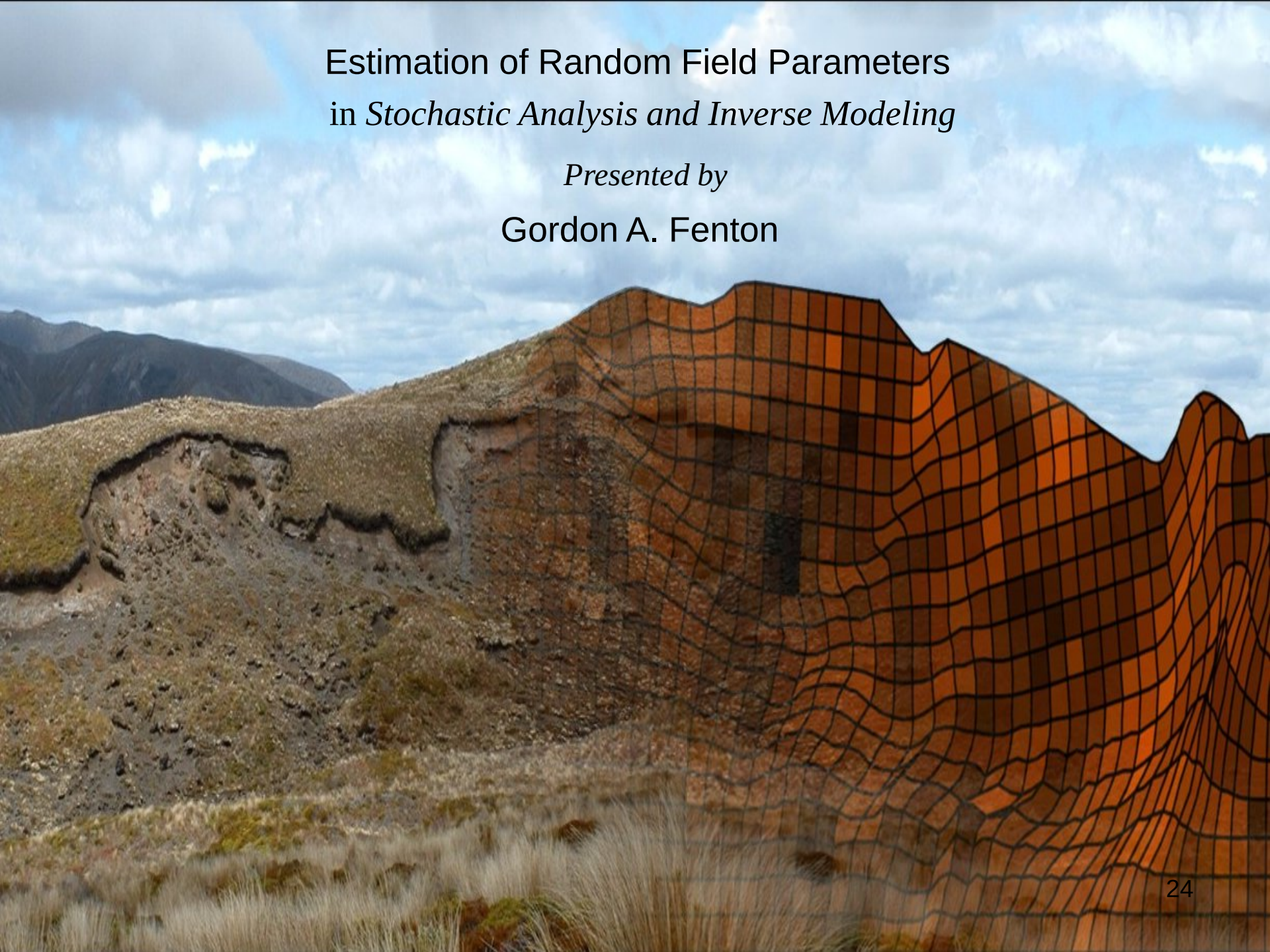
Estimation of Random Field Parameters

in *Stochastic Analysis and Inverse Modeling*

Presented by

Gordon A. Fenton



Introduction

- Our probabilistic models require distributions for the soil properties.
- Each distribution is characterized by parameters such as the mean and standard deviation.
- spatial correlation is characterized by a correlation length.
- the mean is generally easy to estimate, mean trends require somewhat more data.
- the standard deviation requires even more data to estimate – we often depend on the literature for estimates.
- the correlation length requires huge amounts of data to estimate – even difficult to find useable estimates in the literature (design may have to proceed using a *worst case* correlation length).

Interpolation vs. Extrapolation

How the statistical estimates are to be used influences how they are determined;

1. *Interpolation*: the goal here is to use the data to best describe the site at which the data was obtained. In this case all trends should be accounted for. May want to perform simulations conditioned (i.e. “pinned”) on the data. Correlation length is now of “residual” variability.
2. *Extrapolation*: the goal here is to use the data to attempt to characterize other (similar) sites. In this case, trends should only be accounted for if they are expected to recur at the other sites. Mean estimates will be uncertain, variances will be underestimated. Correlation length will be larger than estimated (i.e. generally unknown).

Interpolation vs Extrapolation

Interpolation:

- data is used to characterize site at which data was obtained
- estimator errors *decrease* with increasing correlation between observations, i.e. the more highly correlated the site is, the fewer samples required to characterize it. (Unfortunately, we generally don't know a-priori how highly correlated a site is!)
- the data can be assumed known – uncertainty now occurs between data points, so we need only model this *residual* variability.
- Kriging and/or conditional models can be used to characterize the residual variability.

Interpolation vs. Extrapolation

Extrapolation:

- the data are being used to characterize the soil *population* (i.e. to infer the population parameters for use at other sites).
- estimator error *increases* with increasing correlation between observations, i.e. the more highly correlated a site is, the less representative it is of other sites – you cannot expect to accurately characterize a neighboring site if all of your samples are taken from a (highly correlated) stiff soft clay layer at the current site.
- statistical estimates of population parameters are typically quite inaccurate (due to correlation), especially estimates of correlation length.

Interpolation vs. Extrapolation

- Practicing geotechnical engineers are typically *interpolating*. That is, they sample with the goal to characterize the site at which the samples are observed.
- Published research papers and textbooks are *extrapolating* (or at least they *should* be). That is, they are expressing soil property information that is meant to be useful at sites other than the single site at which the data were obtained.
- Unfortunately, all too often, research papers will provide soil property statistics where the locally observed trend has been removed. This leads to significantly underestimated variabilities (only useful at sites where similar trends occur and have been similarly removed).
- In *extrapolation* trends should be generally considered to be part of the uncertainty being characterized.

Choosing a Distribution

Once the data have been gathered, we need to decide how to best represent the "population". The first step is to decide on a population distribution. There are several possibilities;

1. *Trace driven simulation*: use the data directly in a simulation. This is the least preferable approach since it can only reproduce the data and not all possibilities. This approach is most commonly used in earthquake ground motion simulation.
2. *Empirical distribution*: the data are used to define an empirical cumulative distribution function (e.g. P[$X < x$] is just equal to the number of observed values less than x). This does not allow for the extremes that often control design. That is, most samples will not include those 1/1000 extremes that would lead to failure.

Choosing a Distribution

3. *Fit a Distribution*: A common distribution, such as the normal or lognormal, is fitted to the data. The advantages to this approach are that
   a) irregularities in the data, due to the natural variability in any data set, are smoothed out. That is, we don't end up with a distribution that is skewed by an outlier in the data set.
   b) the known physics of the property can be properly represented (e.g. properties such as porosity, friction angle, and Poisson's ratio have known (or at least almost known) upper and lower bounds, so a bounded distribution would be appropriate).
   c) extremes can also be modeled in a physically reasonable way.

Choosing a Distribution

Extrapolation:

- fit the simplest distribution that you can – you are trying to model the population, not the specific data set.

Interpolation:

- fit a distribution of reasonable complexity – just remember that you still need to capture the range of possibilities that might occur between your observation points (so there is probably little point in employing a 20 parameter distribution).

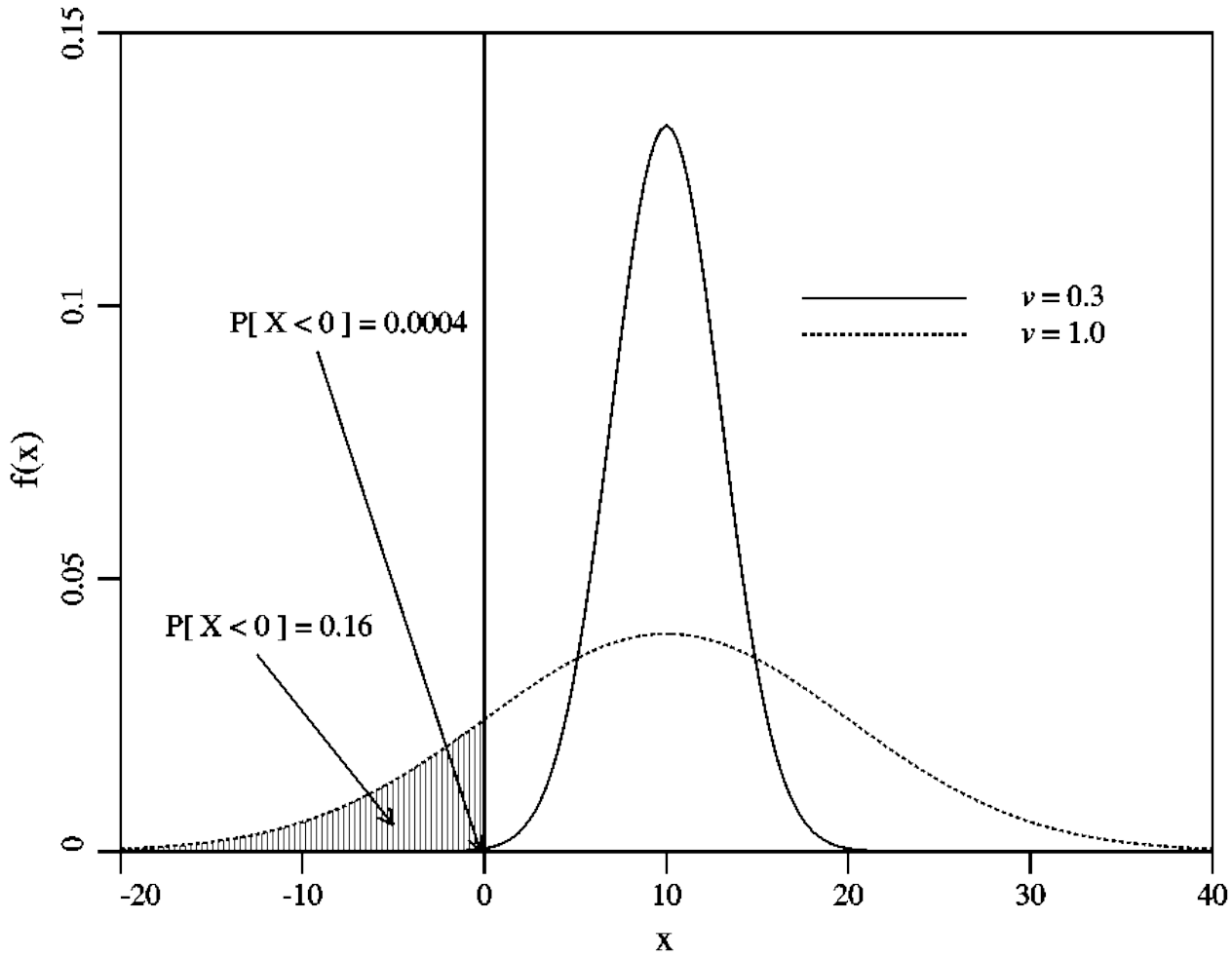
Choosing a Distribution

The normal distribution is a very popular choice, especially if the soil property is a random field (since we then only need to know the mean and covariance structure).

The one major disadvantage to the normal distribution is that its *range is from* $-\infty$ *to* $+\infty$, so for many non-negative soil properties it is not physically possible.

However, if the probability of obtaining a negative value is sufficiently small, the normal distribution might be a reasonable approximation. What is meant by "sufficiently small" depends on the acceptable probabilities of the extremes that might lead to failure.

Choosing a Distribution



Probability that $X < 0$ for coefficients of variation $\nu = 0.3$ and 1.0

Goodness-of-Fit

Once a distribution has been selected and then fit, by estimating its parameters using the collected data, the fit must be assessed. There are two common approaches to measuring how well the assumed distribution fits the data;

1. Frequency comparisons and probability plots, and
2. Goodness-of-Fit tests.

Frequency Comparison

Example:

Suppose that, just after construction, a series of 50 randomly selected one-kilometre long sections of highway through a hilly region were selected to evaluate the annual probability of slope failure under the existing design code. The number of years until an observable slope failure occurred within each one-kilometre length, t_i , was recorded, with the following results

3, 2, 8, 9, 10, 4, 4, 2, 7, 7, 1, 14, 2, 1, 8, 3, 4, 5, 4, 2, 10,

2, 1, 7, 8, 4, 3, 3, 21, 1, 3, 9, 1, 4, 5, 1, 4, 1, 4, 3, 5, 3, 1,

9, 1, 6, 3, 5, 12, 11

A previous analysis of similar data suggested that the annual probability of observable slope failure in each one-kilometre section of highway is 0.2. Assuming that sections fail independently and that each year constitutes an independent trial, how reasonable does the hypothesis that the annual probability of slope failure per km is 0.2 appear to be on the basis of this data?

Frequency Comparison

If sections fail independently, and each year is also independent, then we have 50 independent observations of the ‘number of trials’ (i.e. years) to first failure of a 1-km section. Under the given assumptions, the ‘number of trials to first failure’ follows a geometric distribution.

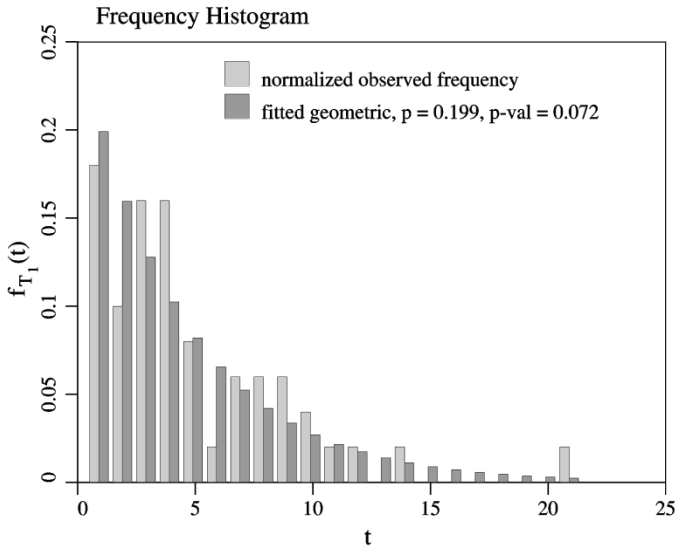
The estimate of the annual probability of slope failure per km is just one (year) over the average time to slope failure;

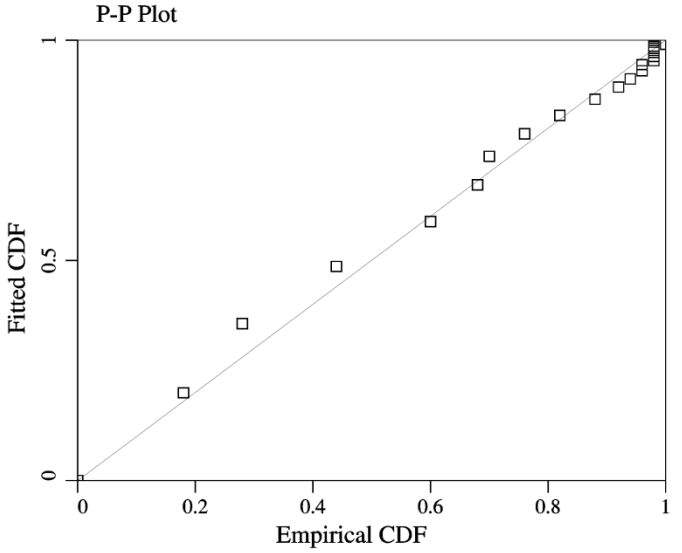
$$\hat{p} = \frac{1}{\left(3+2+\cdots+11\right)/50} = 0.199$$

which is very close to the hypothesized annual probability.

The following page compares the frequency histogram with that predicted by theory along with the empirical and fitted cumulative distribution functions.

Frequency Histogram





Frequency-density plot of annual failure probability and fitted geometric distribution. The lower plot compares the empirical cumulative distribution with the fitted cumulative geometric distribution.

Parameter Estimation

There are many ways to estimate distribution parameters – some are better than others. The common criteria used to compare estimators are

1. *unbiasedness*: E[estimator] = parameter?
2. *consistency*: $\lim_{n\to\infty}$ estimator = parameter?
3. *efficiency*: Var[estimator] small?
4. *sufficiency*: utilizes all pertinent information?

Classical Estimators

Sample Mean: $\hat{\mu}_X = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ estimates the true mean μ_X

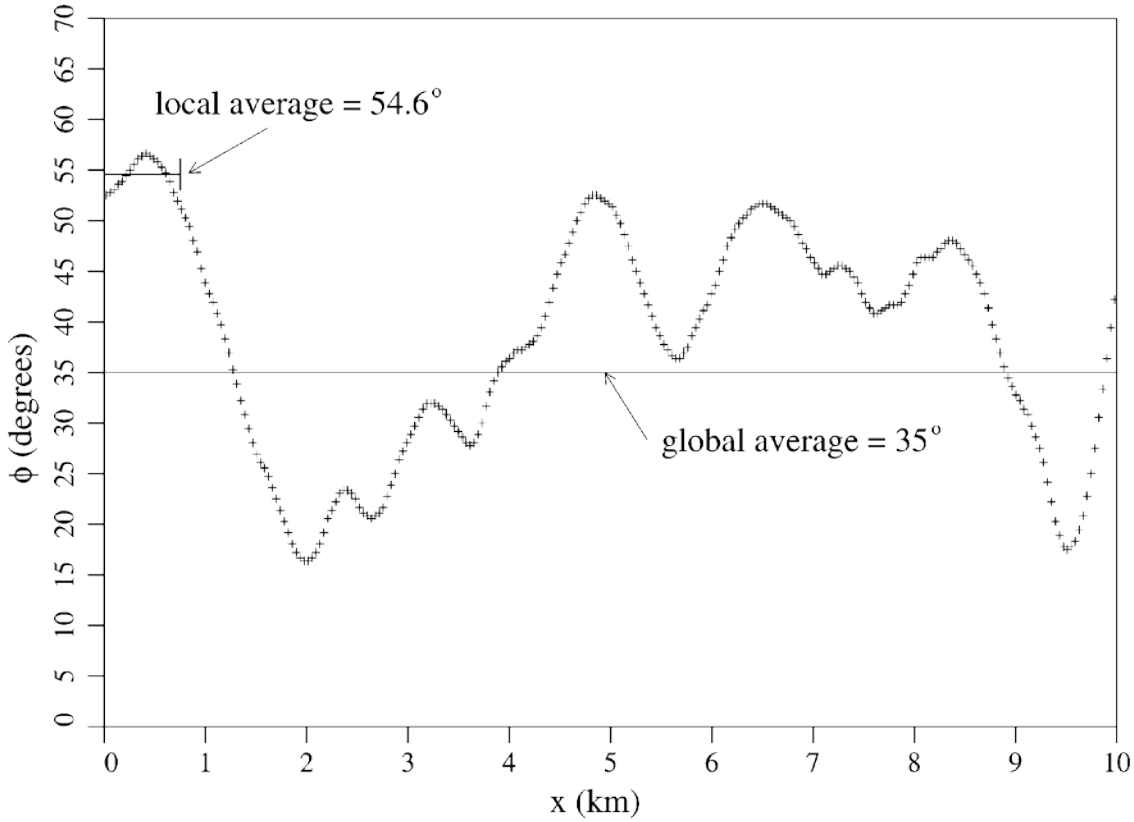
Sample Variance: $\hat{\sigma}_X^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

is an estimate of the true variance σ_X^2

Sample Correlation: $\hat{\rho}_X(j\Delta x) = \frac{1}{\hat{\sigma}_X^2(n-j-1)}\sum_{i=1}^{n-j}(x_i - \overline{x})(x_{i+j} - \overline{x})$

is an estimate of the true correlation ρ_X

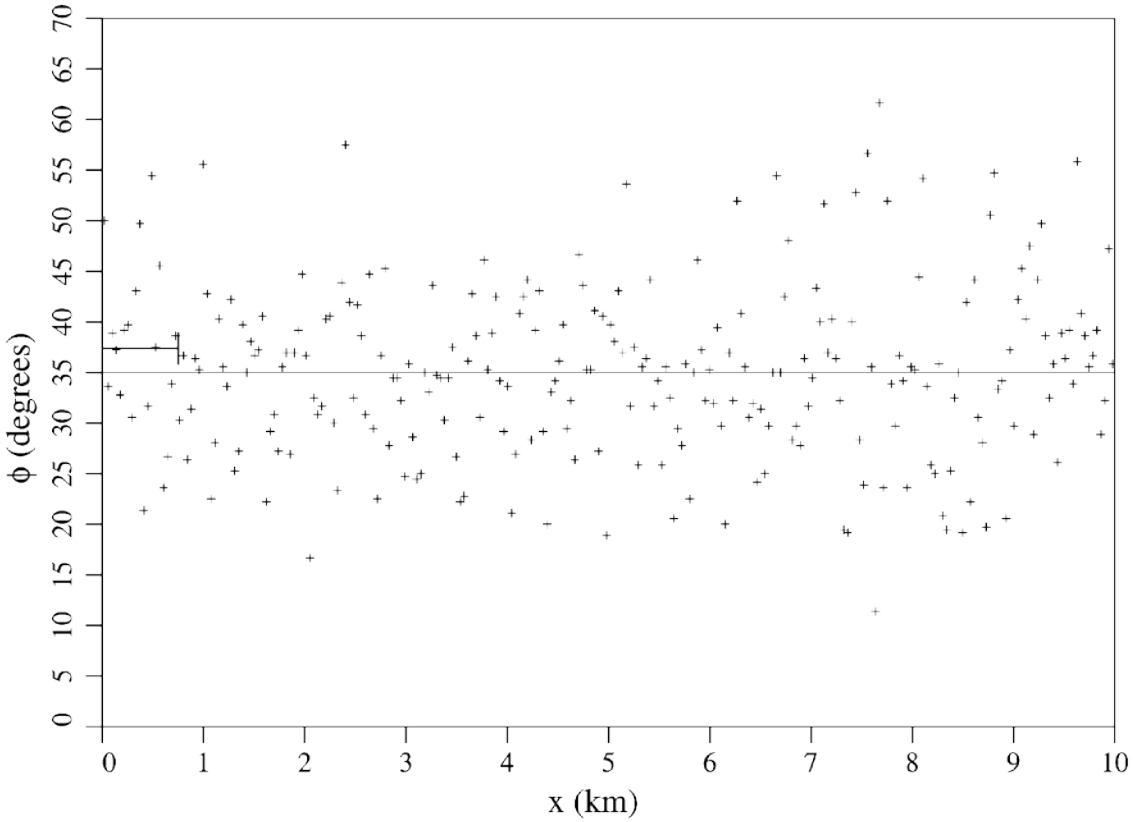
Estimation in the Presence of Correlation



Friction angle measured at regular locations along a 10 km line.

- if we measured only from 0 – 0.75 km, our estimate of the global average would be very poor. In fact, most 1 km segments would give poor results.
- best to sample at widely spaced points (>> θ).
- the variance estimated over 0 – 0.75 km is *far less* than the site variance.

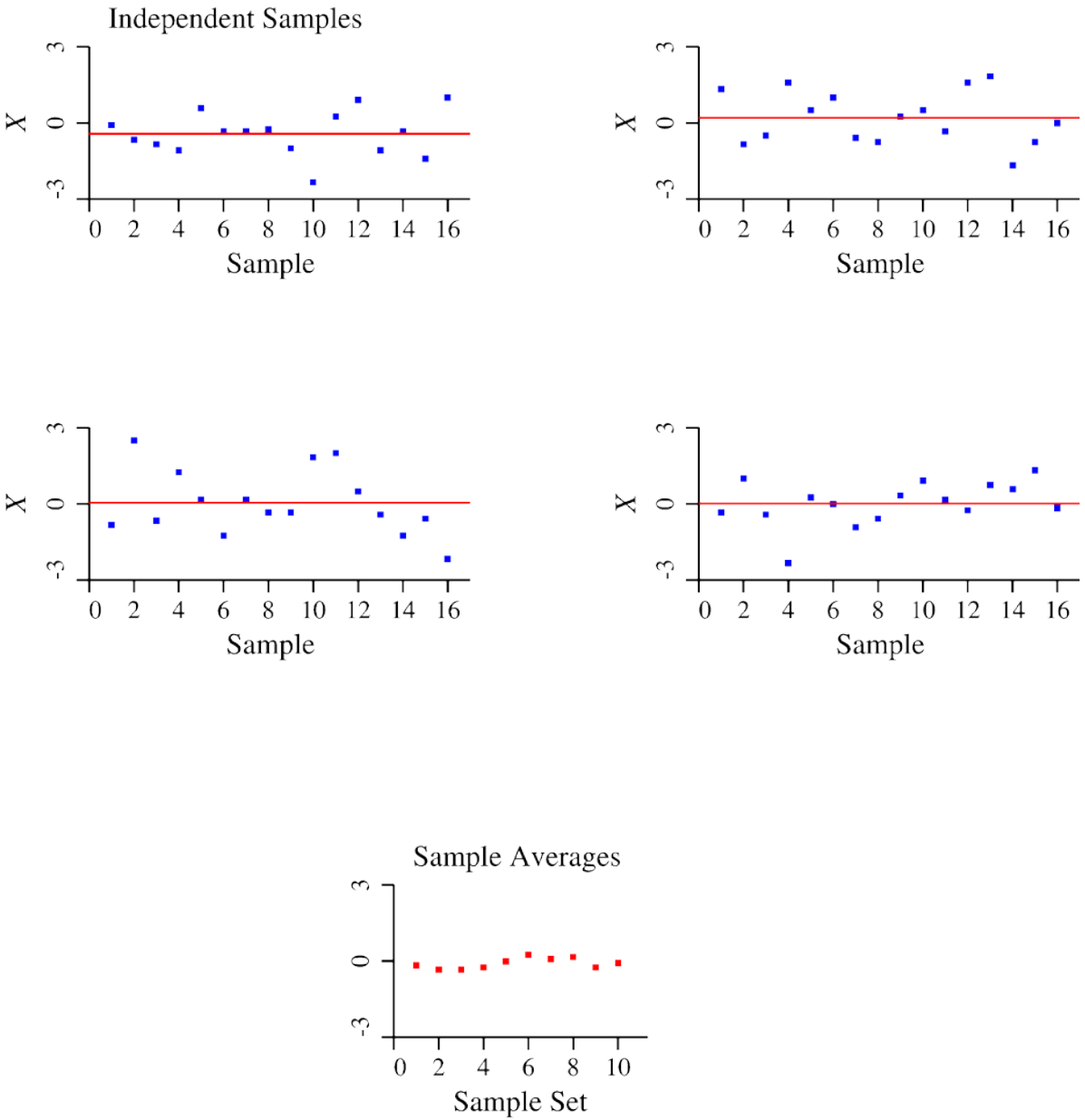
Estimation Without Correlation



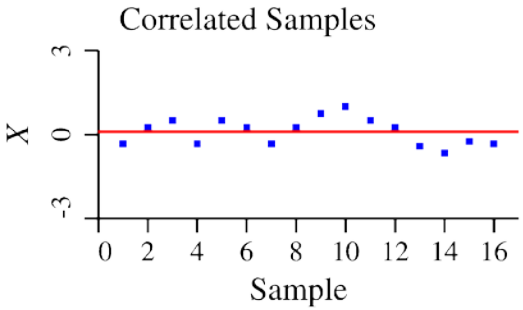
Friction angles measured along a 10 km line where soil properties are largely spatially independent.

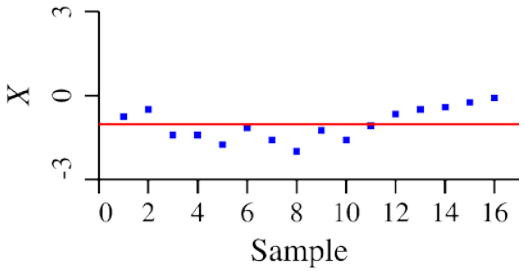
- in this case, both the estimated mean and variance obtained over 0 – 0.75 km are quite representative of the entire 10 km.

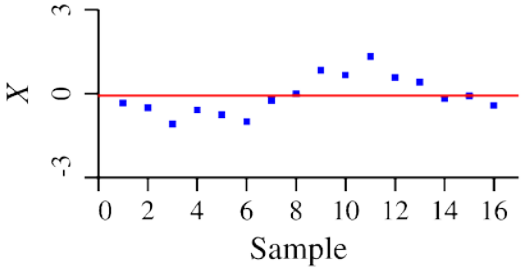
Classical Estimate of the Mean: $\hat{\mu}_X = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

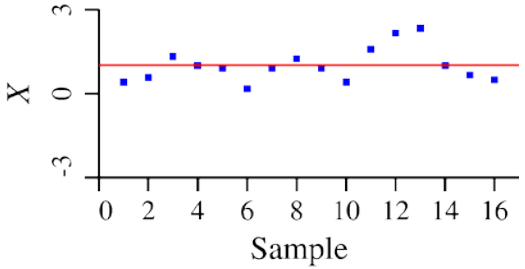


Introduction of Correlation Between Samples

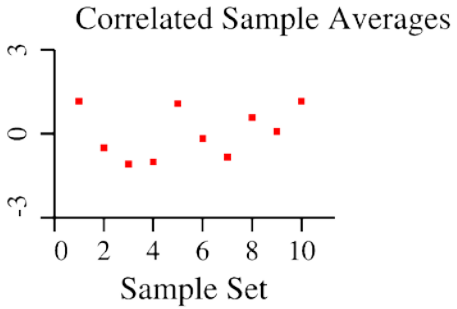


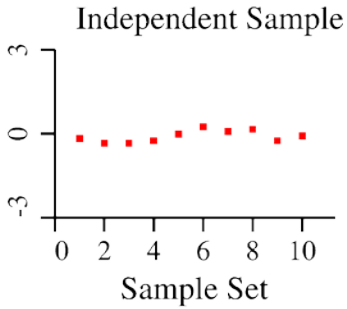




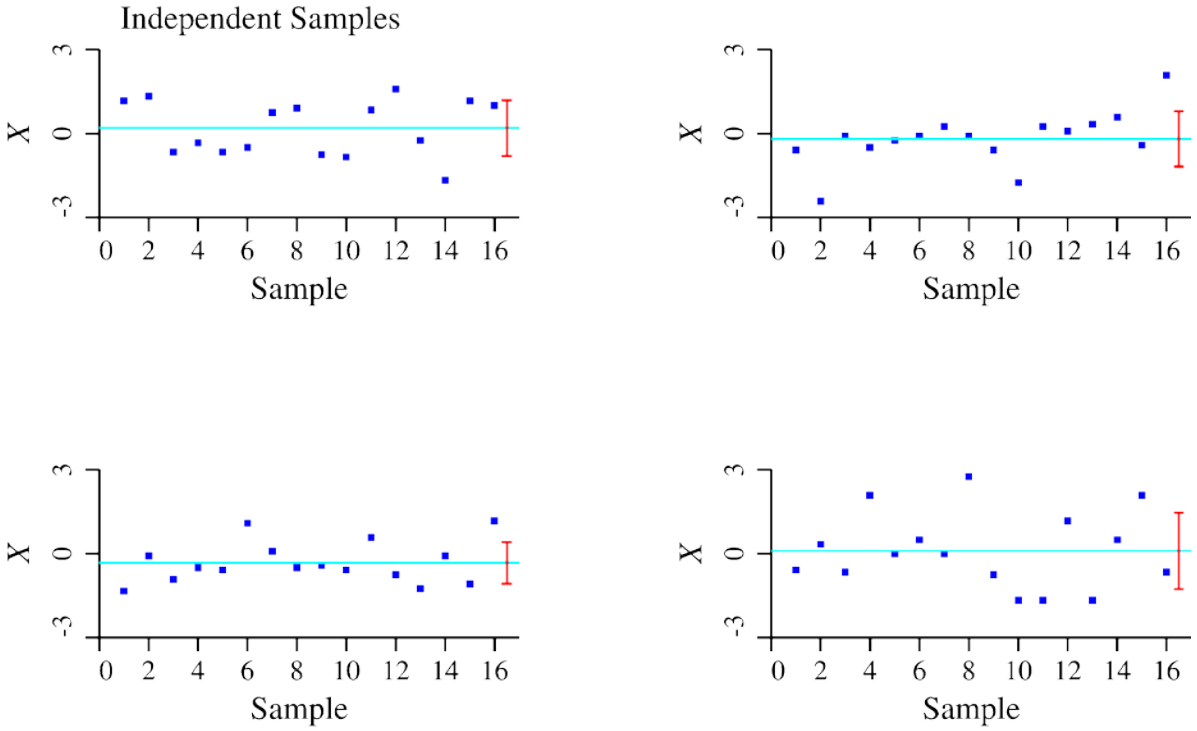


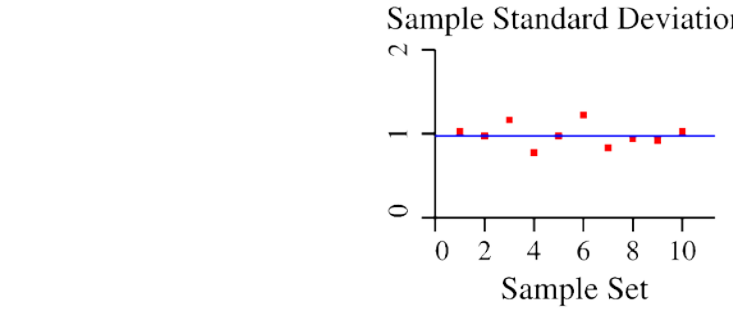
Recall from last slide:



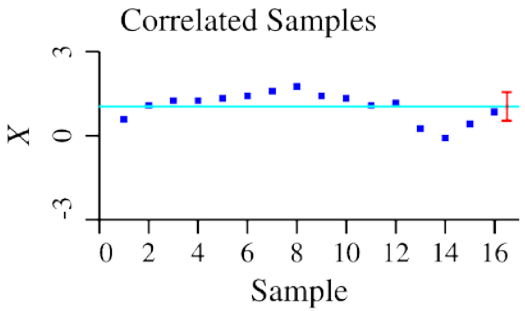


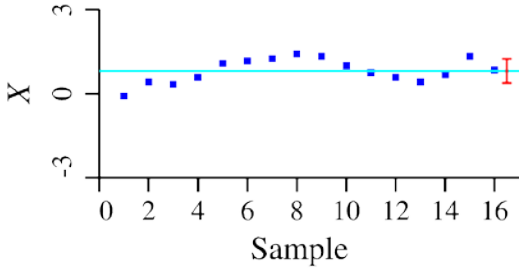
Classical Estimate of the Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

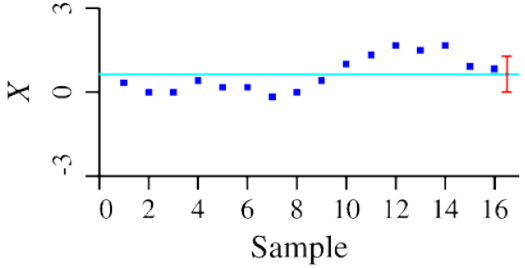


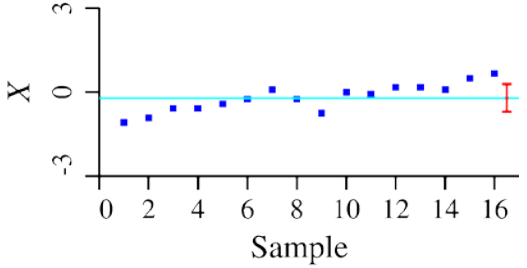


Introduction of Correlation Between Samples

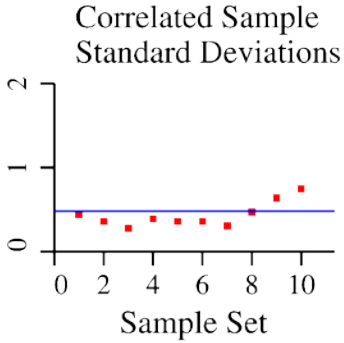


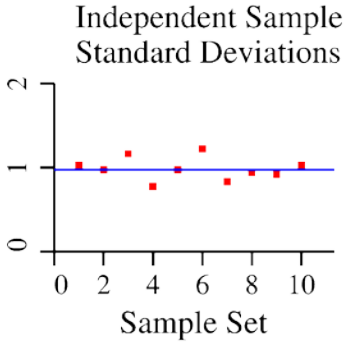






Recall from last slide:





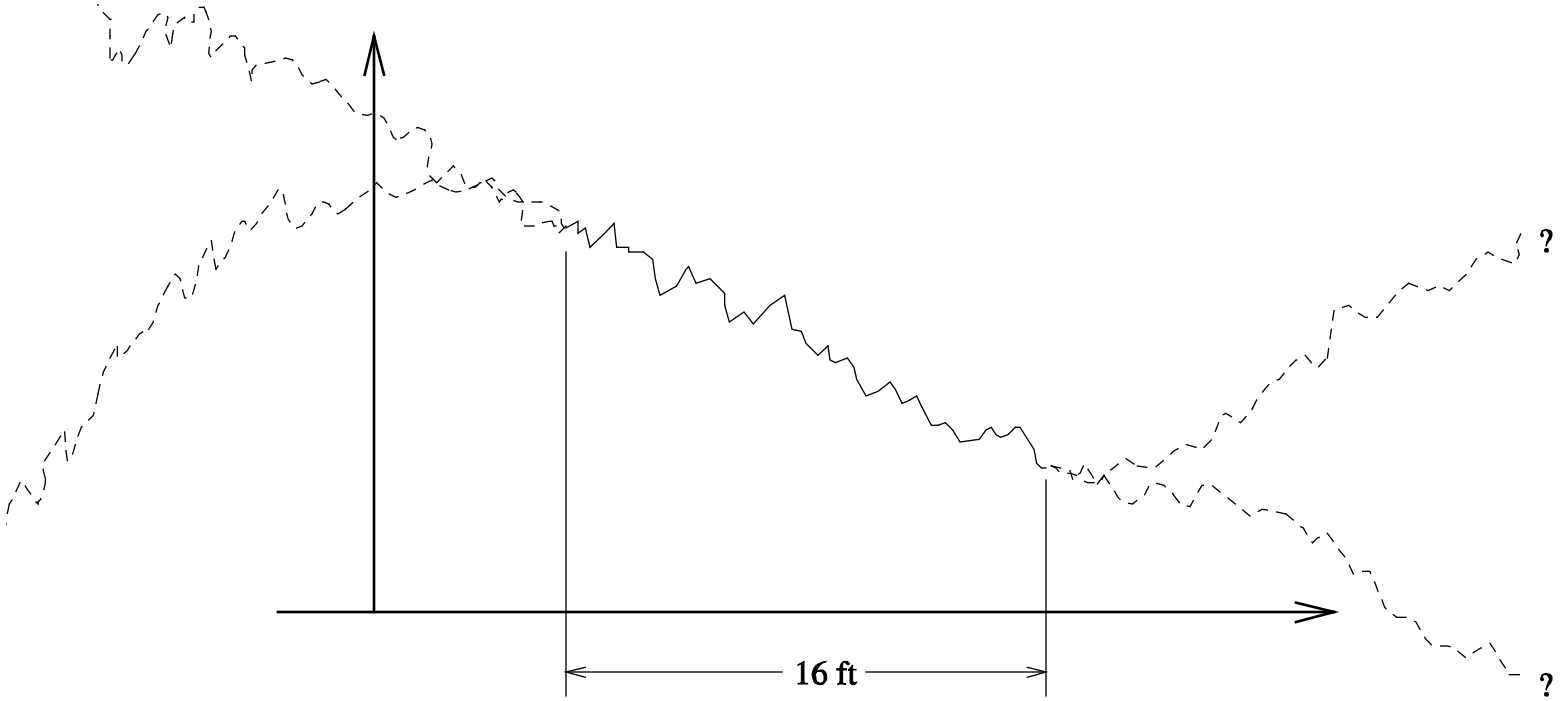
Case 1: Data are Gathered over the Design Site

- we will know the soil properties at the data site locations and will not be attempting to extrapolate beyond the site borders,
- estimates for μ_X , σ_X , and θ_X are "local" and can be be considered to be reasonably accurate
- best estimates for the value and variability of the random field between observation points can be obtained using *Best Linear Unbiased Estimation* (BLUE) or *Kriging*.
- probability estimates should be obtained using a random field conditioned on the data (possibly via conditional simulation)

Case 2: Data are Gathered at a Similar Site

- data gathered at another similar site are used to characterize the design site (extrapolation) – this might occur during preliminary design, e.g. before the site has been cleared.
- much greater uncertainty in applying the resulting statistics to the design site. The sample mean should be viewed with caution and the sample variance should be assumed to be underestimated.
- BLUE and Kriging are not options because no data available at site.
- treatment of trends needs special care – are they likely to recur at the design site?

Characterization of Trends?



- *locally*, trends should be accounted for in design
- *globally*, trends should only be accounted for if expected to continue (or repeat) offsite

Estimating the Mean

Classical sample mean: $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$

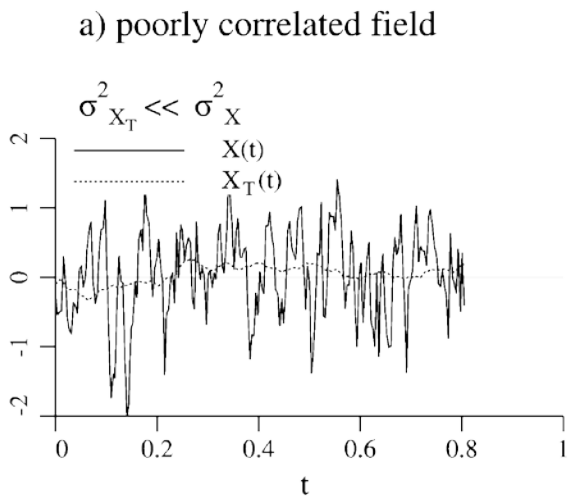
$$\mathrm{E}\left[\hat{\mu}_X\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] = \mu_X \qquad \text{(unbiased)}$$

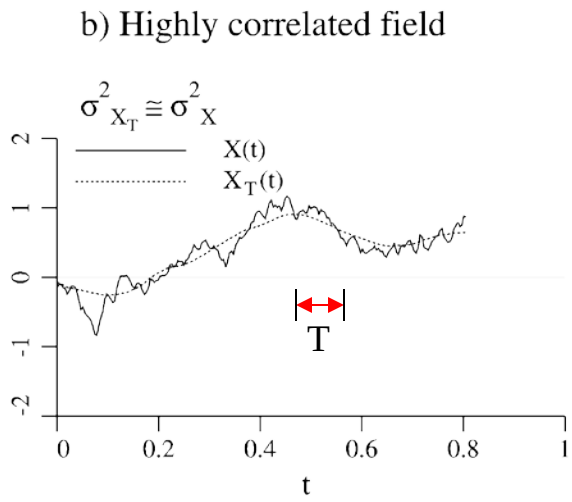
$$\mathrm{Var}\left[\hat{\mu}_X\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left[X_i, X_j\right] = \left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{\mathrm{ij}}\right]\sigma_X^2 \simeq \gamma\left(T\right)\sigma_X^2$$

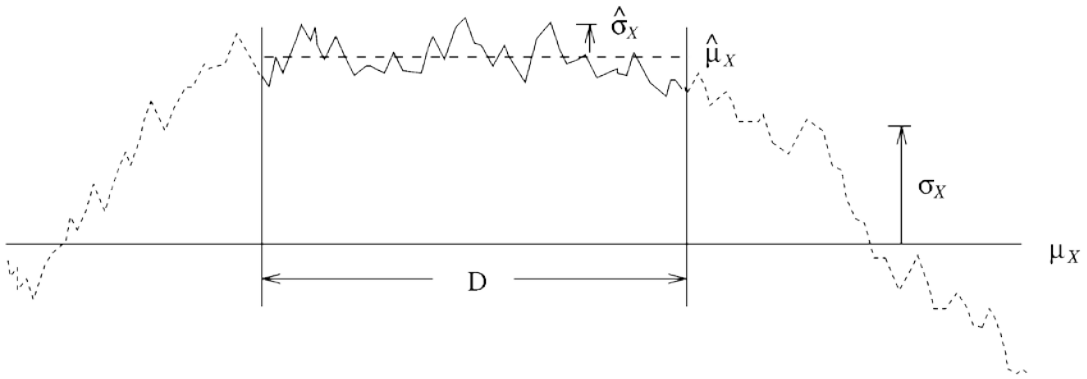
where T is the domain over which the samples are gathered.

For highly correlated samples, $\gamma(T) \approx 1.0$, and the sample mean could be quite variable (i.e. large estimator error).

Effect of Correlation on Statistical Estimates







- mean and variance estimates are *locally* accurate but *globally* poor
- this is one of the reasons why geotechnical engineering is difficult to codify

Estimating the Variance

Classical Maximum Likelihood Estimator:

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}_X\right)^2$$

$$\mathrm{E}\left[\hat{\sigma}_X^2\right] = \sigma_X^2\left[1-\gamma(T)\right]$$

Since $\gamma(T)$ lies between 0 and 1, $\mathrm{E}\left[\hat{\sigma}_X^2\right] \le \sigma_X^2$ (unconservative)

For strongly correlated fields, $\mathrm{E}\left[\hat{\sigma}_X^2\right] \to 0$ (highly biased)

Estimating the Covariance Structure

Consider a sequence of observations $X_1, X_2, \ldots, X_n$, each separated by distance Δx. Then for $\tau_j = j\Delta x$, $j = 0, 1, \ldots, n - j - 1$ we have

$$\hat{C}(\tau_j) = \frac{1}{n - j - 1} \sum_{i=1}^{n-j} (X_i - \hat{\mu}_X)(X_{i+j} - \hat{\mu}_X)$$

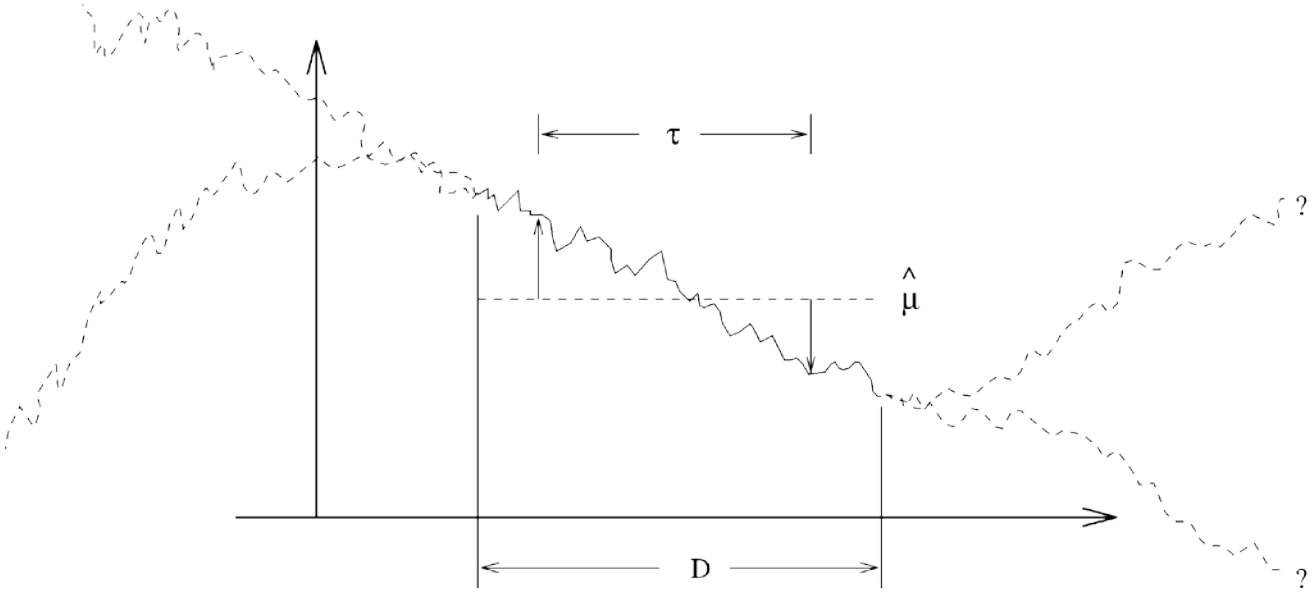
$$\hat{\rho}(\tau_j) = \frac{\hat{C}(\tau_j)}{\hat{\sigma}_X^2} \qquad \text{(where } \hat{\sigma}_X^2 = \hat{C}(0)\text{)}$$

Estimating the Covariance Structure

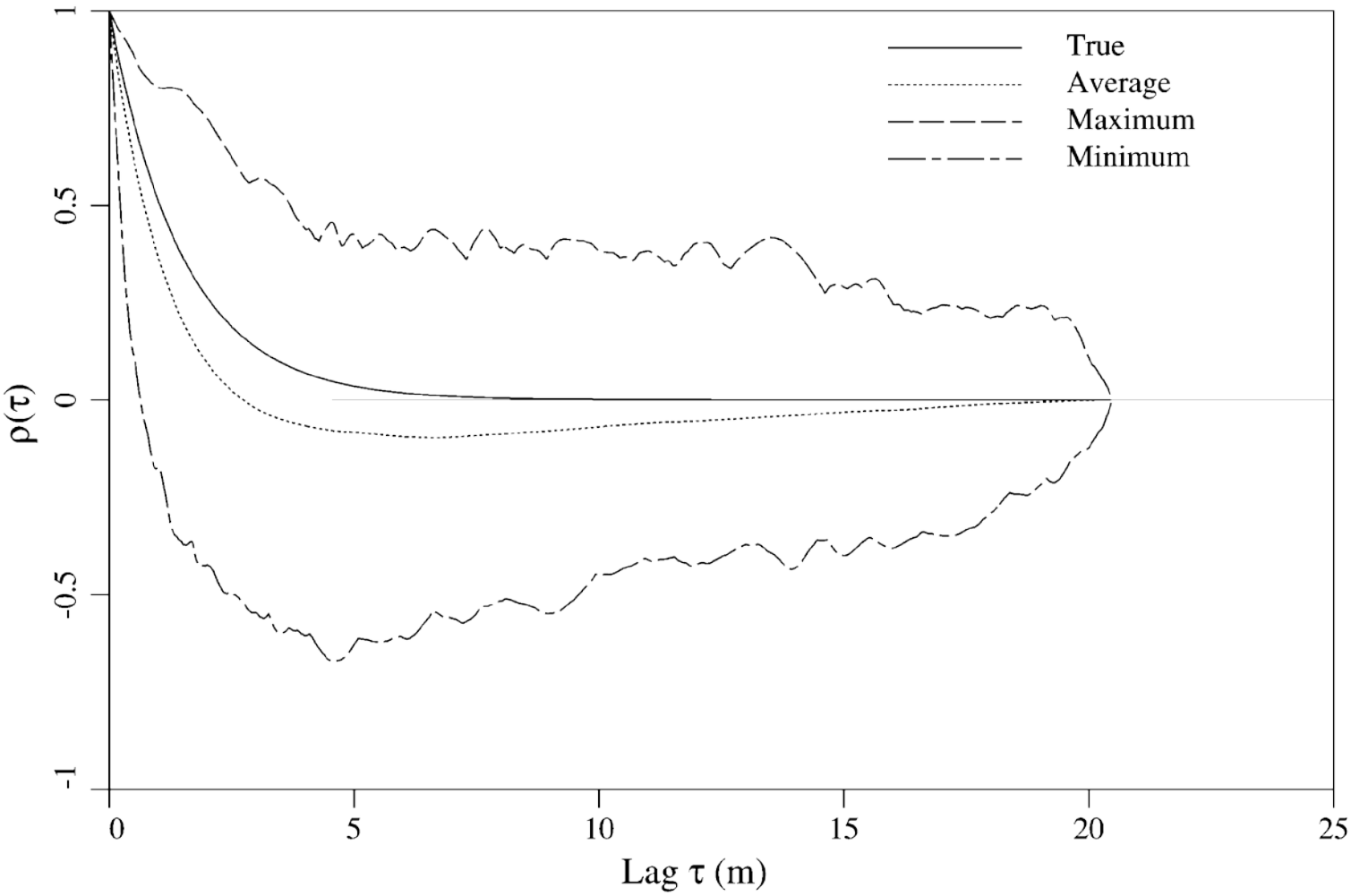
Bias: $$\mathrm{E}\left[\hat{C}\left(\tau_j\right)\right] \simeq \sigma_X^2 \left(\frac{n-j+1}{n}\right)\left[\rho\left(\tau_j\right)-\gamma\left(D\right)\right]$$

$$\mathrm{E}\left[\hat{\rho}\left(\tau_j\right)\right] \simeq \left(\frac{n-j+1}{n}\right)\left[\frac{\rho\left(\tau_j\right)-\gamma\left(D\right)}{1-\gamma\left(D\right)}\right]$$

Note that in a strongly correlated field, $\hat{\rho}\left(\tau_j\right)$ will become negative, often at about the field midpoint.

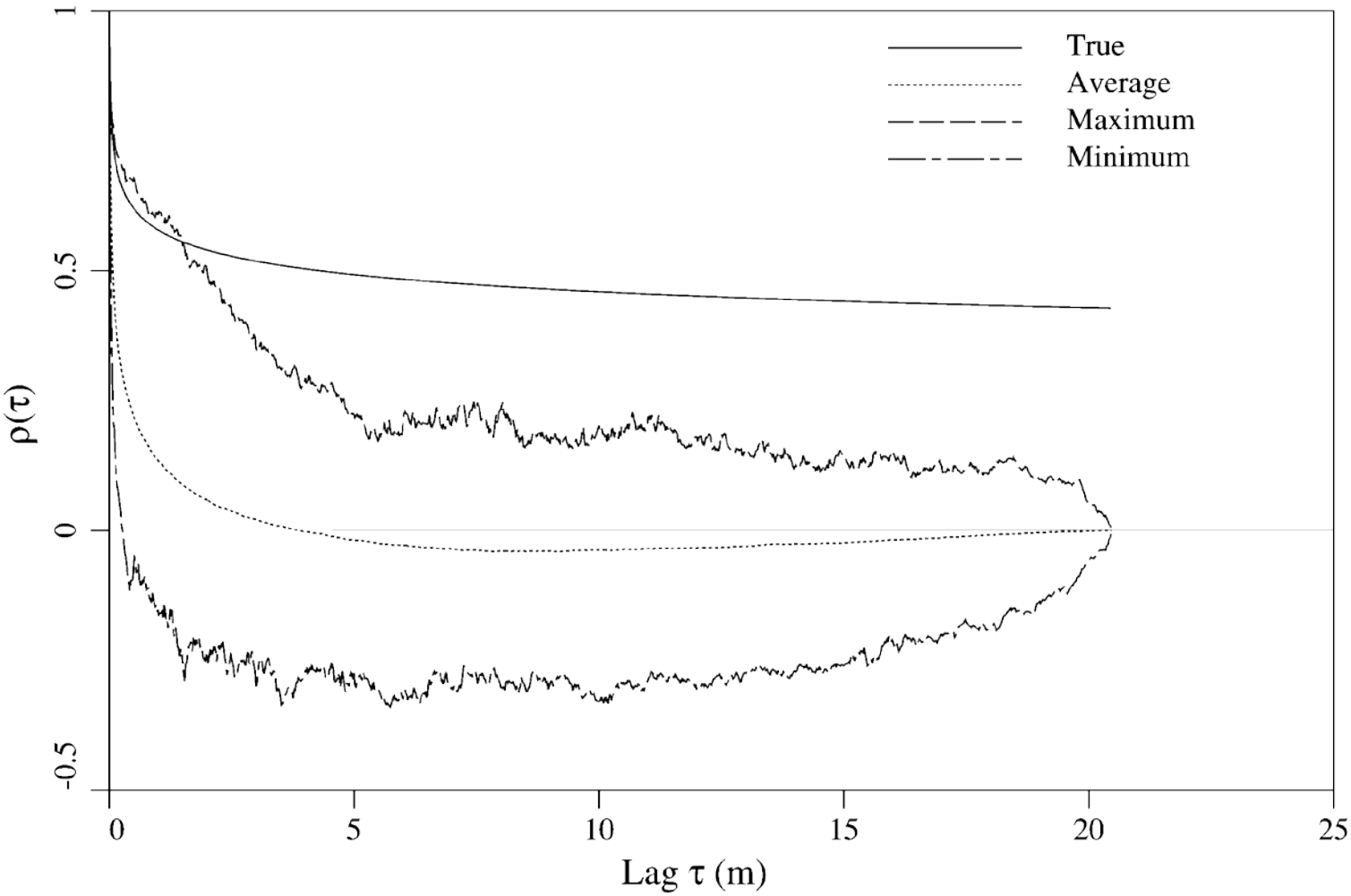


Estimating the Covariance Structure



Correlation function estimates from a finite-scale process ($\theta = 3$)

Estimating the Covariance Structure



Correlation function estimates from a fractal process (H = 0.95)

The Sample Semivariogram

The semivariogram gives essentially the same information as the correlation function since they are related according to

$$V(\tau_j) = \frac{1}{2}\mathrm{E}\left[\left(X_{i+j} - X_i\right)^2\right] = \sigma_X^2\left[1 - \rho(\tau_j)\right]$$

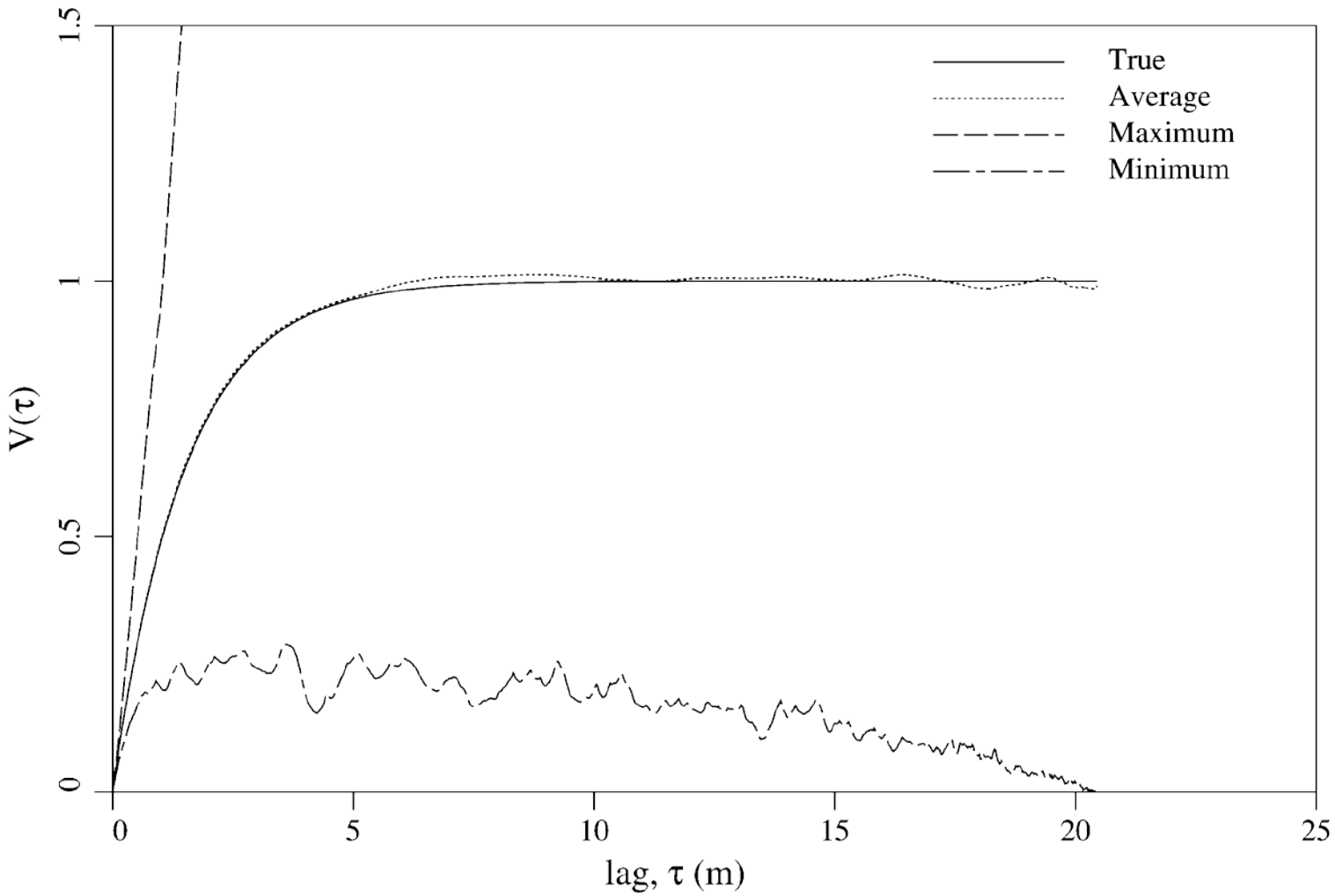
Its estimator is

$$\hat{V}(\tau_j) = \frac{1}{2(n-j)}\sum_{i=1}^{n-j}\left(X_{i+j} - X_i\right)^2, \qquad j = 0,1,\ldots,n-1$$

This estimator does **not** depend on $\hat{\mu}_X$, which is a *significant* advantage. In particular, it means it is unbiased,

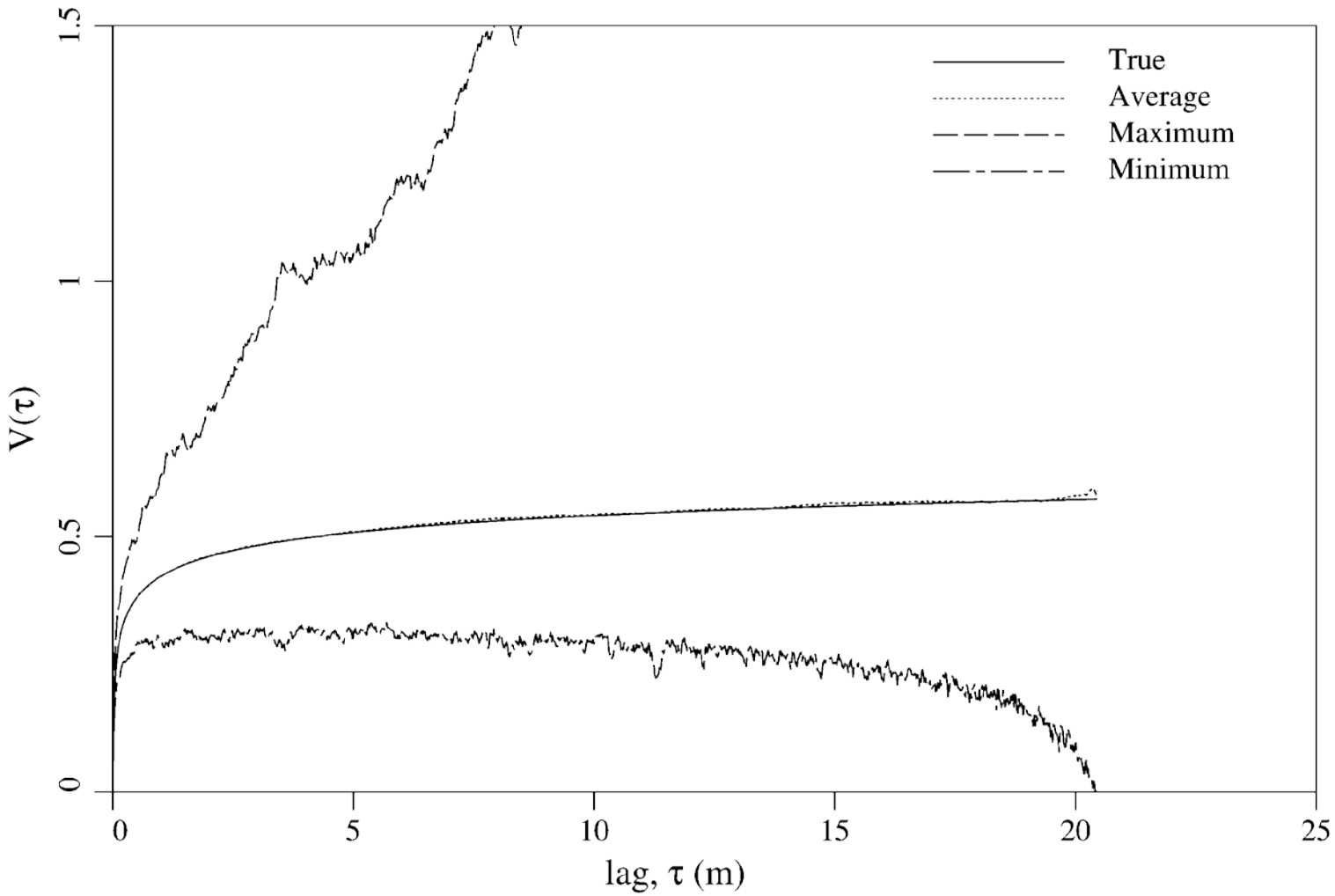
$$\mathrm{E}\left[\hat{V}(\tau_j)\right] = \frac{1}{2}\mathrm{E}\left[\left(X_{i+j} - X_i\right)^2\right]$$

The Sample Semivariogram



Semivariogram estimates from a finite-scale process ($\theta = 3$)

The Sample Semivariogram



Semivariogram estimates from a fractal process (H = 0.95)

Conclusions

- the mean is relatively easy to estimate at a site, the variance less easy, and the correlation length is very hard to estimate.
- interpolation is generally more accurate than extrapolation due to correlation between observations
- use caution when using statistics from the literature – these are generally unconservative due to correlation
- account for trends when interpolating, but not usually when extrapolating
- when estimating the correlation structure
  - correlation function estimates can be highly biased
  - the variogram is approximately unbiased.