# Sampling Design Considerations: Probability Vs Targeted State-wide Vs Regional

Don Stevens

Statistics Department Oregon State University

Presented at

California Aquatic Bioassment Workshop

Sacramento, CA

November 27-28, 2001

The research described in this article has been funded in part by the U.S. Environmental Protection Agency. This document has been prepared through Cooperative Agreement CR82-9096-01 to Oregon State University. It has been not been subjected to Agency review. The conclusions and opinions are solely those of the author and are not necessarily the views of the Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

# Options to characterize a population

- Census: observe every element in population
- Sample: observe selected elements in population and extrapolate properties to population characteristics

# Sampling Design Considerations

**BY A SMALL SAMPLE, WE MAY JUDGE THE WHOLE PIECE**

---- Miguel de Cervantes

# Sampling Design Considerations

**Small Sample Þ Whole Piece**

**INFERENCE / EXTRAPOLATION**

# Sampling Design Considerations

**Observation**

**In the space of one hundred and seventy six years, the Lower Mississippi has shortened itself by two hundred and forty two miles. That is an average of a trifle over one mile and a third every year.**

# Sampling Design Considerations

**Inference/Extrapolation**

**Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million years ago next November, the Lower Mississippi was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing rod.**

**There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.**

**---- Mark Twain**

# Population Inference

- Base the inference on explicit specification of the relationship between the selected sites and the entirety ("Model-based")

- Select sites via a probability sample and use survey sample methods. ("Design-based")

# Model-based Inference

- Model: generic term for the description of the relationship between the subset and the population
- Examples of “Models”
  - Observations are “representative” of the population
  - Choice of sample locations: a sample taken at the basin outflow “integrates” or “averages” conditions within the basin
  - Statistical techniques (spatial statistics, analysis of variance, regression, principal components, ordination) relating sample to population characteristics
  - Process models that explicitly represent physical & biological processes.

# Advantages of Model-based Inference

- Permits very general and precise inference from limited data.

- Inference "borrows strength" from the model: the model structure provides the framework for the inference, and the precision of the inference is judged relative to the model.

# Disadvantages of Model-based Inference

- Model structure provides the framework for the inference
  - If model is not a good description of reality, the inference may have little resemblance to the true population characteristic
- Precision is judged relative to the model
  - There may be no indication that the inference is substantially in error

# Design-based Inference

- Probability-based selection method
- Generality and validity comes from the design
- Inference from a properly executed design is unassailable and irrefutable.

# Design-based Inference

- Can incorporate prior knowledge and understanding in both the design and analysis phases
  - Variable probability used to focus sample.
  - Model-assisted, design-based analysis allows extensive use of models in the analysis while maintaining the design basis
    - Good model improves precision
    - Bad model decreases precision, but inference remains valid

# Sampling Design Considerations Commonly Held Concepts

- An experienced resource manager can pick sites better than a random process can
- Lots of data can substitute for random sampling
- Environmental resources have such large variability that stratification is essential to control it.
- Replicate determinations are essential to overcome the ever present biological variation.

# Commonly Held Concepts TRUE or FALSE?

Many competent, responsible scientists think these statements are generally true. However, for our objective, they are generally FALSE!

# Commonly Held Concepts What is the Problem?

- Statistics courses commonly focus on experimental, not observational, studies.
  - Good experimental design is not good sampling design
- Statistics profession has failed to communicate some concepts as clearly as we/they should have.

# Probability Vs Targeted

An experienced resource manager can pick sites better than a random process can.

- "Typical" sites are usually much more homogeneous than the larger context of interest.
- Non-probability samples can be badly biased for no apparent reason.
- Typical for one set of responses says nothing about typical for any other response, that is, any response not used in determining typical.
- Human reasoning is notoriously poor at integration.
- Most people are Bayesian with very narrow priors

# Probability Vs Targeted

## Lots of data can substitute for random sampling

- Example taken from "*Sample Representativeness: A Must for Reliable Regional Estimates of Lake Condition*", Peterson, et al., *ES&T* **33:**1559-1565.
  - EMAP probability sample of all lakes in northeast US evaluated Secchi transparency
  - Great American Dip-In: 5,000 volunteers in various monitoring programs were asked to evaluate Secchi transparency in "their" lakes.

# Secchi CDF

# Probability Vs Targeted

Environmental resources have such large variability that stratification is essential to control it

- Widely held view, based in experimental design.
- Variability is an intrinsic characteristic of the population, we cannot eliminate it, and should not try.
- Stratification can be detrimental, as next example shows.

# Probability Vs Targeted

- Stratify to control variation: Classify sites as good, fair, or poor, and vary sampling effort by class, according to objective.
  - Cochran (1977) list 4 reasons to stratify:
    - Analytical convenience
    - Administrative convenience
    - Operational convenience
    - Potential increase in precision
  - Dirty secret in sampling: Stratification must be almost perfect for this strategy to increase precision.

# Sampling Streams to Estimate Number of Spawning Salmon

Oregon Department of Fish & Wildlife

- Objective: Estimate number of coho salmon in Oregon coastal streams
- Stratified stream segments into low, moderate, or high quality spawning habitat
- Low was not sampled; high was sampled at three times the rate of moderate.
- Habitat quality was evaluated for each sampled segment

# Sampling Streams for Salmon

| Stratification Class | Observed Class | | |
|---|---|---|---|
| | Low | Moderate | High |
| Low | **NA** | **NA** | **NA** |
| Moderate | **73%** | **17%** | **10%** |
| High | **53%** | **23%** | **24%** |

# State-wide Vs Regional

- Must have common objective
- Similar to stratification issue
  - Independent regional designs are unlikely to result in increased statewide precision
  - Focus on known regional problems does not address objective
  - Justification for regional designs is convenience
    - Administrative
    - Operational
    - Analytical

State-wide Vs Regional

# State-wide Advantages

- Coherent design allows easier combination across regions
  - Watershed that straddles several regions
- Generally better state-wide precision
- Encourages common protocols