Lecture 1: Introduction. Basic probability

Math 586

Course objectives: (after Allan Gutjahr)

- to model variations that occur in space
- to examine applicable data analysis methods
- to develop procedures for estimation
- to study methods that can recreate variations that may occur

Course outline:

- review of probability/statistics
- methods for variability of independent observations (ANOVA, regression)
- random fields and variograms
- kriging (spatial prediction)
- stochastic simulation

Geo. data:

- considered as random variables sampled over space (2-d, 3-d) or time (1-d) or both.
- usually correlated (classical statistics: independent)
- point observation, block observation; hard and soft data.

Software:

- Matlab for low-level matrix routines, calculations etc.
- GSLIB - a collection of Fortran-based routines (see Deutsch and Journel, 1998)
- `geoR` package for R

Literature:

- Kitanidis, *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, 1997
- Caers, *Petroleum Geostatistics*, SPE, 2005
- Chilès and Delfiner, *Geostatistics: Modeling Spatial Uncertainty*, Wiley, 1999.
- Deutsch and Journel, *Gslib: Geostatistical Software Library and User's Guide*, Oxford University Press, 2nd ed., 1997. '
- Isaaks and Srivastava, *Applied Geostatistics*, Oxford University Press, 1997
- Ribeiro and Diggle, *Model-based Geostatistics*, Springer, 2007.
- Wackernagel, *Multivariate Geostatistics* Springer, 3rd ed., 2003
- Wikle, Zammit-Mangion and Cressie, *Spatio-Temporal Statistics with R* Chapman & Hall/CRC, 2019;
  free download at `https://spacetimewithr.org`

Considering both structure and randomness

Consider a spatial process, say the thickness of a geological unit, denoted by $U(x)$, where U is a random variable representing thickness and x is spatial location, in Cartesian coordinates. Suppose you observe thickness at some location x_0, what can you say about the thickness at some nearby location, say $x_0 + h$?

$$\text{Say, given } U(x_0) = 2m, \quad U(x_0 + 1m) = ?$$

Some relationship is to be expected. It is both random and "predictable".

Our aim is to characterize and explain variation and use it to

- Predict, extrapolate and/or interpolate using its statistical features, and
- Reconstruct plausible "histories", "images" or realizations of variables, including the ones honoring the observations.

Probability Review - one variable

- Random variables (r.v.) - "chance magnitude"
  Use capital letters $(X, Y, V, ...)$ for r.v.'s and lowercase letters $(x, y, v, ...)$ for their particular (e.g. observed) values.

- Descriptors:
  - Continuous vs. discrete
  - Probability density (PDF) vs. cumulative distribution functions (CDF): $f(x) = dF(x)/dx$ (continuous)
  - Single r.v. (marginal distribution) vs. Joint

- Example
  Consider an exponential model of travel time of a particle in the environment

  Let T=time (say, in hours) that a particle spends in a mobile phase. Then

$$\text{PDF:} \quad f(t) = \begin{cases} \frac{1}{100} \exp\left(-t/100\right) & t \geq 0 \\ 0 & t < 0 \end{cases}$$

  Probability that mobile phase time exceeds 75 hours is

$$P(T > 75) = \int_{75}^{\infty} \frac{\exp\left(-t/100\right)}{100}\, dt = \exp\left(-75/100\right) = 1 - F(75) = 0.472,$$

  where the CDF $F(t) = P(T \leq t) = 1 - \exp\left(-t/100\right)$.
  Probability that mobile phase time is between 50 and 150 hours is

$$P(50 < T < 150) = \int_{50}^{150} \frac{\exp\left(-t/100\right)}{100}\, dt = F(150) - F(50) = 0.383$$

- Commonly used r.v.'s:
  Uniform, Exponential, Normal(and Lognormal), Binomial

- Summary measures: **Expectation**, $\mathbb{E}\left[\cdot\right]$
  A theoretical average.

$$\boxed{\begin{array}{l} \mathbb{E}\left[X\right] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \quad \text{(continuous)} \\ \qquad\qquad\qquad \mathbb{E}\left[X\right] = \sum_{\texttt{all}\ i} x_i P(X = x_i) \quad \text{(discrete)} \end{array}}$$

  Expectation of a function:

$$\boxed{\begin{array}{l} \mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx \quad \text{(continuous)} \\ \qquad\qquad\qquad \mathbb{E}\left[g(X)\right] = \sum_{\texttt{all}\ i} g(x_i) P(X = x_i) \ \text{(discrete)} \end{array}}$$

  provided that the integral or sum exists. Note that $\mathbb{E}\left[g(X)\right] \neq g(\mathbb{E}\left[X\right])$.

– Example:

Let $R =$ rainfall (m) in a region of the Earth.

$$\text{PDF:} \quad f(r) = \begin{cases} 6r(1-r) & 0 \le r \le 1 \\ 0 & \texttt{otherwise} \end{cases}$$

The expected rainfall is

$$\mathbb{E}[R] = \int_0^1 r \cdot f(r)\, dr = \int_0^1 r \cdot 6r(1-r)\, dr = 0.5m$$

Suppose that crop yield (metric tons/hectare) $= 70\sqrt{r}$. Then the expected yield is

$$\mathbb{E}[70\sqrt{R}] = \int_0^1 70\sqrt{r} \cdot f(r)\, dr = \int_0^1 70\sqrt{r} \cdot 6r(1-r)\, dr =$$

$$= 70 \cdot 6 \cdot \left[\frac{2}{5} r^{5/2} - \frac{2}{7} r^{7/2}\right]_0^1 = 48 \text{ tons}$$

This is not the same as $70\sqrt{\mathbb{E}[R]}$.

Question: what is the probability that yield is below 50 tons?

– *Moments* of X: n-th moment is $\mathbb{E}[X^n]$.

– Properties of expected values:

1. $\mathbb{E}[aX+b] = a\mathbb{E}[x] + b$

2. $\mathbb{E}[X_1 + X_2 + ... + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_n]$
   "Expectation of a sum is the Sum of expectations."

3. If $X_1, ..., X_n$ are statistically independent and $g_1(x_1), ..., g_n(x_n)$ are functions then

   $\mathbb{E}[g_1(X_1) \cdot g_2(X_2) \cdot ... \cdot g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot \mathbb{E}[g_2(X_2)] \cdot ... \cdot \mathbb{E}[g_n(X_n)]$