GEOG 176A: Introduction to Geographic Information Systems

Lecture 11: Spatial Analysis II

(chapter 6)

Rui Zhu
Standard normal distribution

Practice question: the precipitation (in inches) of all cities in a country follows N(40, 10). What is the probability that a city may have a precipitation larger than 60 inches?
Analysis on two attributes (bivariate)

- Certain relationships may exist between two attributes
  - E.g., elevation and temperature
  - E.g., soil PH values and the growth of crops
- We want to verify and find out such relationships based on data
- A handy tool: **scatter plot**
Analysis on two attributes (bivariate)

- **Pearson’s correlation coefficient**: a measurement to evaluate the linear relation between two attributes.

\[
\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]

Most of GIS, and many other statistical softwares, will do it for you! No need to memorize it for this class.
Analysis on two attributes (bivariate)

- Interpretation of Pearson’s correlation coefficient
  - Pearson’s correlation coefficient ranges in \([-1, 1]\)
  - \(-1\): a perfect **negative linear relationship** between the two attributes
  - \(0\): **no linear relationship** between the two attributes
  - \(1\): a perfect positive linear relationship between the two attributes
Analysis on two attributes (bivariate)

- Correlation does not mean causation!

E.g., strong correlation between chocolate consumption and Nobel prizes
From exploratory to inferential study

- So far, what we have done are called **exploratory analysis** (or descriptive statistics)
- Describe what the data look like:
  - Provide some **basic metrics**: mean, median, std, …
  - Provide some **graphics**: histograms, scatter plots, box plots, …
  - Describe the **relationship** between two attributes
- Another major analysis is **inferential analysis** (or inferential statistics)
Inferential analysis on attribute data

- A **population** represents all the targets to be investigated
- A **sample** is a set of values taken from the population
- Most geographic data we have is only a **sample** of the target **population**
- Example:
  - We want to investigate the water consumption of all US cities
  - **Population**: all cities in the US
  - **Sample**: 200 cities surveyed
Inferential analysis on attribute data

- **Parameters**: a characteristic of the population, e.g., population mean $\mu$
- **Statistics**: a characteristic of the sample, e.g., sample mean $\bar{x}$
- **Inferential analysis**: estimate a parameter of a population based on the statistics of sample data
Inferential analysis on attribute data

- The word “statistics” have multiple meanings
  - The **domain and procedures** for analyzing data
  - A **metric** derived from the sample data, e.g., mean derived from the sample is a statistic of the population

- **Geospatial statistics**: the sub-field in GIScience which integrates statistical methods in analyzing geographic data.
  - Spatial regression
  - Spatial autocorrelation
  - Spatial interpolation
  - Spatial simulation
  - Geostatistics
  - Network analysis
Inferential analysis on attribute data

Two forms of inferential analysis

- Estimation of parameters
- Hypothesis testing
Inferential analysis on attribute data (Estimation)

- **Estimation**: using a statistic from a data sample to estimate a parameter of the population
  - Mean of the sample $\rightarrow$ mean of the population:
    \[
    \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \rightarrow \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}
    \]
  - Std of the sample $\rightarrow$ std of the population:
    \[
    s_{n-1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \rightarrow \quad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}
    \]
Inferential analysis on attribute data (Estimation)

- You can estimate a parameter of the population using 100 sample records
- You can also estimate a parameter of the population using 10 sample records
- What’s the difference?
Inferential analysis on attribute data (Estimation)

- **Degree of freedom (df):** the amount of information used to estimate a parameter of a population
- **How to calculate df?**
  - The number of sample records minus the number of parameters which have to be estimated during the intermediate steps
  - E.g., if you have $N$ sample records, the df of estimating population mean is $N$
  - E.g., if you have $N$ sample records, the df of estimating population variance or std is $(N-1)$
- **Different numbers of sample records bring different dfs for the estimation**
Inferential analysis on attribute data (Estimation)

- **Confidence interval:**
  - A more scientific approach to estimate parameters with uncertainty considered.

- **Formula:**
  - $N > 30$: \[ \bar{X} - \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + \frac{s}{\sqrt{N}} \]
  - $N \leq 30$: \[ \bar{X} - t_{N-1} \cdot \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{N-1} \cdot \frac{s}{\sqrt{N}} \]
    - The value of $t$ and $z$ can be acquired from the $t$ and $z$ tables, based on the confidence threshold you defined such as 90%, and 95%.

  - The $t$-distribution incorporates the fact that for smaller sample sizes the distribution will be more spread using degree of freedom. For every change in degree of freedom, the $t$-distribution changes. The larger the sample size, the closer the $t$-distribution mimics the $z$-distribution in shape.
Examples of Estimation

To estimate the average soil PH value of a big farm, you decide to collect some samples:

**Trial 1:** you collected 50 samples, obtained a sample mean as 6 and a sample std as 0.5, you know the $t = 2.01$ when $df = 49$ with 95% confidence. What is the 95% confidence interval for the average PH?

$$
\bar{X} - t_{N-1} \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{N-1} \frac{s}{\sqrt{N}}
$$

$$
6 - 2.01 \times \frac{0.5}{\sqrt{50}} \leq \mu \leq 6 + 2.01 \times \frac{0.5}{\sqrt{50}}
$$

$$
6 - 2.01 \times 0.07 \leq \mu \leq 6 + 2.01 \times 0.07
$$

$$
5.86 \leq \mu \leq 6.14
$$

**Conclusion:** You have 95% confidence to say that the PH value is between [5.86, 6.14], when 50 samples were collected.
Examples of Estimation

To estimate the average soil PH value of a big farm, you decide to collect some samples:

Trial 2: you collected 5 samples, obtained a sample mean as 6 and a sample std as 0.5, you know the $t = 2.78$ when $df = 4$ with 95% confidence. What is the 95% confidence interval for the average PH?

$$\bar{X} - t_{n-1} \cdot \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{n-1} \cdot \frac{s}{\sqrt{N}}$$

$$6 - 2.78 \times \frac{0.5}{\sqrt{5}} \leq \mu \leq 6 + 2.78 \times \frac{0.5}{\sqrt{5}}$$

$$6 - 2.78 \times 0.22 \leq \mu \leq 6 + 2.78 \times 0.22$$

$$5.39 \leq \mu \leq 6.61$$

**Conclusion:** You have 95% confidence to say that the PH value is between [5.39, 6.61], when only 5 samples were collected.
From non-spatial analysis to spatial analysis ...
Spatial analysis

- Spatial descriptive analysis
  - Point pattern analysis
  - Spatial clustering analysis,
  - *Spatial associations* (Moran’s I, Geary’s C, semivariogram, …) → GEOG176B/Geog176C / GEOG172

- Spatial inferential analysis:
  - Regression models
  - Spatial interpolation (in terrain analysis)
Spatial analysis

- Spatial descriptive analysis
  - Point pattern analysis
  - Spatial clustering analysis,
  - Spatial associations (Moran’s I, Geary’s C, semivariogram, …) → GEOG176B/Geog176C / GEOG172

- Spatial inferential analysis:
  - Regression models
  - Spatial interpolation (in terrain analysis)
Descriptive analysis on spatial data

- Min, max and range of coordinates
- MinX, maxX, minY, maxY define the minimum bounding rectangle
Descriptive analysis on spatial data

Practice question:

- You have three points whose coordinates are in the form of (x, y): (1, 8), (4, 7) and (3, 9)
- What is the coordinate of the upper left corner of the minimum bounding rectangle of the three points.
Your tasks

- Finish reading Chapter 6
- Lab 3 due: Sunday, August 26th, 23:55 pm
- Review slides
- Paper about basic statistics on Gauchospace

Next week, we will continue discussing on spatial analysis.