(Semi)Parametric Models
vs
Nonparametric Models

Probability Models

I define parametric, semiparametric, and nonparametric models in the two sample setting

My definition of semiparametric models is a little stronger than some statisticians

- The distinction is to isolate models with assumptions that I think too strong

Notation for two sample probability model

$$\text{Treatment}: \quad X_1, \ldots, X_n \overset{iid}{\sim} F$$

$$\text{Control}: \quad Y_1, \ldots, Y_m \overset{iid}{\sim} G$$

Probability Models

Parametric models

F, *G* are known up to some finite dimensional parameter vectors

$$F(t) = \Psi\left(t, \vec{\Phi}_X\right)$$

$$G(t) = \Psi\left(t, \vec{\Phi}_Y\right)$$

where:

$\Psi(\cdot,\cdot)$ has known form

$\vec{\Phi}$ is finite dimensional and unknown

Probability Models

Parametric models: Examples

$$\text{Normal}: \quad X_i \sim N\left(\mu, \sigma^2\right) \quad Y_j \sim N\left(\nu, \tau^2\right)$$

$$\text{Bernoulli}: \quad X_i \sim B(1, \mu) \quad Y_j \sim B(1, \nu)$$

$$\text{Exponential}: \quad X_i \sim E(\mu) \quad Y_j \sim E(\nu)$$

Probability Models

Semiparametric models

Forms of *F, G* are unknown, but related to each other by some finite dimensional parameter vector

- *G* can be determined from *F* and a finite dimensional parameter
- (Most often: under the null hypothesis, F = G)

Probability Models

Semiparametric models

Forms of *F, G* are unknown, but related to each other by some finite dimensional parameter vector

- *G* can be determined from *F* and a finite dimensional parameter

$$F(t) = \Psi(t, \vec{\Phi}_X)$$
$$G(t) = \Psi(t, \vec{\Phi}_Y)$$

where:

$\Psi(\cdot,\cdot)$ has unknown form (in t)

$\vec{\Phi}_X$ is finite dimensional and known (identifiability)

$\vec{\Phi}_Y$ is finite dimensional and unknown

Probability Models

Semiparametric models: Examples

Shift: $G(t) = F(t - \mu)$

Shift - scale: $G(t) = F\left(\frac{t - \mu}{\sigma}\right)$

Accel failure: $G(t) = F(t\gamma)$

Prop hzd: $1 - G(t) = [1 - F(t)]^{\gamma}$

Probability Models

Nonparametric models

Forms of *F, G* are completely arbitrary and unknown

- An infinite dimensional parameter is needed to derive the form of *G* from *F*
- (I demand that the above hold under <u>all</u> hypotheses, unless the test is consistent when F ≠ G)

Examples of truly nonparametric analyses:

- Kolmogorov-Smirnov test
- t-test with unequal variances (large samples)

The Problem: A Logical Disconnect

"Because the light is so much better here under the streetlamp"

- a drunk looking for the keys he lost half a block away

The Problem

In the development of statistical models, and even moreso in the teaching of statistics, parametric probability models have received undue emphasis

Examples:

- t test is typically presented in the context of the normal probability model
- theory of linear models stresses small sample properties
- random effects specified parametrically
- Bayesian (and especially hierarchical Bayes) models are replete with parametric distributions

The Problem

ASSERTION: Such emphasis is not typically in keeping with the state of knowledge as an experiment is being conducted

The parametric assumptions are more detailed than the hypothesis being tested, e.g.,:

- Question: How does the intervention affect the first moment of the probability distribution?
- Assumption: We know how the intervention affects the 2nd, 3rd, …, ∞ central moments of the probability distribution.

The Problem

Conditions under which an intervention might be expected to affect many aspects of a probability distribution

Example 1: Cell proliferation in cancer prevention

- Within subject distribution of outcome is skewed (cancer is a focal disease)
- Such skewed measurements are only observed in a subset of the subjects
- The intervention affects only hyperproliferation (our ideal)

The Problem

Conditions under which an intervention might be expected to affect many aspects of a probability distribution (cont.)

Example 2: Treatment of hypertension

- Hypertension has multiple causes
- Any given intervention might treat only subgroups of subjects (and subgroup membership is a latent variable)
- The treated population has a mixture distribution
  - (and note that we might expect greater variance in the group with the lower mean)

The Problem

Conditions under which an intervention might be expected to affect many aspects of a probability distribution (cont.)

Example 3: Effects on rates

- The intervention affects rates
- The outcome measures a cumulative state
- Arbitrarily complex mean-variance relationships can result

The Problem

These and other mechanisms would seem to make it likely that the problems in which a fully parametric model or even a semiparametric model is correct constitute a set of measure zero

Exception: independent binary data must be binomially distributed in the population from which they were sampled randomly (exchangeably?)

The Problem

Impact on what we teach about optimality of statistical models

Clearly, parametric theory may be irrelevant in an exact sense (though as guidelines it is still useful)

Much of what we teach about the optimality of nonparametric tests is based on semiparametric models

- e.g., Lehmann, 1975: location-shift models

The Problem

Example: the Wilcoxon rank sum test

Common teaching:

- Not too bad against normal data
- Better than t test when data have heavy tails

More accurate guidelines:

- Above holds when a shift model holds for some monotonic transformation of the data
- If propensity to outliers (mixture distributions) is different between groups, the t test may be better even in presence of heavy tails
- In the general case, the t test and the Wilcoxon are not testing the same summary measure