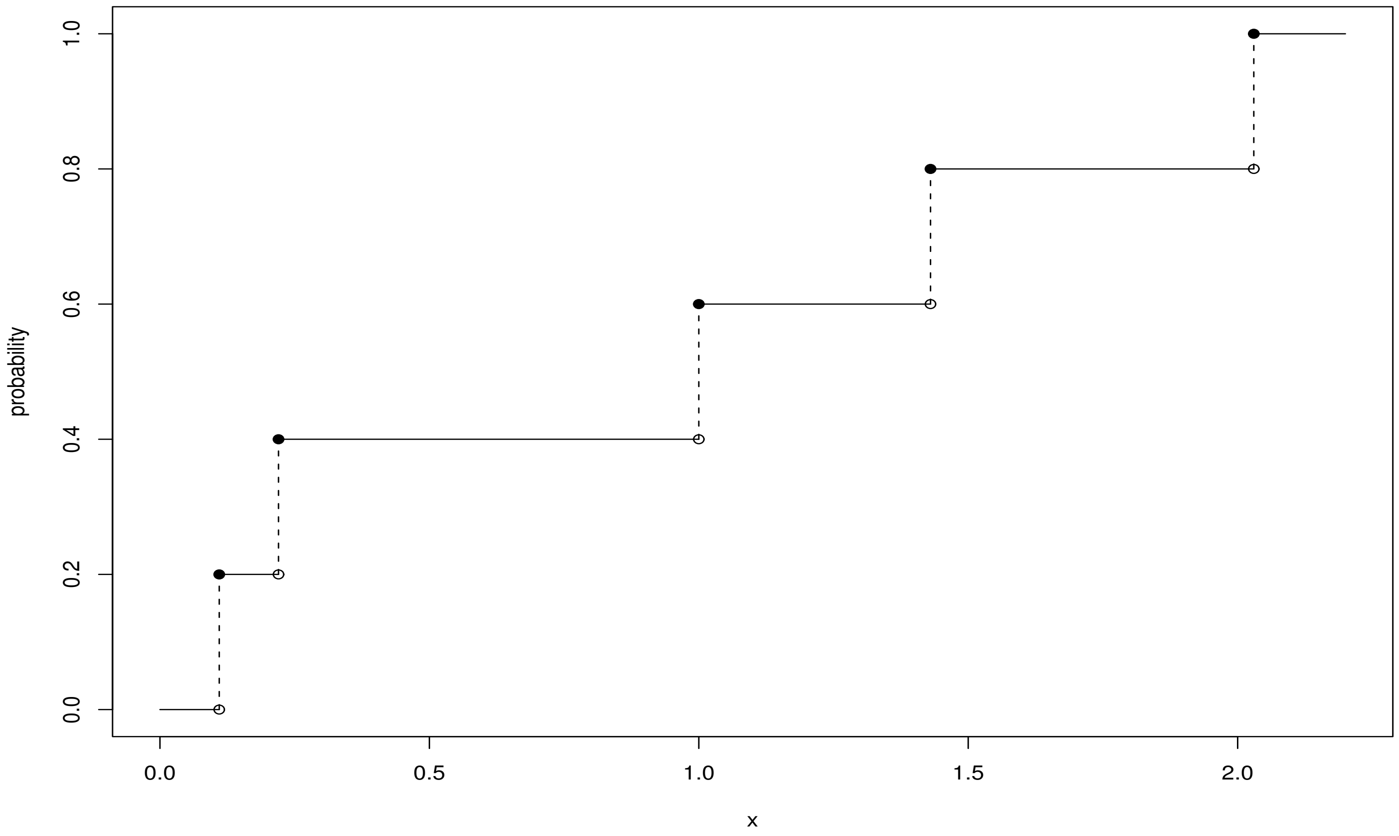
# Examples of Empirical Processes

Let $Y_1, \ldots, Y_n$ be a sample of body mass index measurements, and let

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq x\}.$$

$\mathbb{F}_n(x)$ viewed as a function of $x$ is the empirical distribution function (EDF) and is one of the simplest examples of an empirical process.

Figure 1. Plot of empirical cumulative distribution function.

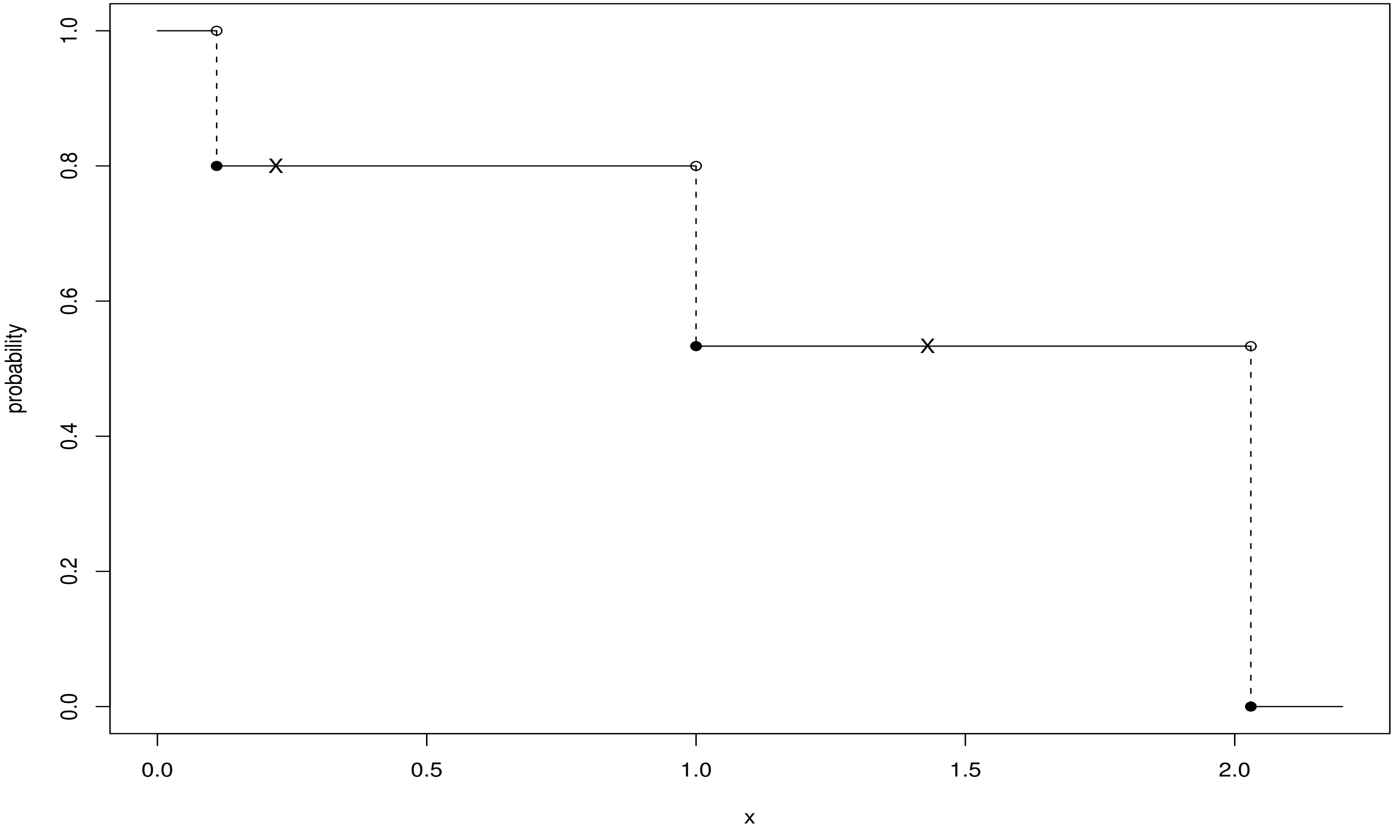


Let $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ be a sample of right-censored failure time observations (with no ties), and let

$$\hat{S}_n(x) = \prod_{i:X_i \le x} \left(1 - \frac{\delta_i}{\#\{j : X_j \ge X_i\}}\right).$$

$\hat{S}_n(x)$ is the Kaplan-Meier estimator of the survival function and is not quite so simple is the EDF but is still fairly simple as far as empirical processes go.

Figure 2. Plot of Kaplan-Meier estimator with censoring marked by “x”.



# Examples of Empirical Process Methods

One example of an empirical process method is the use of the *functional perspective*.

This involves

- viewing an empirical process as a random function or a sample path, rather than just as a set of numbers,
- and then drawing conclusions based on this perspective.

Constructing simultaneous confidence bands for Kaplan-Meier plots rather than just pointwise confidence bands is an example of using the functional perspective.

Viewing the sample median as an inverse of either $\mathbb{F}_n$ or $\hat{S}_n$ is another example of using the functional perspective:

- This approach is not strictly needed for the usual EDF $\mathbb{F}_n$ (there are “simpler” alternatives based on the binomial distribution).
- This approach is quite useful for the more complex Kaplan-Meier $\hat{S}_n$.

Another example of an empirical process method is the use of a *maximal inequality*:

THEOREM 1. *Let $X_1, X_2, \ldots$ be any (infinite) collection of real random variables (not necessarily independent or identically distributed) with “exponential tails,” i.e., for some $K, C < \infty$,*

$$P\left(|X_i| > x\right) \leq Ke^{-Cx},$$

*for all $x > 0$ and all $i \geq 1$.*

*Then there exists a universal constant $B < \infty$ such that*

$$E\left(\max_{1\leq i\leq n} |X_i|\right) \leq B\log(1+n),$$

*for all $n \geq 1$.*

Maximal inequalities, such as those given in Theorem 1,

- can be used to control the maximum of complicated collections of random quantities and data points;
- are used for some of the deepest and most powerful results in empirical processes;
- can lead to surprising results about testing and estimating for high dimensional data, such as microarray data, (an example will be given later).

# So What?

What does this have to do with biostatistics?

How does empirical process knowledge and research fit into the biostatistical paradigm?

“Biostatistics is the science of of obtaining, analyzing and interpreting data in order to understand and improve human health.” (from the departmental web site)

There are four basic “phases” of biostatistical research:

**A:** Human health research in substantive area;

**B:** Applying existing statistical methods (and study designs) to human health research;

**C:** Developing new methods for human health research;

**D:** Establishing foundational theory for developing new methods.

All four phases are human health research and must be done by someone familiar with practical issues in human health research.

Phase C (developing new methods) has three sub-phases:

**C1:** Developing new methods (and study designs);

**C2:** Validating new methods using simulation and other numerical studies: this shows that the methods work in the particular simulation settings evaluated;

**C3:** Validating using foundational theory: this shows the methods work in broad generality.

All of the phases are interdependent:

A$\Leftrightarrow$B: Health science research motivates the choice of methods, while the available selection of methods places restrictions on the kind of health science research that can be done.

B$\Leftrightarrow$C: Limitations in available methods can motivate the development of new methods, while availability of new methods can increase the breadth of health science research achievable using biostatistics.

C1⇔C2⇔C3:

- Sometimes new methods need to be created in spite of theoretical limitations, but those new methods need to be validated;
- problems arising in numerical or theoretical validation studies can motivate changes in methods that lead to better performance;
- simulation studies can inspire theoretical investigations, while theoretical phenomenae can suggest new simulation scenarios to investigate.

C⇔D: Unanswered theoretical questions raised by new methods can stimulate foundational research, while new theoretical results can inspire previously unimagined biostatistical methods.

The contribution of the field of biostatistics is maximized when these interdependencies are minded, managed and leveraged.

Most biostatisticians tend to spend most of their time in Phases B and C, as they should, with different balances depending on individual preferences and scientific needs.

It is also important that at least some biostatisticians spend significant effort in Phases A and D.

Even those statistician not spending effort in D need to be aware of developments in D—and be capable of applying those developments—in order to be successful in C.

Biostatistical participation in Phase A research is important since the uniquely biostatistical perspective can be highly effective in human health science.

Some biostatisticians should participate in Phase D research since

- theorists in other disciplines are not as aware of nor as interested in the theoretical concerns arising in biostatistics and
- biostatisticians are uniquely aware of and intuitively understand the important biostatistical issues arising in health research.

Empirical processes play a crucial role in Phases C3 and D.

Empirical processes have essentially replaced martingales in terms of relative importance in biostatistical research.

The importance of empirical processes will continue to increase during the next decade and beyond.

Empirical processes are quite broadly applicable and useful in both survival and non-survival settings, including

- clinical trials,
- analysis of microarrays,
- regression analysis,
- many other biostatistical areas,
- econometrics,
- many other quantitative disciplines.

# Example 1: Adjudicated Endpoints

In some clinical trials in cancer, cardiology, AIDS, and in vaccine trials, endpoints can be difficult to assess.

For example, in the Coumadin Aspirin Reinfarction Study (CARS) (1997), there were 12 cardiovascular event types, only three of which were *primary endpoint constituents*:

- reinfarction,
- stroke, or
- cardiovascular death.

Unfortunately, it is difficult to determine whether an event at the time of observation was, in fact, a primary endpoint.

An Event Classification Committee (ECC) (also called an Event Adjudication Committee) reviews the medical records and provides a final endpoint classification.

There is typically a significant delay between the original observation and the final classification from the ECC.

Thus, at a given analysis time, there are both adjudicated and unadjudicated endpoints.

In Cook and Kosorok (2004, *JASA*), a method for valid analysis was developed for time-to-event data with incomplete event adjudication.

Empirical processes played an essential role in

- the derivation of the method,
- the validation of the method, and in
- developing diagnostics to assess assumption validity.

Let $N_i(t)$ be the usual counting process in survival analysis that indicates the timing of an observed event for individual $i$:

$$N_i(t) = \delta_i \mathbf{1}\{X_i \le t\},$$

where $(X_i, \delta_i)$ is a “standard” right-censored failure time observation pair.

The basic idea is to compute an estimate $\tilde{N}_i(t)$ from available data for the counting process for the first observed primary event.

$\tilde{N}_i(t)$ starts at zero when $t = 0$ and is updated whenever individual $i$ has a “candidate event” at, say, time $t$.

If the event at time $t$ has been adjudicated, then,

- if it is adjudicated to be a primary event, $\tilde{N}_i(t)$ jumps to 1;
- if not, no jump in $\tilde{N}_i(t)$ occurs.

If the event at time $t$ is unadjudicated, then,

- the probability $q$ of the event being the first occurrence of a primary event is estimated from a regression model based on adjudicated observations and individual-specific covariates;
- $\tilde{N}_i(t)$ jumps with a size $\hat{q}(1 - \tilde{N}_i(t-))$.

Similar calculations are used to obtained an estimated at-risk process $\tilde{Y}_i(t)$.

Both $\tilde{N}_i$ and $\tilde{Y}_i$ are plugged-in to standard counting process formulations to compute Kaplan-Meier and Cox model estimates as well as log-rank statistics, and needed variances are estimated.

Theoretical and simulation studies verify that the approach works quite well.

We are currently collaborating with Merck to extend this methodology to vaccine safety studies with rare adverse events.

# Example 2: High Dimensional Gene Expression Data

In gene expression studies, microarrays are used to evaluate the difference between two (or more) groups in terms of the expression of thousands of genes.

A common question to ask is which of the many thousands of genes are expressed significantly differently (“differentially expressed”) between the two groups.

Within each group, a sample of microarrays are used to measure the expression levels for each of thousands of genes.

For each gene, a two-sample t-test is computed to assess whether that particular gene’s expression levels differ between the two groups.

A p-value based on normal quantiles is then calculated for each gene, and the “significant” genes are then identified based on how small the p-values are.

Typically, there may be $\approx 10,000$ genes being assessed and only around 20 arrays per group (or fewer): this is called the “large p, small n” paradigm.

The *false discovery rate* (FDR) is usually used to assess significance since family-wise error rate is considered much too conservative.

There are two important issues in this setting:

1. The question of whether p-values that rely on the central limit theorem are uniformly valid when the number of p-values is exponentially larger than the sample size.
2. There are often outliers in these settings and methods more robust than t-tests, such as median-based tests, are important.
   - The asymptotic validity question for median-based tests is even more precarious than that for t-tests.

In Kosorok and Ma (2007, *AOS*), minimax inequalities from empirical processes are used to show that the central limit theorem for p-values based on t-tests and median tests for the large-p small-n paradigm is valid in broad generality.

The results are applied in Ma, Kosorok, Huang, et al. (2006, *Biometrics*) to develop robust regression methodology for microarrays in the large-p, small-n paradigm.

These results are important because they evaluate the appropriateness of commonly used procedures for high-dimensional data.

We are currently working on refinements of these procedures to improve practical performance in individual gene expression studies and gene network studies.

# Example 3: Classification and Regression Trees

A classification and regression tree (CART) is a regression model based on cut-points in the predictor variables.

Let $X_1, \ldots, X_d$ be variables to be used to predict an outcome $Y$.

Here are the basic components of CART regression:

- A predictor variable is selected and a cut-point obtained to divide the data into two groups.
- The groups are further divided, possibly using a different predictor, until the process is stopped.
- The final groups that are not further subdivided are called leaves, and the means of the leaves are the predicted values of $Y$ in the regression.
- Decisions on variable and cut-point selection are made sequentially using a least-squares error criterion.

An important question is how to perform inference on the cut-points and regression predictions.

The distribution of the cut-point estimator depends heavily on whether the predictor variable has a true cut-point with the mean of $Y$ differing above and below the cut-point.

Whether or not there is a true cut-point, the limiting distribution is not Gaussian, and the derivation of the limiting distribution requires careful empirical process arguments.

Moreover, the rate of convergence (usually this is $\sqrt{n}$) is different for these two situations, with one rate being $n$ (when there is a true cut-point) and the other being $n^{1/3}$ (when there is no true cut-point).

The calculations involved make use of very delicate empirical process “entropy” techniques which precisely quantify the complexity of the underlying statistical model.

What happens in this relatively simple setting is still not known, and we are currently working on sorting this out.

Ideally, one would like to be able to perform inference in this simple setting without having to know what the true distribution of the data is.

# Example 4: Personalized Medicine Trials

- These are a new kind of trial for finding and validating personalized treatment regimens.
- The basic idea is to find a rule based on prognostic data (including possibly genetic and/or genomic data) to create a “table” which tells a clinician which treatment is best for which patients.
- Statistical inference and design questions (including sample size formulas) are very difficult in this setting, and empirical process methods seem to be essential to making progress in this area.
- We are currently applying this methodology to treatment of non-small cell lung cancer, colon cancer, and cystic fibrosis.

# Empirical Processes and Semiparametric Inference

The entire field of semiparametrics depends heavily on empirical process methods:

- Semiparametric models are models with both a parametric (finite dimensional) and a nonparametric (infinite dimensional) component.
- The term nonparametric is usual reserved for models with only infinite-dimensional components (or for statistical procedures that do not require knowledge of underlying distributions).
- In semiparametric models, the parametric part is for scientific interpretability, while the nonparametric part is for flexibility.
- A major question for semiparametric models is how to perform efficient inference: this is the quest of semiparametric inference.

# What Is Semiparametric Inference?

*Semiparametric inference* is the study of inference for semiparametric models.

A major, defining component is the question of efficiency.

*Semiparametric inference methods* are a set of specialized techniques and ways of thinking about semiparametric models that enable deriving and evaluating inferential procedures, assessing efficiency, and comprehending scientific properties.

The key technical tools for semiparametric inference are empirical processes and Hilbert space methods.

# Example 1 of a Semiparametric Model

Consider the semiparametric model

$$Y = \beta' Z + e, \tag{1}$$

where

- $\beta, Z \in \mathbb{R}^p$ are restricted to bounded sets (for convenience),
- $(Y, Z)$ are the observed data,
- $E[e|Z] = 0$ and $E[e^2|Z] \leq K < \infty$ almost surely,
- $E[ZZ']$ is positive definite, and
- the joint distribution of $(e, Z)$ is otherwise unrestricted (nonparametric).

Given an i.i.d. sample of data generated by (1), $(Y_i, Z_i), i = 1, \ldots, n$, we are interested in

- estimating $\beta$ (the component of scientific interest)
- in a flexible manner (reflected in the nonparametric joint distribution of $(e, Z)$).

There are several estimators of $\beta$ to choose from: how do we decide between them?

Consider first the least squares estimator

$$\hat{\beta} = \left[ \sum_{i=1}^{n} Z_i Z_i' \right]^{-1} \sum_{i=1}^{n} Z_i Y_i,$$

for which $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with mean zero and bounded variance $V_0$.

Is $V_0$ the lowest possible variance among “reasonable” estimators?

The answer is:

- Yes, if $e$ is independent of $Z$ and mean zero Gaussian.
- No, if $E[e^2|Z = z]$ is a nonconstant function of $z$.

Some of the goals of semiparametric inference are to:

- Define clearly what “reasonable” (regular) means.
- Find the asymptotic form of the efficient estimator.
- Find an estimator that achieves this asymptotically.

Now suppose we want to figure out whether the residual $e$ is Gaussian.

We could estimate nonparametrically the distribution function $F$ for $e$ with

$$\hat{F}(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\left\{Y_i - \hat{\beta}' Z_i \leq t\right\}.$$

Empirical process methods allow us to determine that $\sqrt{n}(\hat{F} - F)$ converges weakly to a Gaussian process.

We can use this to develop a hypothesis test of Guassianity.

# Example 2 of a Semiparametric Model

Suppose we observe the random triplet $X = (Y, Z, U)$, where $Z \in \mathbb{R}^p$ and $U \in \mathbb{R}$ are covariates that are not linearly dependent, and $Y$ is a dichotomous outcome with

$$E\{Y|Z,U\} \quad = \quad \nu\left[\beta' Z + \eta(U)\right], \qquad (2)$$

where $\beta \in \mathbb{R}^p$, $Z$ is restricted to a bounded set, $U \in [0,1]$, $\nu(t) = 1/(1+e^{-t})$, and $\eta : [0,1] \mapsto \mathbb{R}$ is an unknown smooth function.

We assume that the first $k-1$ derivatives of $\eta$ exist and are absolutely continuous, with

$$J^2(\eta) \equiv \int_0^1 \left[\eta^{(k)}(t)\right]^2 dt < \infty.$$

Given an i.i.d. sample of $X_i = (Y_i, Z_i, U_i)$, $i = 1, \ldots, n$, we wish to estimate $\beta$ and $\eta$.

The conditional density at $Y = y$ given the covariates $(Z, U) = (z, u)$ is

$$p_{\beta,\eta}(x) = \left\{\nu\left[\beta' z + \eta(u)\right]\right\}^y \left\{1 - \nu\left[\beta' z + \eta(u)\right]\right\}^{1-y} :$$

- Direct maximization yields a function $\hat{\eta}$ with $\hat{\eta}(u_i) = \text{sign}(y_i - 1/2)\infty$;
- We need to restrict $\hat{\eta}$ somehow, e.g., via penalization.

We can use the penalized log-likelihood

$$\tilde{L}_n(\beta, \eta) = n^{-1} \sum_{i=1}^{n} \log p_{\beta,\eta}(X_i) - \hat{\lambda}_n^2 J^2(\eta),$$

where $\hat{\lambda}_n$ is a *smoothing parameter*, where:

- Large $\hat{\lambda}_n$ forces $\hat{\eta}$ to be smoother;
- Small $\hat{\lambda}_n$ forces $\hat{\eta}$ to fit better.

If we choose the smoothing parameter so that $\hat{\lambda}_n = o_P(n^{-1/4})$ and $\hat{\lambda}_n^{-1} = O_P(n^{k/(2k+1)})$, then:

- $\sqrt{n}(\hat{\beta} - \beta)$ converges to a mean zero Gaussian vector;
- $\hat{\eta}$ is uniformly consistent for $\eta$, i.e.,

$$\|\hat{\eta} - \eta\|_{[0,1]} \equiv \sup_{u \in [0,1]} |\hat{\eta}(u) - \eta(u)| = o_P(1);$$

- The $L_{P,2}$ norm of $\hat{\eta} - \eta$ is $O_P(n^{-k/(2k+1)})$, i.e.,

$$\|\hat{\eta} - \eta\|_{U,2} \equiv \left\{P\left[|\hat{\eta}(U) - \eta(U)|^2\right]\right\}^{1/2} = O_P(n^{-k/(2k+1)});$$

- $\hat{\beta}$ is efficient.

In this example, the nuisance parameter $\eta$ is not $\sqrt{n}$ consistent, but $\hat{\beta}$ is both $\sqrt{n}$ consistent and efficient:

- We first need to establish consistency of the estimators (not necessarily uniform);
- We then need to obtain the appropriate convergence rates for all parameters;
- We then need to show (if possible) efficiency for $\sqrt{n}$ consistent estimators;
- Then we need to find a way to conduct inference (e.g., bootstrapping).

# The Big Picture

- Empirical processes are tools for studying complex statistics;
- Semiparametrics is an important area of application for empirical processes;
- Semiparametric models are very important in biostatistics (and econometrics, etc.);
- We will begin by giving an overview of the key concepts without proofs;
- We will then reexamine concepts in greater detail with some proofs;
- Do not worry too much about understanding everything (be patient).