Lecture 1

What is an Environment?

- An "environment" refers to the surroundings or conditions in which a person, animal, or object lives, operates, or exists.
- The environment is composed of all living and non-living things that surround us, consisting of four primary physical domains (the Earth's spheres) and dynamic living systems.

The Earth's spheres

- ✓ Earth is an interconnected system divided into four major subsystems, or "spheres".
- ✓ These four spheres—the **geosphere** (land), **hydrosphere** (water), **atmosphere** (air), and **biosphere** (life)—constantly interact to support and sustain the delicate balance of life on our planet

Dynamic living systems

- ✓ Dynamic living systems are open, self-organizing networks of interacting components.
- ✓ Driven by continuous flows of energy, substances, and information, these biological structures maintain stability (homeostasis) while constantly adapting, evolving, and responding to their environment.

Components of the Environment

The Environment is divided into

1. **Abiotic** (non-living),
2. **Biotic** (living), and
3. **Socio-cultural** components.

1. Abiotic Components (Physical Elements)

- These non-living physical and chemical factors provide the foundation for life and are typically divided into three main spheres:

Lithosphere: The solid, outermost crust of the Earth. It includes rocks, soil, and minerals, providing a solid foundation for land-based life and agriculture.

Hydrosphere: All water on Earth, including oceans, rivers, lakes, groundwater, and atmospheric water vapor. It regulates climate and is essential for all living processes.

Atmosphere: The thick, life-sustaining layer of gases (primarily nitrogen and oxygen) that blankets the Earth. It protects the planet from harmful solar radiation and maintains habitable temperatures.

2. Biotic Components (Biological Elements)

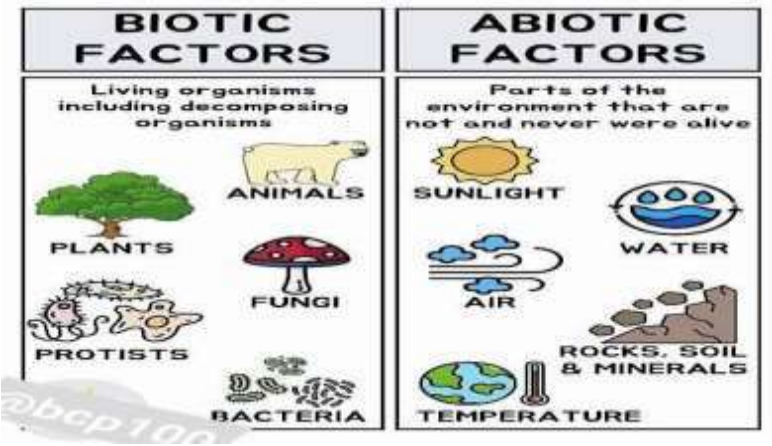
- These are **all the living organisms** that inhabit the Earth and **interact with the physical environment**, commonly referred to as the **Biosphere**. They are grouped by their role in the ecosystem:

Producers (Autotrophs): Organisms that generate their own food through photosynthesis (e.g., plants, algae, and certain bacteria).

Consumers (Heterotrophs): Organisms that rely on producers or other living creatures for food (e.g., animals, birds, humans).

Decomposers: Organisms that break down dead and decaying matter, recycling nutrients back into the ecosystem (e.g., fungi, bacteria).

Decaying matter, or decomposition, is the natural process where dead plants and animals are broken down by **decomposers** (like bacteria, fungi, and insects) into simpler substances. This essential cycle recycles nutrients like carbon and nitrogen back into the soil, acting as nature's ultimate recycling system.



3. Cultural Components (Human-Made)

- Because human activities significantly alter the planet, scientists and sociologists also categorize the environment through human interactions:

  Built Environment: Physical structures created by humans, such as cities, roads, bridges, and agricultural farms.

  Socio-Economic & Political Environment: The societal, cultural, economic, and institutional systems that humans have developed, which influence how we manage natural resources.

  - ✓ Societal influence on the environment encompasses how collective human choices, economic systems, cultural values, and governance models shape ecological outcomes.
  - ✓ Through urbanization, industrialization, and mass consumption, societies significantly alter ecosystems, driving climate change, deforestation, and pollution, while also holding the power to implement sustainable solutions.
  - ✓ Culture shapes how societies perceive, value, and interact with the natural world. Beliefs, traditions, and social norms dictate whether nature is viewed as a resource to be exploited or a sacred system requiring stewardship, ultimately driving patterns of consumption, conservation, and policy-making.
  - ✓ Economic activities significantly influence the environment by extracting natural resources, generating waste, and altering ecosystems to meet human demands. These impacts drive critical global challenges like **climate change**, **biodiversity loss**, and **pollution**. Balancing these activities with sustainability is vital to prevent severe environmental degradation.
  - ✓ Institutional systems—comprising the political, economic, legal, and social frameworks that dictate human behavior—are the primary drivers of environmental outcomes. They shape how natural resources are extracted, regulated, and consumed by defining property rights, market incentives, and the enforcement of ecological protections.

Lecture 2

Collecting environmental data

Environmental Data

- Environmental data include information on **air quality**, **temperature**, **humidity**, **pollution levels**, **noise levels**, and other relevant parameters in an individual's surroundings.
- **Continuous monitoring** of the environment for **changes** in air quality, temperature, humidity, pollution levels, noise levels, industrial by-products, and other relevant parameters within an individual's surroundings is essential.
- It is essential **in smart healthcare** by delivering insights into
  - ✓ the **influence** of environmental elements **on individual health and well-being**,
  - ✓ allowing healthcare professionals to **provide specific recommendations and interventions** that improve patient outcomes.
- This process can take the form of **regular sampling of a fixed set of sites**, often arranged roughly on a grid or network.
  - ✓ **Network sampling is relevant** to this problem.
- Factors to consider while **establishing** an environmental change network consist of **ensuring a regular supply of data** on both the physical as well as possible social environments.
- For example, we have to decide
  - ✓ where **best to locate** the network sites: **spatial regularity** (predictable, organized, or equally-spaced distribution) is naturally appealing but may not yield the **statistically most informative allocation.**
- Furthermore, if economies are sought, **guidelines** must be applied
  - ✓ to **determine** which sites to **eliminate and**
  - ✓ where **new** sites should be **added**.
- Environmental data **encompasses information** about
  - ✓ the natural world and
  - ✓ how humans interact with it.
- Examples include climate data like
  - **temperature** and **rainfall**,
  - **air and water quality** measurements,
  - **biodiversity data** like species populations, and
  - data on natural resources like soil and water.

- It also includes data on **human activities** that impact the environment, such as **pollution levels, deforestation rates, and energy consumption**.

➢ Here's a more detailed breakdown:

1. Meteorological Data:

➢ **Temperature:** Daily, monthly, or yearly average, maximum, and minimum temperatures.
➢ **Precipitation:** Amount of rainfall, snowfall, or other forms of precipitation.
➢ **Wind Speed and Direction:** Data on wind patterns and strength.
➢ **Atmospheric Pressure:** Measurements of air pressure.
➢ **Humidity:** Levels of moisture in the air.

2. Air Quality Data:

✓ **Pollutant Levels:** Concentrations of pollutants like particulate matter ($PM_{2.5}$, PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO).

➢ **Air Quality Index (AQI):** A standardized index indicating the overall air quality.

3. Water Quality Data:

➢ **pH Levels (Potential of Hydrogen):** Acidity or alkalinity of water. pH measures the acidity or alkalinity of a water-based solution on a scale from 0 to 14. Values below 7 are acidic, values above 7 are basic (alkaline), and 7 is perfectly neutral.
➢ **Dissolved Oxygen:** Amount of oxygen in the water, important for aquatic life.
➢ **Nutrient Levels:** Concentrations of nitrogen, phosphorus, and other nutrients.
➢ **Presence of Pollutants:** Detection of **heavy metals** in water (toxic contaminants like **lead, arsenic, and mercury** that originate from industrial discharge mining, or plumbing corrosion), **pesticides**, and other **harmful substances.**

4. Biodiversity Data:

➢ **Species Presence and Abundance:** Counts of different species in a given area.
➢ **Population Trends:** Changes in population sizes over time.
➢ **Habitat Data:** Information about the types of habitats available and their condition.

5. Land and Soil Data:

- **Soil Composition:** Types of soil (clay, sand, loam) and their properties.
- **Soil Moisture:** Amount of water held in the soil.
- **Land Cover:** Information on different land uses, such as forests, agricultural land, urban areas, etc.
- **Deforestation Rates:** Changes in forest cover over time.

6. Natural Resources Data:

- **Water Resources:** Data on surface water and groundwater availability.
- **Energy Resources:** Information on sources of energy (renewable and non-renewable).

7. Socioeconomic Data Related to the Environment:

- **Population Density:** Number of people living in a specific area.
- **Poverty Rates:** Economic conditions of communities and their potential impact on environmental practices.
- **Access to Resources:** Availability of clean water, sanitation, and other essential resources.
- **Environmental Health Data:** Information on diseases related to environmental factors.

8. Environmental Impact Data:

- **Pollution Levels:** Data on air, water, and soil pollution.
- **Waste Generation:** Amounts of different types of waste generated and their disposal methods.
- **Energy Consumption:** Data on energy usage by households, industries, and transportation.
- **Noise Pollution:** Measurements of noise levels in different environments.

Lecture 3

Inaccessible Data

- Organizations and researchers **commonly encounter difficulties** when trying to **access, use, or share** important data.
  - ✓ This phenomenon is referred to as data **inaccessibility.**
- Data inaccessibility occurs when users or organizations cannot **retrieve, utilize, or share critical information**.
- It is typically caused by
  - ✓ technical failures,
  - ✓ intentional security measures,
  - ✓ human error, or
  - ✓ organizational fragmentation.
- This problem **appears** in many ways, **creating obstacles** that **block** the easy exchange of information.
  - **Hardware & Infrastructure Failures:** Crashed servers, failed hard drives, or network outages make physical and cloud-hosted data unreachable.
  - **Security & Privacy Restrictions:** Strict cybersecurity protocols, strict governance mandates, and data privacy regulations often lock down data tightly to prevent breaches/attacks.
  - **Software Issues & Corruption:** Bugs, incompatible formats, and corrupt file systems or databases render stored information unreadable by both humans and software.
  - **Data Silos & Fragmented Systems:** Incompatible software ecosystems and decentralized storage isolate information in specific departments, obstructing cross-functional collaboration and search.
  - **Human Error:** Accidental deletions, misconfigured access permissions, and forgotten passwords are among the most common reasons data cannot be accessed.
  - **Malware & Ransomware:** Cyberattacks that delete, modify, or heavily encrypt data prevent legitimate users from accessing their own systems.

Sensitive Data

- **Sensitive data** refers to information that must remain **confidential** and **not available to the public.**
- Exposing it puts **at risk the individual** and may result **in harassment, harm, identity theft**, and similar consequences.

- This type of data includes
    - **personally identifiable information** (PII),
    - **business data**, and
    - **sensitive category** data.
  - ✓ Personally Identifiable Information (PII) is any data that can be used to **identify, locate, or contact an individual**, either on its own or when combined with other data.
  - ✓ It is a critical concept **used in cybersecurity and data privacy** to protect consumers from identity theft and fraud.
- When an **individual's personal data is exposed**, it creates severe risks including
    - **identity theft**,
    - **financial fraud**, and
    - **emotional distress**.
- **Stolen Personally Identifiable Information** (PII)—like passwords, medical records, or banking details—can be used by bad actors to **steal money or permanently damage a person's credit score.**

The adverse effects of data exposure often present in many fundamental forms:

- **Financial and Identity Theft:** Hackers utilize exposed databases or dark web marketplaces to open credit cards, take out loans, or drain bank accounts in the victim's name.
- **Targeted Phishing:** Exposing details like phone numbers, addresses, or workplace information allows scammers to craft highly convincing social engineering and phishing attacks.
- **Loss of Confidentiality:** Leaks of special category data (such as health conditions, political opinions, or legal history) can lead to public stigma, discrimination, or targeted harassment.

Encountered Data

- We must always consider **how to obtain environmental data** and remain aware for **potential bias**.
- Let **us start** with such problems of **access** and of **sources of bias**.
- Data are the **essential raw material** of any environmental study.
- To investigate an environmental issue, we need **objective observations** of what is going on.
  - ✓ *The objective observer will seek to record simply what they see without offering any opinion.*

- ✓ *In this type of observation results should be the same among individuals*.

- The **classical methods of survey sampling** can be important.
- The long-established methods of **designed experiments** can also be valuable.
  - ✓ But such approaches **need observations** to be taken **at random** and under prescribed circumstances.
- **Addressing environmental problems** frequently **limits** complete solutions;
  - ✓ we must **rely on the available forms** and
  - ✓ confined observations that **occur at certain times and locations**.
- **Climatological variables** recorded
  - across time, particularly in the past,
  - remain limited to data gathered by meteorological stations;
  - floods and tornadoes appear at their own occurrence.
- **Pollution levels** are
  - often **measured and reported** irregularly and selectively, typically **during site visits**, possibly **due to concerns** that levels may **have risen significantly.**
- **Accessibility** is a key factor; **measurements may only be carried out** when
  - permitted,
  - within financial constraints,
  - when equipment is readily available, and
  - when they have been previously recorded, and so on.
- A number of data **fit to the characterization** of **encountered data**:
  - they describe what is encountered - 'the investigator enters the field, observes, and records what he observes . . . what he encounters'.
- It is important to recognize that **encounter sampling cannot be considered a statistical sampling technique**;
  - by the inherent nature of encountered data,
  - a **formal and statistically interpretable** approach to data collection or analysis is unattainable.
- A **more limited concept called 'encountered data'** emerges when we **utilize** a **specified statistical sampling method**, which inherently **restricts or distorts** the scope or relevance of the data.
- **Encounter sampling** is also used to **describe survey sampling** where individuals are chosen by **casual encounter** as an interviewer traverses a **random route** through a survey area.
  - For **example**, with **transect sampling** we traverse a 'transect line' observing the subject of interest as we move along. Clearly, we are more

likely to **see subjects close to the line** but, on the other hand, we may have **frightened them off before we see them**.

- Again, if we **sample fish in a pond** by catching them **in a net**, there will be another form of 'encounter bias' (more usually called size bias). This is because the mesh size reduces the probability of catching smaller fish, since some will slip through the net.

LENGTH-BIASED OR SIZE-BIASED SAMPLING AND WEIGHTED DISTRIBUTIONS

- Every simple **step in statistical inference** is guided by some kind of **assumptions** whose existence are **essential for a valid inferential statement.**
- Let us consider the **sort of difficulties** which have to be faced in the problem of **sample fish in a pond by catching them in a net**.
- Related conceptual difficulties can arise **if we were to sample harmful industrial fibres** (in monitoring adverse health effects) by examining fibres on a plane sticky surface by **line–intercept methods.**
- In the **two cases** our data would consists of
  - ✓ the **sizes of fish that are caught in the net, and**
  - ✓ **the lengths of fibres** crossed by the intercept line, respectively.

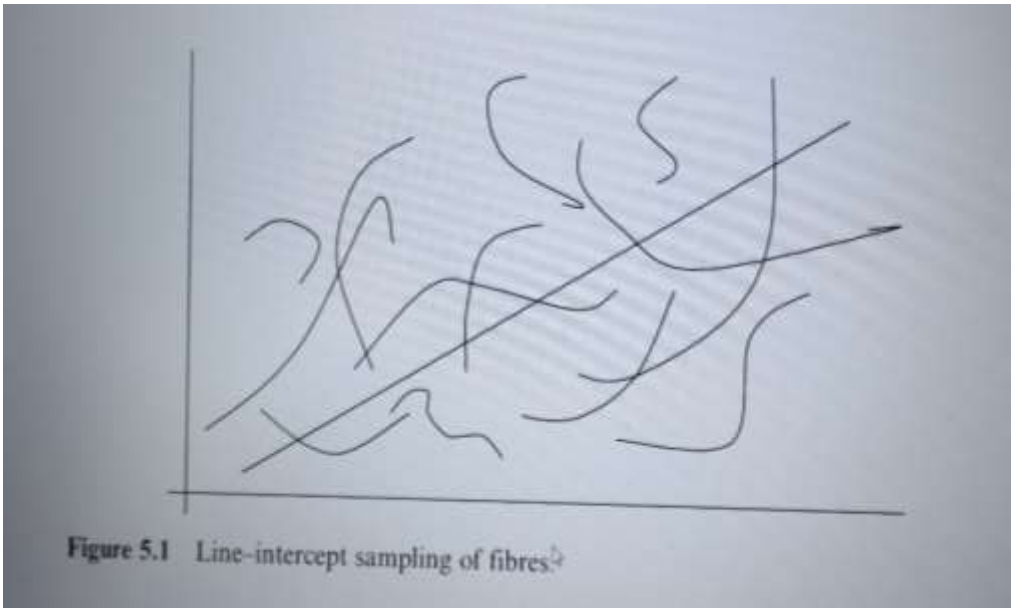


Figure 5.1 Line-intercept sampling of fibres

- One **interest will be in the distribution of sizes** (of the fish or of the fibres), but the **classical sampling methods** are clearly likely to produce **seriously biased results**, since almost **all small fish will slip through the net** – and the **longer the fibre the larger the probability of it crossing** the intercept line, that is, of our sampling it.
- So, we are bound to **obtain what are known as length-biased or size-biased** samples, and statistical **inferences drawn from such samples will be seriously flawed** because **they relate to the distribution of measured sizes** and not to the population at large, which is our real interest.
- Thus, we will typically **overestimate the mean both in the fish and in the fibre examples,** possibly to a serious extent.
- The typical relationship **between populations of measured values and of values at large** is as shown in Figure 5.2.

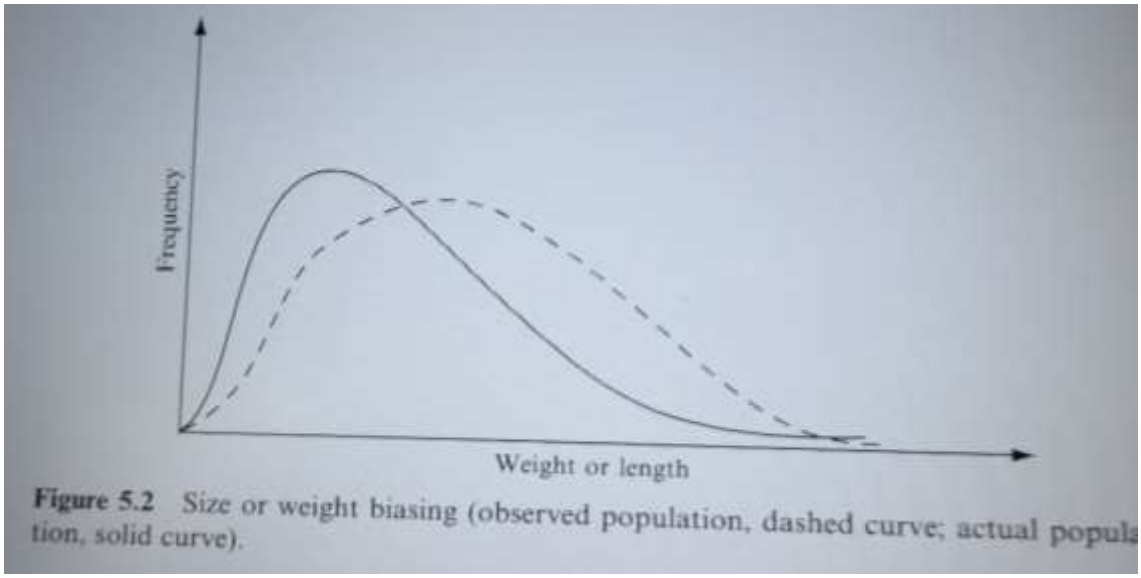


Figure 5.2 Size or weight biasing (observed population, dashed curve; actual popula tion, solid curve).

Weighted distribution methods

- Rao introduced **weighted distribution methods** which can help to **remove such length and size biases**.
- Such methods **also have wider application** to forms of sample bias not just based on observation of length or size.
- Suppose X is a non-negative random variable with p.d.f. $f_\theta(x)$, but an observation x drawn from this distribution is only retained in the collected

sample in accordance with a relative retention distribution having p.d.f. $g_\beta(x)$. Thus, what we actually sample is a random variable X^* with a distribution having p.d.f.

➢

$$f^*_{\theta,\beta}(x) = k f_\theta(x) g_\beta(x), \qquad (5.1)$$

where k is the normalizing constant satisfying

$$k^{-1} = \int f_\theta(x) g_\beta(x) dx.$$

X^* is said to be a *weighted version* of X with *weight function* $g_\beta(x)$. If we know the form of $g_\beta(x)$ in a given situation, we can try to adjust for its effect.

In the special case where $g_\beta(x) \propto x$, X^* follows what is called the sized-biased distribution with p.d.f.

$$f^*_\theta(x) = x f_\theta(x)/\mu, \qquad 5.2$$

where $\mu = E(X)$. The variable actually sampled has expected value

$$E(X^*) = \int [x^2 f_\theta(x)/\mu] dx = \mu(1 + \sigma^2/\mu^2). \qquad 5.3$$

where $\sigma^2 = Var(X)$. So if we take a random sample of size n, $\bar{x}^*$ (the sample mean of the observed data) is biased upward by a factor$1 + \sigma^2/\mu^2$. This could be vast in its effect. For example, if $\sigma^2/\mu^2 = 9$ (a coefficient of variation of just 3) we find that $E(X^*) = 10\mu$ rather than $E(X^*) = \mu$ as we would hope!

But we have another problem: σ^2 and μ are typically unknown, so we have no direct way of correcting the bias. If we knew μ we would not be estimating it; if we knew σ^2 we still could not make proper allowance for the bias of $\bar{x}^*$ in view of (5.3).

A different approach shows an interesting effect. Suppose we consider the reciprocal of the random variable X*. It is interesting to note from (5.2) that

$$E(1/X^*) = \int [f_\theta(x)/\mu]dx = \frac{1}{\mu},$$

so that in this size-biased case (but certainly not generally) its mean is the reciprocal of the mean of X.

This has implications for inference about μ. It is clear that the sample mean reciprocal value $(\sum_{i=1}^{n} 1/x_i^*)/n$ is unbiased for $1/\mu$. More importantly, however, since

$$E(X^*)E(1/X^*) = 1 + \frac{\sigma^2}{\mu^2}, \qquad 5.5$$

we note that $g = \bar{x}^*(\sum_{i=1}^{n} 1/x_i^*)/n$ provides an intuitively appealing estimate of the bias factor $1 + \frac{\sigma^2}{\mu^2}$ without the need to know μ or σ^2. Thus, we can seek, at least informally, to compensate for the bias in (5.3) by using the estimator

$$\bar{x}^*/g = n(\sum_{i=1}^{n} 1/x_i^*)^{-1}.$$

Let us consider a more specific situation. Suppose now that X has a Poisson distribution with mean μ, that is $X \sim P(\mu)$, and again we obtain observations subject to size-biased sampling. Then the distribution of X* has probability function $q(x^*)$ proportional to $xp(x)$ where $p(x) = \frac{e^{-\mu}\mu^x}{x!}$ for x = 0, 1, 2,……

Thus

$$q(x^*) \propto \frac{e^{-\mu}\mu^{x^*}}{[\mu(x-1)!]} = \frac{e^{-\mu}\mu^{x^*-1}}{(x-1)!},$$

so that $X^* - 1$ has a Poisson distribution with mean μ.

So, size-biased sampling here has a straightforward and interpretable effect. The observed outcomes X^* have the distributional form X + 1 where $X \sim P(\mu)$. (Clearly X^* cannot take the value 0 since a population member of value x = 0 will be suppressed totally by the size-biasing process.) Thus, we have an un biased estimator of μ in the form $x^* - 1$.

We confirm this also from examination of (5.3). Since σ^2/μ=1 for the Poisson distribution $P(\mu)$, (5.3) shows that $\bar{x}^*$ has expectation $1+\mu$ and hence that $\bar{x}^* - 1$ is unbiased for μ.

Lecture 4

Ranked-set Sampling

- Ranked-set sampling is becoming widely used for sampling in such contexts, where **measuring environmental risk factors is expensive**.
- It could operate here in the following way.
- If we want a sample of size 5 we would choose five sites at random, but **rather than measuring pollution at each of them we would ask a local expert** which would be likely to give the **largest value**. (Alternatively, we may use a **cheaply observable** concomitant variable, like the **water's opacity**, to select the candidate with the highest value.)
- We then repeat the process by **selecting a second random set of five sites** (and a second expert to guide us) and **seeking to measure the second largest pollution level** amongst these, and so on, until we seek the lowest pollution level in the final random set of five sites.
- The resulting ranked-set sample of size 5 is then used for the **estimation of the mean**.

- Such an approach can also be used to estimate a **measure of dispersion, a quantile or even to carry out a test of significance or to fit a regression model.**
- The **gains can be dramatic**: efficiencies relative to simple random sampling **may reach 300%.**

- The ranked-set sampling approach may be described in the following way.
  - ✓ We first consider a set of **n conceptual random samples**, of observations of a random variable X.

- These would yield observations, say, of the form

$$\begin{matrix} x_{11}, x_{12}, \ldots, x_{1n} \\ x_{21}, x_{22}, \ldots, x_{2n} \\ \vdots \\ x_{n1}, x_{n2}, \ldots, x_{nn} \end{matrix}$$

  if we were to actually measure the outcomes.
- Instead, each subsample yields only one measured observation $x_{i(i)}$: the ith ordered value in the ith sample (i = 1, 2, . . . , n).

- The ranked-set sample is then defined as

$$X_{1(1)}, X_{2(2)}, \ldots, X_{n(n)}:$$

- The studies of ranked-set sampling proposed that the mean μ of the underlying distribution should be estimated by

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i)}$$

which is, for obvious reasons, known as the ranked-set sample mean.

Ranked-set sample mean and variance

- In forming the ranked-set sample, we assume that we **have accurately arranged** the potential observations inside each conceptual sub-sample. This is referred to as perfect ordering.
- If **ordering is not entirely correct**, then, benefits still arise **unbiasedness and increased efficiency** compared with $\bar{X}$ provided there is **non-zero correlation** between the observed sample values and their claimed order.
- The **advantage disappears** once the ranking becomes equivalent to mere randomization.

- Suppose we adopt a **more structured approach** and assume that observations come from a distribution of random variable ***X*** with distribution function in the form $F[(x-\mu)/\sigma]$, having mean μ_x and variance σ_x^2.
- We know that $\bar{X}$ **is unbiased**, with expected value μ_x and variance σ_x^2/n.
- If it is symmetric we have the special case where $\mu_x = \mu$ and $\sigma_x^2 = \sigma^2$.

Consider $\bar{\bar{X}}$.

- The ranked-set sample values $x_{1(1)}, x_{2(2)}, \ldots\ldots, x_{n(n)}$ are the values of the order statistics $X_{(1)}, X_{(2)}, \ldots\ldots, X_{(n)}$ in a sample of size n except

that, because each comes from separate sample, they are independent (rather than correlated as is usually the case).

- Consider the reduced order statistics

$$U_{(i)} = \left(X_{(i)} - \mu\right)/\sigma$$

with means and variances denoted by α_i *and* v_i respectively (i = 1, 2,…..,n), which are independent of μ and σ.

Then

$$E(X_{(i)}) = \mu + \sigma\alpha_i$$

and

$$Var(X_{(i)}) = \sigma^2 v_i$$

Thus

$$E(\overline{\overline{X}}) = \mu + \sigma\overline{\alpha}$$

and

$$Var\left(\overline{\overline{X}}\right) = \left(\sum v_i\right)\sigma^2/n^2,$$

where $\overline{\alpha} = \left(\sum \alpha_i\right)/n.$

- Clearly, we have the simple relationship,

$$\sum X_i = \sum X_{(i)},$$

so that by taking expectations of each side we have

$$n\mu_x = E\left(\sum X_i\right) = E\left(\sum X_{(i)}\right) = n\mu + \sigma\sum\alpha_i$$

Thus

$$\mu_x = \mu + \sigma\overline{\alpha}$$

Which shows that $\overline{\overline{X}}$ is unbiased for μ_x.

Now consider the identity

$$\left[\sum (X_i - \mu)\right]^2 = \left[\sum (X_{(i)} - \mu)\right]^2$$

Taking expectations, we now obtain

$$n\sigma_x^2 = E\left\{\left[\sum\left(X_{(i)} - \mu - \sigma\alpha_i + \sigma\alpha_i\right)\right]^2\right\}$$
$$= E\left[\sum\left(X_{(i)} - \mu - \sigma\alpha_i\right)^2\right] + \sigma^2\sum\alpha_i^2$$

So that

$$\frac{\sigma_x^2}{n} = \frac{\sigma^2}{n^2}\Sigma v_i + \frac{\sigma^2}{n^2}\Sigma\alpha_i^2 \quad \text{.....(1)}$$

That is,

$$Var(\bar{X}) = Var(\bar{\bar{X}}) + \frac{\sigma^2}{n^2}\Sigma\alpha_i^2$$

And the essentially positive nature of the term $(\sigma^2/n^2)\Sigma\alpha_i^2$ shows, as required, that

$$Var(\bar{X}) \geq Var(\bar{\bar{X}}).$$

The relative efficiency of $\bar{\bar{X}}$ *and* $\bar{X}$ is thus

$$e = \frac{Var(\bar{X})}{Var(\bar{\bar{X}})} = 1 + \frac{\Sigma\alpha_i^2}{\Sigma v_i}$$

If X is symmetric $\sigma_x^2 = \sigma^2$, so that from (1) we have $n = \Sigma v_i + \Sigma\alpha_i^2$ and the relative efficiency can be written in terms of the α_i as

$$e = \left(1 - \Sigma\alpha_i^2/n\right)^{-1}$$

It is possible to show that $1 < e \leq (n+1)/2$, so that the potential efficiency gains from using the ranked-set sample mean can be very high.

Lecture 5

Line Transects and Variable Circular Plots

- Estimation of **abundance and density** of animals and plants population is essential for **conservation and management in ecology**.
- The study **introduces** the methods of **line transect sampling**; a type of **distance sampling method** used in ecology.
- **Line transect sampling** was first developed by R.T. Kings **to estimate abundance of animals and plants**.

Sampling Technique

- In a **line transect survey of an animal or plant species**, an observer **moves along a selected line** and **notes the location relative to the line** of every individual of the species detected.
- Line transects methods have been **used for many types of populations**, including
  - ✓ **bird,**
  - ✓ **mammal, and**
  - ✓ **plant species**
  - ✓ as well as **other objects** for which **detectability depends on location relative to the observer.**
- **For surveys** of some species, the **observer walks along the transect**.
- Line transects methods have also been **applied to aerial surveys, surveys from research vessels, and sighting of animals from automobiles**.

- The statistical sampling methods are inevitably **based on assumptions** which are major **simplifications of the real-life circumstances**.

- For example, the **search objects** are assumed (in the simplest applications):

  - ✓ to be stationary (i.e. not moving);
  - ✓ to be similarly and independently able to be observed;
  - ✓ to be seen at right angles to the transect line;
  - ✓ to be seen on one occasion only,
  - ✓ to be unaffected (e.g. neither repelled nor attracted) by the observer.

Limitations

- It typically occurs in such surveys that
  - ✓ **more individuals are detected close to the line than far away from it**, not because
  - ✓ abundance is higher near the line but because the **probability of detection is higher** near the line than far from it.
- A line transect is characterized by a **detectability function** giving the probability that an animal (or plant) at a given location is detected.
  - ✓ In most situations, the **probability of detection can be expected to decrease** as distance from the transect line increases.
  - ✓ In many cases, **detectability on the line** itself can be assumed **perfect**.
  - ✓ In other cases, **avoidance by the animals** of the observer can result **in detectability reaching a maximum** at some distance from the line.
- To estimate the **abundance or density of the species** in the study area from one or more such transects, **this nonconstant detectability** must be taken into account.

Density Estimation Methods for Line Transects

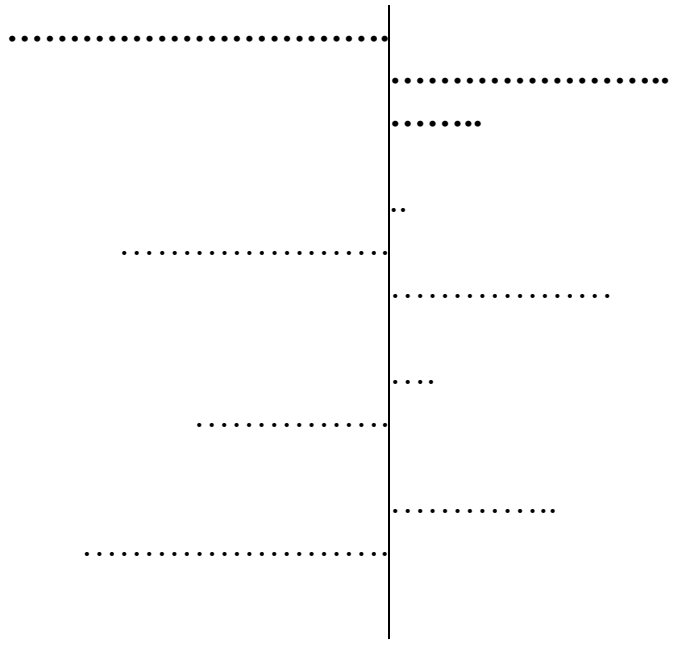


Figure: Observations of animals or other objects from a transect line.

- Figure 1 depicts **observations of animals** or other objects from a **segment of a line transect.**

- The **perpendicular distances** from the objects to the transect are indicated with **dashes lines**.
- **Given the set of distances** of observed animals **from one or more transect lines**, for which detectability is virtually perfect near the line but decreases with distance from the line,
  - ✓ **it may not be immediately apparent how to estimate the abundance or density** of the animals in the population.

- In the methods that follow, the **object is to estimate the density of animals or other objects in a study region of area A**.
- **For the *ith* transect** in the sample, the variable of interest y_i **is the number of animals observed.**
- The **sample size n** refers to the number of **transects selected** (not the variable of interest).
- The **total number of animals** in the study region is **denoted** τ , and the **density** of animals is D = τ / A .

Narrow-Strip Method

- Although the **detectability of animals far away from the transect line may be imperfect**, there may be **some narrow strip** along the line in which **detectability is virtually perfect**.
- Then by **using only those observations** within the strip and **ignoring the more distant observations**, one may **consider the strip** a **conventional unit or plot** and estimate the population total or density in the usual way.

- Let **L denote** the **length** of the transect and let w_0 be the **maximum distance** from the line to which **detectability is assumed perfect**.
- Then the **width** of the strip is $\boldsymbol{2w_0}$ and its **area** is $\boldsymbol{2w_0L}$
- Let $\boldsymbol{y_0}$ be the **number of animals** detected within the narrow strip.
- To estimate the **density D**, that is, the number of animals **per unit area**, one may use the number of animals in the narrow strip divided by its area:

$$\widehat{\boldsymbol{D}} = \frac{\boldsymbol{y_0}}{\boldsymbol{2w_0L}}$$

- If the study region has area A, the **total number of animals** in the study region is estimated as

$$\hat{\tau} = A\hat{D} = \frac{Ay_0}{2w_0L}$$

- The distance w_0 is generally **smaller than the maximum distance** to which animals have been detected, and hence the number of animals y_0 used to estimate density is generally **fewer than total number y detected**.

Problem: On a line transect of length L = 100 meters, a total of y = 18 birds were detected at the following distances (in meters) from the transect line: 0, 0,1, 3, 7, 11, 11, 12, 15, 15, 18, 19, 21, 23, 28, 33, 34, 44. Estimate the density of birds in the study region.

Solution: The frequency distribution is

Class boundary	0-10	10-20	20-30	30-40	40-50
Frequency	5	7	3	2	1

Plotting the numbers of birds detected in each 10-meter distance interval, we find that 5 were seen within 10 meters of the line, 7 were seen between 10 and 20 meters, 3 between 20 and 30 meters, 2 between 30 and 40 meters, and 1 between 40 and 50 meters. Choosing $w_0 = 20$ as the distance beyond which sightings drop off markedly, this narrow strip was $y_0 = 12$.

The estimate of population density is

$$\hat{D} = \frac{y_0}{2w_0L} = \frac{12}{2(20)(100)} = 0.003$$

Smooth-By-Eye Method

- In making a histogram to proximate a probability or probability density function f, one first choose an interval width and then sets the **height** $\hat{f}$ of the histogram for a given distance x by the following formula:

$$\hat{f}(x) = \frac{(number\ of\ observations\ in\ the\ \text{int}erval\ containing\ x)}{(total\ number\ of\ observations)(Interval\ width)}$$

- Note that, in keeping with its **probability** interpretation, **the area under the histogram adds to one.**

- The **histogram for distance *x*** from the transects line may be viewed as approximating a **smooth probability density function f(x)** that would describe the **distribution of detection distances** that would be obtained if one ran an **infinite number of randomly selected transect lines** from the species in question.

NONPARAMETRIC METHODS

- To **avoid assumptions about the shape of the unknown detectability functions,** non-parametric density estimation methods can be used.
- Using observations of random variables from a probability density function *f*, the **methods use smoothing techniques** to estimate the value $f(x)$ of the density function at any given value of x.
- With line transect sampling, **the probability density function of interest** is the density of observed detection distances.
- **Detectability on the transect line is assumed perfect**, so that the estimate has the form

$$\widehat{\boldsymbol{D}} = \frac{\boldsymbol{y\hat{f}(0)}}{\boldsymbol{2L}}.$$

Estimating *f* (0) by the Kernel Method

- In the **extensive statistical literature on the estimation of probability density functions (PDFs)**, the dominant trend is **kernel estimation**, a nonparametric smoothing approach.
- The method employs a ***kernel function K*(*x*),** which integrates to 1; that is,

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

- The kernel estimator of the PDF *f* at *x* is

$$\hat{f}(x) = \frac{2}{yh}\sum_{j=1}^{y} K\left(\frac{x - x_j}{h}\right)$$

  where *h* is called the *window width* and x_j is the value of the *i*th observation (i.e., the distance from the transect line to the *j* th animal) and *y* is the number of observations (i.e., the number of animals detected).
- The coefficient 2 arises when the density of unsigned distance, without regard to which side of the line, is used.

To estimate $f(0)$ with a symmetric kernel, the estimator becomes

$$\hat{f}(0) = \frac{2}{yh}\sum_{j=1}^{y} K\left(\frac{x_j}{h}\right)$$

With the normal kernel, for example,

$$K\left(\frac{x_j}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}\right)\left(\frac{x_j}{h}\right)^2}$$

Silverman gives a simple rule for choosing the window width h:

$$h = 0.9ay^{-1/5}$$

where $a = \min(s, Q/1.34)$, in which s is the sample standard deviation of the x's observed and Q is their interquartile range.

But when dealing with positive distances only, one should use the median distance in place of the interquartile range.

The estimator of the population density of animals is then

$$\widehat{D} = \frac{y\hat{f}(0)}{2L}$$

Problem: On a line transect of length L = 100 meters, a total of y = 18 birds were detected at the following distances (in meters) from the transect line: 0, 0,1, 3, 7, 11, 11, 12, 15, **15**, 18, 19, 21, 23, 28, 33, 34, 44. Estimate the density of birds in the study region.

Solution: Here

median absolute distance =15/1.34= 11.19 and

s =12.56.

Since 11.19 is less than the sample standard deviation s = 12.56, Silverman's rule for choosing window width h gives

$$h = 0.9(11.19)(18)^{-\frac{1}{5}} = 5.65$$

The normal kernel estimate of $f(0)$ is

$$\hat{f}(0) = \frac{2}{18(5.65)\sqrt{2\pi}}\left[e^{-0^2/2(5.65)^2+\cdots\ldots\ldots\ldots+e^{-44^2/2(5.65)^2}}\right]$$
$$= 0.0376$$

The estimate of bird density is

$$\widehat{D} = \frac{y\hat{f}(0)}{2L} = \frac{18(0.0376)}{2(100)} = 0.0034$$

bird per square meter, or 34 birds per hectare.

Fourier Series Method

The Fourier series method of estimating *f(0)* is

$$\hat{f}(O) = \frac{1}{w^*} + \sum_{k=1}^{m} \hat{A}_k$$

where w* is the maximum distance at which animals can be observed and the coefficients $\hat{A}_k$ are given by

$$\hat{A}_k = \frac{2}{yw^*}\left[\sum_{i=1}^{y} cos\left(\frac{k\pi x_i}{w^*}\right)\right]$$

The number m of terms to use in the approximation is somewhat arbitrary, but the following rule of thumb has been recommended: Starting with m = 1, choose the first whole number m such that

$$\frac{1}{W^*}\sqrt{\frac{2}{y+1}} \geq \left|\hat{A}_{m+1}\right| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In determining the maximum detectability distance w∗, Burnham et al. (1980) and Crain et al. (1979) recommend using some distance less than the greatest distance at which an animal was detected, throwing out the largest 1%–3% of the observed distances as outliers.

Problem: On a line transect of length L = 100 meters, a total of y = 18 birds were detected at the following distances (in meters) from the transect line: 0, 0,1, 3, 7, 11, 11, 12, 15, 15, 18, 19, 21, 23, 28, 33, 34, 44. Estimate the density of birds in the study region by applying Fourier series method.

Solution: In applying the Fourier series method, the largest observation, the detection at 44 meters, is thrown out as an outlier and the next largest, 34 meters, is used as w∗.

Thus

$$\frac{1}{W^*}\sqrt{\frac{2}{y+1}} = \frac{1}{34}\sqrt{\frac{2}{17+1}} = (.029)\sqrt{.1111} = .0097$$

The inequality of the rule of thumb [Equation (1)] is satisfied for the value m = 1, so only one term is needed. Thus, only one coefficient, $\hat{A}_1$, needs to be computed, but it involves 17 terms (the number of observations after discarding the largest). The coefficient A_1 is calculated from the Equation:

$$\hat{A}_1 = \frac{2}{yw^*}\left[\sum_{i=1}^{y} cos\left(\frac{k\pi x_i}{w^*}\right)\right]$$

$$= \frac{2}{17(34)}\left[\cos\frac{1(3.1417)(0)}{34} + \cdots \ldots \ldots + \cos\frac{1(3.1417)(34)}{34}\right]$$

=0.0091

The estimate of *f(0)* is calculated as

$$\hat{f}(O) = \frac{1}{w^*} + \sum_{k=1}^{m}\hat{A}_k = \frac{1}{34} + 0.0091 = 0.0385$$

The estimate of population density is

$$\widehat{D} = \frac{y\hat{f}(0)}{2L} = \frac{17(0.0385)}{2(100)} = 0.0033$$

Or 33 birds per hectare.

RANDOM SAMPLE OF TRANSECT

A random sample of n transects in the study area will be selected as follows.

- A straight baseline of length B is drawn across (or below) the study region on a map. The study area need not be regular in shape.
- The length of the baseline is the width of the perpendicular projection onto the line of every point in the study area—that is, the width of the shadow cast by the study area onto the line.
- A random sample of n transect locations $v_1, v_2, \ldots\ldots, v_n$ is selected from the uniform distribution on the interval [0, B].
- Transect lines perpendicular to the baseline are then located through each selected point v_i.
- The transects either run completely across the study area, or if a maximum transect length L is desired that is less than the distance across the study region, parallel baselines may be drawn distance L apart and starting points selected from the total length of baseline.
- Note that sampling is with replacement, although since transect locations are selected from a continuous distribution, there is zero probability of selecting precisely the same transect twice.
- In selecting a random sample of transects, some biases may be introduced due to boundary problems—that is, slightly lower average detection probability for an animal near the boundary of the study area.
- Such biases would tend to be small in any case when the study area is large relative to the effective width of transects.

RANDOM SAMPLE OF TRANSECT

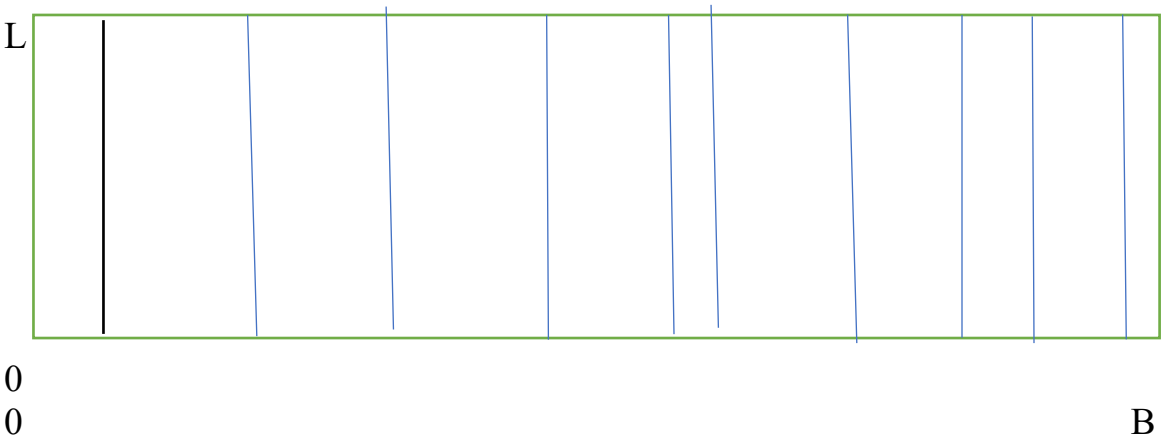


Figure 1. Random sample of 10 line transects in a study region.

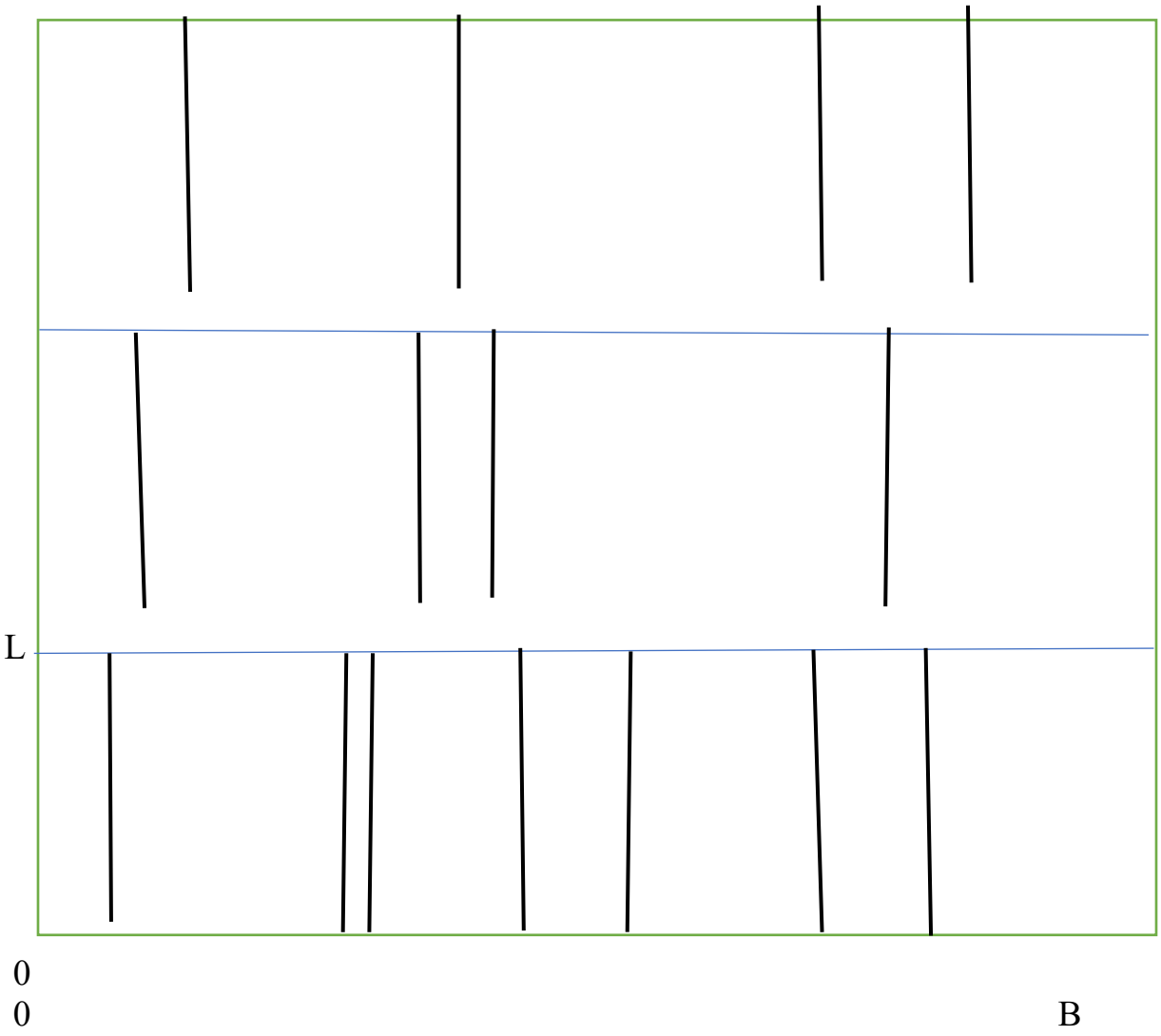


Figure 2. Random sample of 15 line transects of length L in a study region of width wider than L

Unbiased Estimator

Because of irregularities in the shape of the study region, the length L_i of the ith transect is a random variable, with expected value $E(L_i) = A/B$, where A is the area of the study region and B is the length of the baseline. Let y_i denote the number of animals seen from the ith transect. If the effective width w or the density $f(0)$ is known, an unbiased estimator of density, based on the ith transect, is

$$\widehat{D}_i = \frac{B}{A}\left(\frac{y_i}{2w}\right) = \frac{B}{A}\left(\frac{y_i f(0)}{2}\right) = \frac{y_i f(0)}{2E(L)}$$

Hence an unbiased estimator based on the n transects is

$$\widehat{D} = \frac{1}{n}\sum \widehat{D}_i = \frac{B}{A}\left(\frac{\bar{y}}{2w}\right) = \frac{B}{A}\left(\frac{\bar{y}f(0)}{2}\right)$$

When *w* or *f(0)* in the expression for $\widehat{D}$ is estimated, for example, by one of the methods just given, the estimated value $\widehat{w}$ or $\hat{f}(0)$ is substituted in the expression for $\widehat{D}$ and the unbiasedness holds only approximately.

If individual estimates $\widehat{w}_1, \widehat{w}_2, \ldots\ldots\ldots, \widehat{w}_n$ or $\hat{f}_1(0),\ \hat{f}_2(0), \ldots\ldots., \hat{f}_n(0)$ are made independently for each transect and $\widehat{D}_i = \frac{By_i}{2A\widehat{w}_i}$ or $\widehat{D}_i = \frac{By_i\hat{f}_i(0)}{2A}$, then $\widehat{D}_1, \widehat{D}_2, \ldots\ldots., \widehat{D}_n$ are independent and an unbiased estimator of variance is

$$\widehat{var}\left(\widehat{D}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\widehat{D}_i - \widehat{D}\right)^2$$

However, one often gets a better estimate of w or f (0) by pooling the distance data from all transects in the survey. With $\widehat{D}_i = \frac{B}{A}\left(\frac{y_i}{2\widehat{w}}\right)$ or $\widehat{D}_i = \frac{By_i\,\hat{f}(0)}{2A}$ using the pooled estimates, the $\widehat{D}_i$ are not independent and $\widehat{var}\left(\widehat{D}\right)$ tends to underestimate the true variance of the estimator.

With the pooled estimates, a better estimate of variance could be obtained through a resampling method such as the bootstrap or jackknife. For the bootstrap method, the sample of n transects is treated as a population in itself. A bootstrap sample is obtained by selecting n of these transects from the sample at random with replacement, and bootstrap estimate $\widehat{D}^{*1}$ is computed from the bootstrap sample by the same method as used for the original sample. Note that the bootstrap sample may differ from the original sample because of the with-replacement sampling. Repeating the procedure to obtain M independent bootstrap values $\widehat{D}^{*1}, \ldots\ldots\ldots\ldots., \widehat{D}^{*M}$, the bootstrap estimate of variance is

$$\widehat{var}_b\left(\widehat{D}\right) = \frac{1}{(M-1)}\sum_{m=1}^{M}\left(\widehat{D}^{*M} - \widehat{D}_b\right)^2$$

where $\widehat{D}_b = (1/M)\sum_{m=1}^{M}\widehat{D}^{*M}$

The jackknife estimate is obtained by systematically deleting one transect at a time from the sample. Let $\widehat{D}_i$ be the estimate obtained from the n − 1 remaining transects in the sample after deleting the ith transect, and let $\widehat{D}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{D}_i$. Note that for each of the n jackknife samples, each consisting of n − 1 transects, the entire process of making a pooled estimate of *w* or *f(0)* and then estimating density is repeated. The jackknife estimate of variance is

$$\widehat{var}_j(\widehat{D}) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{D}_i - \widehat{D}_{(.)}\right)^2$$

Note that for each method, the resampling is in terms of transects, which are independent because of the initial random selection, not in terms of observed distances, which cannot be assumed independent without assumptions about the population itself.

Approximate 100(1 − α)% confidence intervals based on the jackknife method are usually of the form $\widehat{D} \pm t\sqrt{\widehat{var}_j(\widehat{D})}$, where t is the upper α/2 point of the t distribution with n − 1 degrees of freedom. With the bootstrap method, confidence intervals are typically based on percentiles of the distribution of the bootstrap estimates $\widehat{D}^{*m}$. From any estimate $\widehat{D}$ of population density with estimated variance $\widehat{var}(\widehat{D})$, an estimate of the total abundance in the study region is $\hat{\tau} = A\widehat{D}$ with variance estimate $A^2\widehat{var}(\widehat{D})$.

Lecture 6

Environmental Standards

What is an Environmental Standard?

- Environmental standards are **approximations to unknown efficient levels** of the associated environmental commodities.
- All human activities **generate pollution** which **contaminates** one or more environmental "compartments" (air, water, land, ecosystems).
- When the concentration of a given pollutant in the environment becomes too high, it can cause serious environmental damage and/or health impacts depending on its toxicity.
- In order to **protect both health and the environment**, it becomes necessary to **regulate the amount of the pollutant** that can be released **without causing unacceptable harm** to health or the environment.
- Regulation in this way involves **setting a permissible limit, called a standard**, on the concentration of a harmful or potentially harmful pollutant which can be released into the environment.
- Standards are also **imposed on all kinds of manufactured goods**, processes, consumer goods and services, other commodities.

- As a typical example, the EC Drinking Water Directive specifies that the **maximum permissible concentration of iron in drinking water** shall not exceed 200 μg^{-1}. This limit of 200 μg^{-1} is a "standard" in this case.

- It is to be noted that a single-substance standard such as this is seldom stipulated: usually standards cover a range of pollutants. For example, the aforementioned directive sets a standard (permissible limit) for several properties of drinking water, including acidity, color, turbidity, and iron, manganese, aluminum, and lead content.

- According to the USEPA regulation with what are called "primary" standards is carried out to protect health, "secondary standards" are intended to prevent damage to property and the environment.
- A **geographic area** that meets the **primary standard criteria** (or exceeds them) is called an **"attainment area;"** an area that does not is called a "non-attainment area."

- Clearly **checks should be made** on polluters to determine whether or not they are complying with a standard, and this is normally the **responsibility** of the **regulatory agency concerned**.
- It is also clear that those **who violate a standard must be punished**, and that all **polluters should be encouraged** (through incentives, for example) to keep emission levels **below the set standard**, for it is only then that the best results can be achieved.
- In order to be **effective, standards must** therefore be **set and operated** within a regulatory framework, and **supported by the due processes of law**.

THE STATISTICALLY VERIFIABLE IDEAL STANDARD (SVIS)

- The SVIS in principle **links an ideal standard** (which reflects uncertainty and variation in the population at large) **with a compliance criterion** to be used in practice to seek to **verify the standard** at some prescribed **level of statistical assurance.**
- Typical specific examples of SVIS might require, on the one hand, that
  - ✓ (i) 'the 0.95 quantile of the pollutant distribution must not exceed level L, to be demonstrated with 95% statistical assurance' or, on the other hand, that
  - ✓ (ii) 'the mean of the pollutant distribution must not exceed level M, to be demonstrated with 99% statistical assurance'.
- Characteristically, such standards are expressed **in terms of 'levels'**, that is, **acceptable limit** values for **appropriate summary statistics or parameters** of a pollution or contamination distribution.

- We start with example (ii), where we aim for 99% assurance that the mean pollution level does not exceed a value M. We can express this more specifically as follows.
  - ✓ The expected value μ_X of the distribution of the pollution level X should not exceed M. If this is so, the probability of misclassification as in non-compliance (i.e. of concluding that $\mu_X > M$) should not exceed a specified small value a (0.01 in this case).
- This is readily set up in terms of significance.

- ✓ Suppose, for the moment, that we have $X \sim N(\mu_X, \sigma^2)$ with σ^2 $k$nown, and we have a random sample $x_1, x_2, \ldots, x_n$ of n pollution readings with sample mean $\bar{x}$. Then we need only conduct a level-α test of the hypothesis $H: \mu_X \leq M$ against $\overline{H}: \mu_X > M$, concluding that the standard is violated if

$$\bar{x} > M + \frac{z_\alpha \sigma}{\sqrt{n}} \ldots\ldots\ldots\ldots (1)$$

where z_α is the one-tailed α -point of the standardized normal distribution. For '99% assurance' as specified, we would use a 1% test and (1) becomes

$$\bar{x} > M + \frac{2.576\sigma}{\sqrt{n}}$$

- If σ^2 is not known, as is commonly the case, we would merely replace (1) with the form for the standard t-test.

- It is clear to see, however, that **standards expressed in terms of quantiles** will often be more relevant than those expressed in terms of means; we frequently want to **limit 'extreme' values** which will be found in the tails of the distribution of X.
  - ✓ The most direct **quantile based SVIS is essentially 'non-parametric'** in form in that it does not depend explicitly on the distribution of the pollution level, X.
- Example (i) requires $p = P(X > L) \leq 0.05$. So, if we take sample observations of pollution levels, each will be either greater than L (with probability p) or less than L (with probability 1-p). Thus, if r out of n observations exceed L then r is an observation from a binomial distribution *B(n, p)* and we have only to employ the standard 5% significance test of the hypothesis $H: p \leq 0.05$ to implement the SVIS.

- Of course, we could base the SVIS in example (ii) more specifically on an assumed distribution for X, and we might expect to obtain a more efficient version in this more structured situation. Let us consider how this can be done. We can state the standard as follows:

- ✓ The $1-\gamma$ quantile, $\xi_{1-\gamma}$, of the distribution of the pollution level X should not exceed L. If this is true, the probability of misclassification as in noncompliance (i.e. of concluding that $\xi_{1-\gamma} > L$) should be no more than some specified small value α.

- This again provides a 'level of statistical assurance' of $1-\alpha$, and such a SVIS can, be expressed as a level- α test of significance, this time of the null hypothesis $H: \xi_{1-\gamma} \leq L$ against the alternative hypotheses $\overline{H}: \xi_{1-\gamma} > L$. Thus, using the statistical machinery of significance testing, we would satisfy the requirement for statistical assurance at level $1-\alpha$.
- Barnett and Bown (2002) discuss best linear unbiased quantile estimators based on random samples (and also, indeed, on ranked-set samples) as the basis for carrying out such a SVIS in the case of normally distributed pollution levels.
- Let us briefly examine how the quantile-based SVIS would work in the case of $N(\mu_X, \sigma^2)$ pollution levels X, where σ^2 is assumed known. Notice that the quantile $\xi_{1-\gamma}$ is defined by $P(X < \xi_{1-\gamma}) = 1-\gamma$ so that, in this situation, we have

$$\xi_{1-\gamma} = \mu_X + z_{z\gamma}\sigma.$$

- Thus, the null hypothesis $H: \xi_{1-\gamma} \leq L$ is equivalent to $H: \mu_X \leq L - z_{z\gamma}\sigma$ so that the significance test is again just the standard test of a normal mean.

- Of course, σ^2 is unlikely to be known in practice and no such simple approach will be feasible.
- Barnett and Bown (2002) discuss this issue, comparing different quantile estimators based on moments or order statistics and the significance tests relevant to them; they also consider the implications of using ranked-set sampling rather than random sampling and demonstrate impressive efficiency gains from ranked-set sampling.
- We note that this **significance test assigns the benefit of the doubt** to the **polluter** (i.e. the polluter could meet the standard even with a sample estimate of the quantile that is in excess of L). Such standards could of course be set the other way round, **with the benefit of the doubt assigned to the regulator**. We have only to **interchange the null and alternative hypotheses**.
- Let us examine this benefit-of-the-doubt principle in more detail.

- ✓ There may be good reason for such imbalance; on the one hand, it might have been preserved in original negotiation or, on the other, perhaps it was imposed by legislation.
- ✓ But the imbalance cannot be regarded as entirely satisfactory.
- ✓ It requires one party to operate on different conditions than the other.
- ✓ For example, with the quantile standard, sample levels of the estimated quantile can be well in excess of L before the regulator can claim violation.

- With the hypotheses reversed, the potential 'polluter' will need to demonstrate sample values well below L to be 'in the clear'. Methods for implementing a SVIS when, as is likely, pollution distributions are non-normal have also been considered. Often pollution levels will have positively skewed distributions; the gamma family may provide a reasonable model for such cases. Bown (2002) derives various appropriate forms of SVIS for gamma-distributed pollution levels and discusses their application.

GUARD POINT STANDARDS

- We have already discussed a **crucial point in relation to environmental standard-setting** using hypothesis-testing principles (e.g. as a basis for an SVIS): namely, that there is implicit assignment of '**benefit of the doubt' either to the regulator or, for complementary hypotheses, to the producer of the potential pollution (the 'complier').**

- Let us review this matter in the case where **the aim** of the environmental standard is to **limit the effects of a pollutant** by **imposing a constraint** on the expected value, μ, of the distribution of the pollution level, X.
- We will assume that the **upper level of concern for μ is μ_0 $(= 120)$.**
  - ✓ With the form of SVIS, the test either assumes the default status of the **population to be 'compliance'** and the **regulator** will then need strong evidence (i.e. a high observed mean pollution level) **in order to reject this**, or assumes 'non compliance' ('violation') at the outset and the **complier will need to produce at a level** far below μ_0 in order to escape the risk of classification as 'non-compliant'.
- This leads to inevitable **conflicts of interest** and cannot be regarded as ideal; it seems to imply an **intrinsic unfairness to one party or the other**.

- Barnett and O'Hagan (1997) suggested a basic means of **overcoming this problem**, and Bown (2002) has developed a corresponding **compromise procedure** where **it is not necessary** for one party or the other to take a rigid **'benefit of the doubt' stance**.
- This she describes as a guard point standard, and it operates on the following principles.
  - ✓ Suppose that both parties (the regulator and the complier) are prepared to compromise.
  - ✓ Instead of **setting the single standard level**, $\mu_0 = 120$, the regulator is prepared to set an upper guard point **above $\boldsymbol{\mu_0}$, at $\boldsymbol{\mu_0 = 130}$ say**, provided **there is assurance** that pollution at this level will be detected with probability $1 - \alpha$ (for α small, e.g. $\alpha = .005$).
  - ✓ Correspondingly, the **complier will accept a lower guard point below $\boldsymbol{\mu_0}$, at $\boldsymbol{\mu_0 = 110}$**, say, if assured that when pollution levels are as low as this there will probability $1 - \beta$ (for β small) of correct classification as compliant.

- We suppose that in any application, $\mu_1 < \mu_0 < \mu_2$ and that α and β can be appropriately chosen to be regarded **as 'fair' to both parties**.
  - ✓ Then if $\mu > \mu_2$, the probability of failing the standard is at least $1 - \alpha$, whilst if $\mu < \mu_1$, the probability of satisfying the standard is at least $1 - \beta$.
  - ✓ Mean pollution levels **between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$** are regarded as **relatively unimportant:** the guard point **compromise principle** defines such situations as **'uncertain'** (or 'too close to call').
  - ✓ Thus, there are now **three possible outcomes**: compliance, violation and uncertainty (where more information may need to be sought).

- Bown (2002) suggests the following statistically verifiable ideal guard point standard (SVIGPS), for specified $\mu_1 < \mu_0 < \mu_2$, α and β:
  - ✓ if a population has **$\boldsymbol{\mu}$ in excess of $\boldsymbol{\mu_2}$,** the probability that the population is **declared as non-compliant** should be no less than $1 - \alpha$;
  - ✓ **if $\boldsymbol{\mu}$ does not exceed $\boldsymbol{\mu_1}$,** the probability that the population is declared **as compliant** should be no less than $1 - \beta$.
- Bown (2002) develops a testing procedure for the SVIGPS case of normally distributed X, and adapts it for setting a SVIGPS for a population quantile

rather than mean. It is further developed to apply to an asymmetrically distributed pollutant level, specifically from the gamma family.

STANDARDS ALONG THE CAUSE–EFFECT CHAIN

- Consider a situation in which an industrial plant lies alongside a waterway such as a river.
- When not fully under control, the plant might discharge effluent into the river from where it flows downstream, possibly flooding farmland, running past a school in a neighboring village and flowing eventually into a lake which is home to fish and wildfowl.
- There are various pollution risks and corresponding prospects for standards to be operated.
  - ✓ For example, we might impose a standard on the possible cause of trouble: the level of pollution initially released in the factory discharge (we call this the 'entry stage'). Various so-called 'pollution contact' effects are possible: damage to crops in the flooded field is a hazard, as are skin complaints among the schoolchildren, even death of fish or birds in the lake.
- Thus, we consider the situation typified by Figure 9.1
-

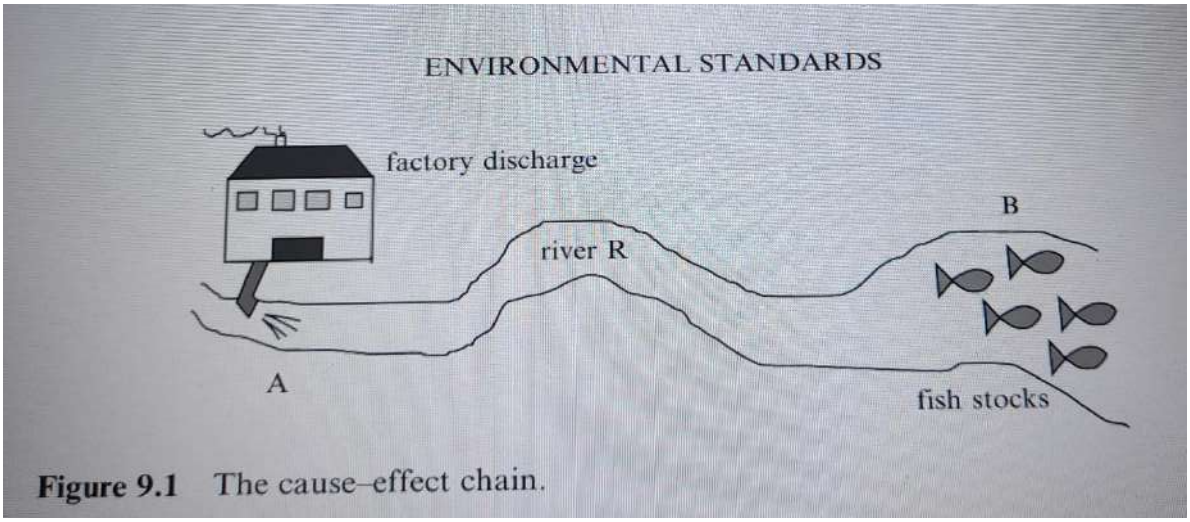


Figure 9.1 The cause–effect chain.

- This is a typical example of the **cause–effect chain of environmental risk**.
- Standards might be set at **various points in the chain**: on the levels of source pollution (at the entry stage), on its levels as it travels downstream, or on its effects (on schoolchildren, on fish or on birds).
- Let us distinguish **two distinct emphases**.

- ✓ The first concerns **pollution levels** (the causes), whether at entry or in contact (possibly later) with vulnerable subjects. The methods described above using the SVIS form of standard are relevant to this interest.
- ✓ Note, however, that we now have a new concern.
- ✓ Surely any standard set at the entry stage should be **mutually compatible** with any standard set at a contact stage: it should represent the same or equivalent levels of concern and control.
- ✓ So we need **harmonization of standards over distance (space) and time**.
- ✓ The **second emphasis concerns effects**, such as the probability that a particular form of fish is killed– again possibly at different distances from the industrial plant.
- ✓ So once more we need to **strive for harmonization over distance and time**, but now we have the added difficulty of determining what form of standard it is appropriate to apply to the probability of death of a fish.

- ➢ We might need to **develop such standards in either or both of the absolute and conditional forms**, for example for the '**marginal probability of death'** or for the '**probability of death when we have pollution level X**= x'.
- ➢ Such standards will be different in form from those described earlier, although the principle of the SVIS will still be appropriate.

- ➢ We have explained the need for **harmonization** (mutual consistency) of standards **at different points** (in distance and time) of the cause–effect chain.
- ➢ But a new consideration now enters.
  - ✓ We also need **to harmonize** (to seek mutual consistency for) standards **set on different functions**, for example, on pollution level (a cause) or on 'probability of death' (an effect).
  - ✓ In general terms, what we are saying is that any requirements we might set on the levels of discharge at the factory must surely be in harmony with requirements on resulting levels downstream and with what we require of the protection of crops, children, fish or fowl.

- ➢ Let us **formalize some of these ideas**.
  - ✓ Referring to Figure 9.1, suppose that pollution levels at the factory (A) are represented by a variable X; those at a fishery (B) by a random variable Y, distributed differently from X but induced by X.

- ✓ Setting a SVIS on X and monitoring it at A is relatively straightforward (based on the distribution of X or using a 'non-parametric' approach).
- ✓ Now let D be some appropriately defined distressing outcome for some vulnerable individual– for example, the pollution-related death of a trout.
- ✓ To examine consistency of standards we will need to know about not just the distribution of X but also about the corresponding absolute or conditional pollution measures (effects)

$$P_A(D) = P(D\ at\ A)$$

and

$$P_A(D/X) = P(D\ at\ A/X)$$

involving, almost inevitably, an appropriate dose–response.

- These three quantities, X, $P_A(D)$ and $P_A(D/X)$, are all measured at point A.
- We also need to examine corresponding transferred levels and effects at the point B.
- Various prospects for harmonization arise: harmonization of standards between functions and between sites.
- There are $3 \times 3 = 9$ such possibilities for any two sites A and B.
- One possible concern (from the nine prospects) is to seek to harmonize standards on X (at A) and the corresponding $P_B(D/Y)$: the probability of a 'distressing outcome' at B given the pollution level at B.
  - ✓ In this example we seek to harmonize over distance and over function. Clearly, this will involve two forms of modelled relationship; firstly, that of how the pollution level distribu tion transfers from that of X at A to that of Y at B, and then of the level/effect (dose–response) relationship $P_B(D/Y)$.

- The easier version of this is to consider dose–response modelling for mutually consistent standard-setting between different pollution measures within the site, A, focusing on the pollution level X and the probability of the distressing outcome given X, $P_A(D/X)$,

Lecture 7

Spatial Methods for Environmental Processes

Spatial Prediction or Kriging
Situation for Spatial Sampling

- In geological studies it is frequently desired to **predict** the amount of ore or fossil fuel **that will be found at a site.**
- The prediction may be based on **values observed at other sites** in the region, and these other sites may be irregularly spaced in the region.
- It may be desired also to **predit or estimate the total amount of ore** in the region.
- Similarly, in pollution monitoring, **measurements of pollutant concentration at a few sites** may be used **to predit the concentration at a new site** or to estimate the **mean concentration** over a large area.
- In ecological studies also, **observations of animals or plants at a sample of sites** can be used **to predict the abundance in the vicinity of a new site** or in the **entire study region.**

- In such situations, it is useful to think of the value y_t of ore, pollutant, or animal abundance at location t as a random variable.
- From the observed values $y_1, y_2, ..., y_n$ at n sites $t_1, t_2, ..., t_n$, we wish **to estimate or predict the value of a new random variable y_0.**
  - ✓ The variable y_0 may be either the **value of the variale at a new site** or the **total of the variable** of interest over a larger geographic region.
- Since **y_0 is viewed as a random variable**, the inference problem is referred to as **prediction rather than estimation**, even though the prediction may be over **space rather than time**.

  - ❖ **(An *estimator* uses data to guess at a parameter while a *predictor* uses the data to guess at some random value that is not part of the dataset**).

- The **spatial prediction problem and its solution**—termed **kriging in geostistics**—are essentially the same as the model-based prediction approach to survey sampling with auxiliary information.

- The **prediction equations** can be written equivalently either with **covariances or with variances of differences (the variogram).**
- The **covariance approach** has been traditional in **statistics** and **time series**, while **variogram** has been traditional in **geostatistics**.
- The **two approaches** are **exactly equivalent** provided the covariance function-or the variogram—is known.

- **Slight differences** can arise when the **covariance function or variogram** is estimated using usual methods.

The Spatial Covariance Function

- In both ecological and geological surveys, the values of the variable of interest at different sites are typically **not independent of each other.**
- Rather, the **values at sites in close proximaity tend to be related to each other.**
- One summary of the **relationship** between the variables y_1, and y_2 associated with sites t_1 and t_2 is covariance,

$$\text{cov}(y_1, y_2) = E[y_1 - E(y_1)][y_2 - E(y_2)]$$

- **When the covariance between two sites** depends only **on their relative positions**, and not their exact locations within the study area, this relationship may be summarized with a covariance function c(h), giving the covariance of the value y-values for any two sites whose relative positions are separated by h, that is,

$$c(h) = \text{cov}(y_{t+h}, y_t)$$

- For a two-dimentional study region, the location t = (t_1, t_2) gives the coordinates in two dimentions of a site.
- The **displacement h** between the sites is also vector-valued, so that the **covariance function** may **dpend on direction as well as distance between sites**.
- If the expected value of the variable of interest is the same at every site and the covariance **depends only on the displacement between sites**, the process is called **second-order stationary**.
- If the covariance **depends only on the distance d between the sites and does not depend on direction,** the process is said to be **isotropic**, and the covariance function will be written c(d).

Linear Prediction (Kriging)

- Suppose observations have been observed at a sample of n sites, and one wishes to predict what will be found at another site.
- For notational simplicity, write y_i for y-value observed at the *ith* site in the sample, for *i=1, 2, ……..,n* and write c_{ij} for $cov(y_i, y_j)$, the covariance between the *y*-values at the *i*th and *j*th sites.
    - ✓ Note that $c_{ii} = var(y_i)$.
- From observed y-values at sites t_1, t_2,……..,t_n, it is desired to predict the value of the random variable y_0 at the site t_0.
    - ✓ For simplicity, the means $E(y_i)$ are assumed equal.

- The object is to find function $\hat{y}_0$ of the n observed y-values that is unbiased for y_0, that is,

$$E(\hat{y}_0) = E(y_0)$$

and which minimizes the mean square prediction error

$$E(y_0 - \hat{y}_0)^2$$

- In general, the **best estimator** is the conditional expectation of y_0 given the observed values $y_1, y_2, \ldots\ldots\ldots\ldots, y_n$.
- Determining $E(y_0 / y_1, y_2, \ldots\ldots\ldots\ldots, y_n)$ depends on the exact joint distribution of the y-values and may be difficult.

- A general more practical criterion is to find a linear function of the observed y-values that is unbiased and minimizes the mean square prediction error.

Writing the linear estimator as

$$\hat{y}_0 = \sum_{i=1}^{n} a_i y_i$$

- The problem is to find values a_1, a_2,……,a_n that **minimize** the mean square prediction error subject to unbiasedness.

- The **solution** (obtained by Lagrange multiplier method) is given in matrix form by

$$f = G^{-1} h$$

where the vector f and h are

$$f = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m \end{pmatrix} \qquad h = \begin{pmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{n0} \\ 1 \end{pmatrix}$$

and the matrix G is

- $$G = \begin{pmatrix} c_{11} & c_{12} \ldots\ldots\ldots c_{1n} & 1 \\ c_{21} & c_{22} \ldots\ldots\ldots c_{2n} & 1 \\ \vdots & & \\ \vdots & & \\ c_{n1} & c_{n,2} \ldots\ldots\ldots c_{nn} & 1 \\ 1 & 1 \ldots\ldots\ldots 1 & 0 \end{pmatrix}$$

- The constant ***m***, which is obtained along with the coefficient a_i, is the Lagrange multiplier and is used in calculating the mean square prediction error.

- The **best linear pridictor** $\hat{y}_0$ is called the **kriging predictor** in geology.

- The **mean square prediction error** of $\hat{y}_0$, also called the kriging variance, is

$$E(y_0 - \hat{y}_0)^2 = c_{00} - \sum_{i=}^{n} a_i c_{i0} - m$$

- If the stochastic process giving rise to the **population is Gausian**, that is, the **joint distribution** of any finite set of y-values is **multivariate normal**, **then**
  - ✓ the **best linear predictor** $\hat{y}_0$ is the **preditor**, having **lowest mean square prediction error** of **any function** of the n observed y-values.

- The **optimality of the prediction results** depends on the **covariance function** being known.
- In practice, however, the **covariance function is estimated** from data—either data from the present survey or a larger store of data from past surveys.
- For a process that is **stationary and isotropic**; the covariance of sites distance d apart can be estimated using a sample covariance based on the n_d distinct pairs of sites in the data that are distance d (or approximately distance d) apart.
- A simple covariance estimator is

$$\hat{c}(d) = \frac{1}{n_d} \Sigma (y_{t_i} - \bar{y})(y_{t_j} - \bar{y})$$

  where the summation is over the distict pairs of observations that are approximately distance d apart, and n_d is the number of such distinct pairs.
- A **smooth curve** may then be **fitted** to the estimated covarnces, using a method such as **nonlinear least squares**, to obtain an estimated covariance fuction and hence an estimate of covariance for any distance.

Example: A srimp survey was conducted to estimate the **spaptial covariance function**, which in turn will be used to predict the amount of catch **at a new location**. The catch data were originally recorded in pounds (lbs), with distances measured in nautical miles (nmi). A research vessel made tows of a trawl net approximately **one nautical mile apart in a grid pattern**. **Sample covariances** were **computed** using pairs of data lumped into distance intervals. Then a curve of the form $a\exp(-bx)$ was fitted by **nonlinear least squares** to the covariance estimates. The fitted covariance function was

$$c(x) = 5.1e^{-0.49x}$$

Suppose one tows has been made with a catch of $y_1 = 5.526$ (units are in 1000s of lbs) and second tow 6 nmi. away produced $y_2 = 1.417.$ What would be the predicted catch y_0 at a location 1 nmi from the first tow and 5.4 nmi from the second?

Solution: The variance is c(0) = 5.1. The covariance for the two tows 6 nmi apart is $c_{12} = 5.1\{\exp[-0.49(6)]\} = 0.3$. The covariances with the new site are $c_{10} = 5.1\{\exp[-0.49(1)]\} = 3.1$ and $c_{20} = 5.1\{\exp[-0.49(5.4)]\} = 0.4$. The predicted equation is

$$\begin{pmatrix} a_1 \\ a_2 \\ m \end{pmatrix} = \begin{pmatrix} 5.1 & 0.3 & 1 \\ 0.3 & 51 & 1 \\ 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 3.1 \\ 0.4 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.104 & -0.104 & 0.5 \\ -0.104 & 0.104 & 0.5 \\ 0.5 & 0.5 & -2.7 \end{pmatrix} \begin{pmatrix} 3.1 \\ 0.4 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.78 \\ 0.22 \\ -0.95 \end{pmatrix}$$

So $a_1 = 0.78,\ \ a_2 = 0.22 \ \ and \ \ m = -0.95$

The predicted catch is

$$\hat{y}_0 = 0.78(5.526) + 0.22(1.417) = 4.622$$

Or 4622 lbs at new location.

The predition mean square error is

$$E(y_0 - \hat{y}_0)^2 = c_{00} - \sum_{i=}^{n} a_i c_{i0} - m = 5.1 - .78(3.1) - 0.22(0.4) + 0.95 = 3.5$$

giving a root mean square error of 1.9.

The Variogram

In geological sciences, spatial variation has traditionally been summarized using the variogram in place of the covariance function. The variogram is defined as the variance of the difference of y-values at separate sites:

$$\text{var}(y_{t+h} - y_t) = 2\gamma(h)$$

The function $\gamma(h)$ is called the semivariogram. When the process is second order stationary, the covariance function and the variogram contain equivalent information, since

$$\gamma(h) = c(0) - c(h)$$

Note that c(0) = var(y_t), the variance of y at any site t. Cressie points out that the variogram exits even for some processes that are not second-order stationay, and hence is more general than the covariance function.

Assume that the process is also stationary in the mean, that is,

$$E(y_t) = E(y_s),$$

For any sites t and s in the study region. Then $\text{var}(y_{t+h} - y_t) = E(y_{t+h} - y_t)^2$ and the simple method for estimating the semivariogram is

$$2\hat{\gamma}(d) = \frac{1}{n_d}\Sigma(y_{t_i} - y_{t_j})^2$$

Where the summation is over all distinct pairs of sites in the sample that are distance d apart, and n_d is the number of pairs that distance apart. For irregularly spaced data, pairs of sites that approximately the same distance apart may be lumped together. Finally, a smooth curve is fitted to the variogram estimates to obtain values of the variogram for all distances. Estimation of the semivariance function γ rather than of the covariance function c is recommendated by Cressie (1991), who points out that $\hat{\gamma}(d)$ is an unbiased estimator of $\gamma(d)$, while $\hat{c}(d)$ is not unbiased for c(d) and that the semivariance estimates do better than the covariance estimates in the presence of an undetected linear trend in the variable of interest.

The **prediction equations** may be written in terms of semivariogram. Suppose that observations have been made at a sample of n sites, and one wishes to predict what will be found at another site. For notational simplicity, write y_i for the y-value observed at the ith site in the sample, for i=1,2,….,n, and write γ_{ij} for $\gamma(y_i - y_j)$, the semivariogram value for the diffrenece between the ith and the jth sites. For observed y-values at sites $t_1, t_2, \ldots\ldots\ldots, t_n$, it is desired to predict the value of the random variable y_0 at the site t_0.

The object is to find a function $\hat{y}_0$ of the n observed y-values that is unbiased for y_0, that is,

$$E(\hat{y}_0) = E(y_0)$$

and that minimizes the mean square prediction error

$$E(y_0 - \hat{y}_0)^2$$

Writing the linear estimator as

$$\hat{y}_0 = \sum_{i=1}^{n} a_i y_i$$

The problem is to find values $a_1, a_2, \ldots\ldots\ldots, a_n$ that minimize the mean square prediction error subject to unbiasedness.

The solution is given in matrix form by

$$a = \Gamma^{-1}\gamma$$

where

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m^* \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ 1 \end{pmatrix}$$

and

$$\tau = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \gamma_{2n} & 1 \\ & & & \\ \gamma_{n1} & \gamma_{n2} & \gamma_{nn} & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

The mean square prediction error of $\hat{y}_0$ is

$$E(y_0 - \hat{y}_0)^2 = \sum a_i \gamma_{i0} + m^*$$

Note that $m^* = -m,$ where m is the constant obtained using the covariance-based predition equation.

Example: A srimp survey was conducted to estimate the semi-variogram function, which in turn will be used to predict the amount of catch at a new locations. The catch data were originally recorded in pounds (lbs), with distances measured in nautical miles (nmi). A research vessel made tows of a trawl net approximately one nautical mile apart in a grid pattern. The fitted semi-variogram function was

$$\gamma(x) = 5.1(1 - e^{-0.49x})$$

Suppose one tows has been made with a catch of $y_1 = 5.526$ (units are in 1000s of lbs) and second tow 6 nmi. away produced $y_2 = 1.417.$ What wold be the predicted catch y_0 at a location 1 nmi from the first tow and 5.4 nmi from the second?

Solution: The semivariance for the two tows 6 nmi. apart is

$$\gamma_{12} = 5.1\left(1 - e^{-0.49(6)}\right) = 4.8$$

The semivariances with the new site are $\gamma_{10} = 5.1\{1 - \exp[-0.49(1)\} = 2.0$ and $\gamma_{20} = 5.1\{1 - \exp[-0.49(5.4)\} = 4.7.$

The prediction equation with the semivariance is

$$\begin{pmatrix} a_1 \\ a_2 \\ m^* \end{pmatrix} = \begin{pmatrix} 0 & 4.8 & 1 \\ 4.8 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2.0 \\ 4.7 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.104 & 0.104 & 0.5 \\ 0.104 & -0.104 & 0.5 \\ 0.5 & 0.5 & -2.4 \end{pmatrix} \begin{pmatrix} 2.0 \\ 4.7 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.78 \\ 0.22 \\ 0.95 \end{pmatrix}$$

So $a_1 = 0.78, \quad a_2 = 0.22, \quad and \quad m^* = 0.95$

The predicted catch is

$$\hat{y}_0 = 0.78(5.526) + 0.22(1.417) = 4.622$$

Or 4622 lbs at the new location, and the prediction mean square error is

$$\hat{E}(y_0 - \hat{y}_0)^2 = 0.78(2.0) + 0.22(1.417) + .95 = 3.5$$

Predicting the Value Over a Region

In many spatial sampling situations, one wishes to predict the mean (or total) of the y-values over a region A. Denote this mean by y_0, where

$$y_0 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

if A is a study region partitioned into N sites, or

$$y_0 = \frac{1}{|A|}\int_A y_t dt$$

where $|A|$ is the area of the study region, if samples can be taken at points t throughout a continuous study region. The population regional mean y_0 is a random variable, since the y-value at each site is random.

Then the predition solution is

$$\gamma_{i0} = \frac{1}{N}\sum_{j=1}^{N}\gamma_{ij}$$

in the discrete case, and

$$\gamma_{i0} = \frac{1}{|A|}\int_A \gamma(y_i - y_t)dt$$

in the continuous case. In each case, γ_{i0} is the average semivariance between site I and sites in the region A.

The minimized mean square prediction error is

$$E(y_0 - \hat{y}_0)^2 = \sum_{i=1}^{n} a_i\gamma_{i0} + m^* - \gamma_{00}$$

where the average semivariance γ_{00} is

$$\gamma_{00} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{ij}$$

in the discrete case, and

$$\gamma_{00} = \frac{1}{|A|^2}\int_A\int_A \gamma(t-v)dtdv$$

in the continuous case.