DESIGN OF SURVEYS

Temporal heterogeneity

-- populations change through time
--potentially confounds inferences
--distinguish systematic population changes from those occurring as a result of random environmental variation
--approaches:

1) random replication at points in time (impractical)
2) modeling time effects

Spatial heterogeneity

-- sample so as to include spatial heterogeneity: estimate and model its effects
-- concentrate on a single spatial component, model it, but restrict inferences to that cohort (patch, etc.)

Sampling variability

Even if spatial and temporal heterogeneity are modeled, still need to consider effects of estimation based on random samples. Depends on assumptions of sample design, leading to particular models for sample distributions

Detectability

random sampling often is insufficient by itself to guarantee reliable sample-based estimators. 1 reason is because of incomplete detectability

consider sample counts C_i

$$E(C_i) = \beta_i N_i.$$

***Complete detectability*:**

$$\beta_i = \beta = 1.$$

In this situation individuals are completely detectable over time, space, or

other dimensions, so that the sample count is actually identical to N_i.

***Less than complete but constant detectability*: $\beta_i = \beta < 1$. The count C_i is a biased estimate of N_i by the factor $\$_i$, but the bias is uniform over time, space, or other dimensions. May be used as an index.**

***Variable detectability*: $\beta_i < 1$ and $\beta_i \neq \beta_j$. Estimates of N_i are biased, and the bias is non-uniform over time, space, or other dimensions, adding variability to the estimates in addition to bias. May also mask experimental effects, or result in illusion of real effects.**

Distinguish between the *target population*, the object of an investigation, and the *sampled population* from which samples are actually taken (Fig. 5.1).

***Replication* — necessary to assess the variability of sample estimates**

***Randomization* — protects against the systematic influence of unrecognized sources of variation required for inferences about the population from which samples are taken, and it allows for the estimation of sample-to-sample variance**

***Control of variation*-- 1)increases the precision of parameter estimates.**
2) increase in the power of hypothesis tests.
Methods:
(i) the use of stratification or blocking to eliminate systematic variation;
(ii) the use of ancillary variates to eliminate nuisance sample-to-sample variation; and
(iii) increasing sample size so as to increase estimator precision

FINITE SAMPLES

***N* sampling units (e.g., individual organisms, kinship units, plots of land), each of which can be characterized by some measurable attribute y_i.**

Thus, the population is represented by the set $\{y_1,..., y_N\}$. Sampling produces a sample set of these values, which then can be combined into estimates of population parameters.

We focus below on the population mean

$$\bar{Y} = \sum_{i=1}^{N} y_i / N,$$

the population total

$$Y = N\bar{Y},$$

the population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N},$$

Simple Random sampling

***n* sampling units are selected from a list of *N* total units in the sampled population (*sampling frame*) includes the whole population under investigation *simple random sampling*, in which the units are drawn so that each unit has the same probability *n/N* of being selected**

(See estimation formulas in Chapter 5)

***Example*. Counts on sample quadrats are used to estimate abundance for cottontail rabbits (*Sylvilagus floridanus*) on a 1000-ha study area.**

The area is divided into 1-ha plots, and 100 plots are selected at random. Each plot is surrounded by a wire barrier, and field workers drive the rabbits

into an enclosure on the plot, where a complete count for the plot is made.

The results from the sample plots provide a sample mean $\bar{y}=16$ **and variance** $s^2=40$**. Application of equation (5.6) produces the estimate for total abundance of**

$$\begin{aligned}\hat{Y} &= N\bar{y} \\ &= 1000(16) \\ &= 16000,\end{aligned}$$

with variance given by equation (5.7):

$$\begin{aligned}v\hat{a}r(\hat{Y}) &= \frac{(1000-100)(1000)}{100}(40) \\ &= 360{,}000.\end{aligned}$$

Because of the large sample size, an approximate

approximate 95% confidence interval for $\hat{Y}$ **based on a normal distribution is**

$$\left(\hat{Y}-z_{.05}\sqrt{v\hat{a}r(\hat{Y})}\ ,\ \hat{Y}+z_{.05}\sqrt{v\hat{a}r(\hat{Y})}\right) = (14824{,}17176)$$

Stratification and stratified random sampling

- **Stratify population into K relatively homogenous strata, take sample from each stratum.**
- **Reduces variation of mean or total estimate**
- **Disperses sample through population**

Proportional allocation– allocate samples proportional to size of strata

Optimal allocation– takes into account per-stratum cost and variance

OTHER SAMPLING APPROACHES

! *Cluster sampling*

- " **investigate a *sample* of the clusters in order to estimate population parameters. That is, the clusters themselves become (primary) sampling units. (unlike stratified sampling in which all strata must be sampled)**
- " **once a sample of primary units is selected, *all* the secondary units from each primary unit are included in the sample**

! *Systematic sampling*

- " **an initial unit is selected, typically at random, from the first $k=N/n$ units. Then every k^{th} subsequent unit is selected, until n total units are selected.**
- " **equivalent to random sampling if the ordering of the individuals is independent of the attribute being measured and the ordering does not result in a nonrepresentative sample being drawn.**
- " **must be used with caution in ecological populations, because it is often impossible to rule out non-random ordering of the sample units**

! *Double Sampling*

- " **Often the variables of interest in a sampling design are difficult or costly to measure, but correlated auxiliary variables can be identified that are cheaper or easier to measure.**
- " **measure an auxiliary variable (x_i) on a sample of n' units, and measure the primary variable (y_i) on a subsample of size n of these units, where n typically is much smaller than n'.**
- " **-Ratio or regression methods then can be used to predict values of y_i for the larger sample, and if x_i and y_i are highly correlated, the precision of population estimators based on the predicted values can be improved substantially.**

- **Adaptive sampling**
  - **involves the selection of sampling units from a population consisting of N units with associated values $\{y_1,...,y_N\}$.**
  - **assigns a probability to every possible sample.**
  - **the selection probabilities at each point depend on the values for previously selected sampling units.**

PROBLEMS IN SAMPLING DESIGNS

Failure to properly define target and sampled populations

3 possible courses of action.

- **Redesign the survey so that the sampled and target populations correspond.**
- **use auxiliary information to establish a predictive relationship between the sample elements in the target population and the sampled subset. (special case of double sampling in which all portions of the target population are ultimately involved in the sample)**
- **Redefine the target population.**

Must avoid in any approach is *ad hoc* selection of the areas to be sampled, and *ad hoc* selection of the sampling units within areas

Lack of replication or "pseudo-replication"

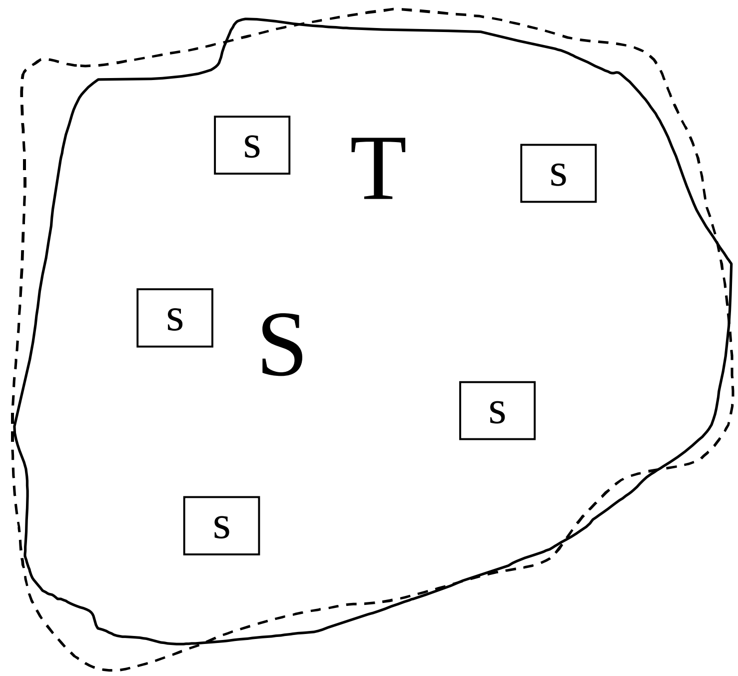
Examples:

- **Measuring a single point in space over time**
- **Single sample is divided into multiple sub-samples. For example a single 100-m transect along which animal counts are made could be divided into 10-m intervals, with a "variance" computed based on this**

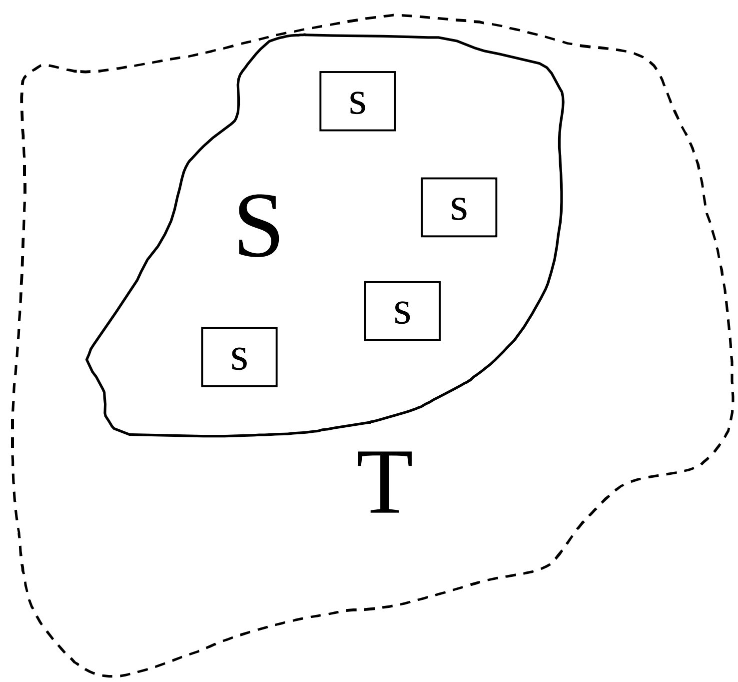
Misinterpretation of pattern as cause and effect

- **Patterns (e.g., transects) correlated with gradients**
- **Statistical correlations in estimates**

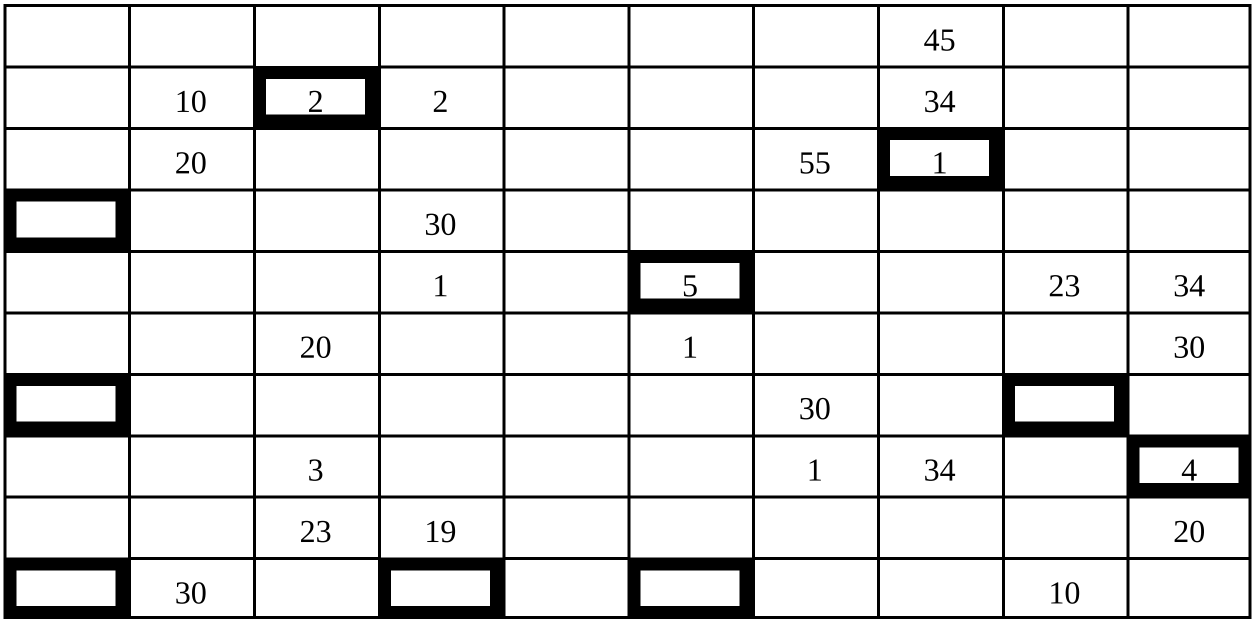
a



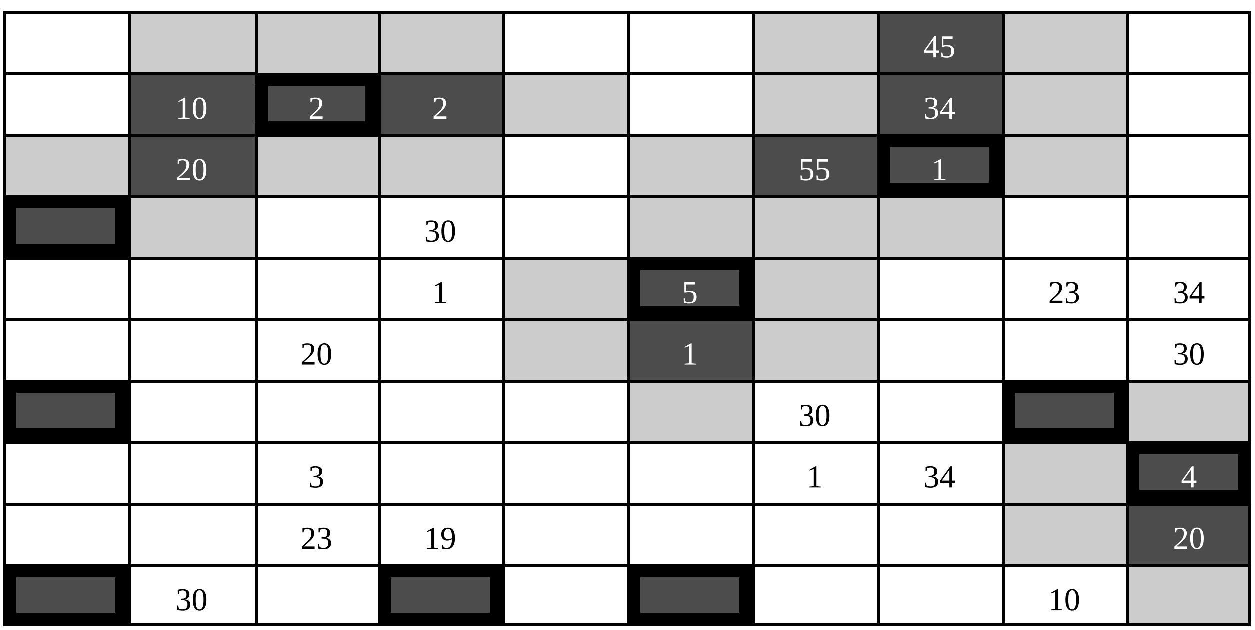
b



a


45
10 2 2 34
20 55 1
30
1 5 23 34
20 1 30
30
3 1 34 4
23 19 20
30 10

b


45
10 2 2 34
20 55 1
30
1 5 23 34
20 1 30
30
3 1 34 4
23 19 20
30 10