Midterm scores

Mean: 74.74
Median: 76
Standard deviation: 11.15
GEOG 176A: Introduction to Geographic Information Systems

Lecture 11: Spatial Analysis II
(chapter 6)

Rui Zhu
Recall: Visualization

- **Data visualization**: to explore the data intuitively
- Typical visualizations:
  - Histogram
  - Bar chart, pie chart
  - Line chart
  - Box plot
  - Scatter plot
  - Radar plot
  - ...
Scatter plot

- Visualize the **potential relations** between two attributes
Radar plot

- Visualize **directional information**

Data distributions

Attribute data can be distributed in different ways:
Normal distribution

- A very common distribution
- Also called **Gaussian distribution**
- Widely used in various statistical methods/assumptions
Normal distribution

- A normal distribution is determined by two parameters:
  - Mean ($\mu$) and standard deviation ($\sigma$)
- If the data $X$ follows a normal distribution, then:
  $$X \sim N(\mu, \sigma)$$
- If $X \sim N(0, 1)$, $X$ follows a standard normal distribution
Standard normal distribution

- A standard normal distribution $X \sim N(0, 1)$ allows us to calculate the probability of different value ranges:
  - What is the probability for having values in $[0, 1]$?
  - What is the probability for having values in $[0, 2]$?
  - What is the probability for having values $> 1$?
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?
Example in ArcGIS
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 (bivariante)

- **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
Inferential analysis on attribute data

Two forms of inferential analysis

- Estimation of parameters
- Hypothesis testing
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: \[ \frac{X - z\cdot s}{\sqrt{N}} \leq \mu \leq \frac{X + z\cdot s}{\sqrt{N}} \]
  - N \leq 30: \[ \frac{X - t_{N-1}\cdot s}{\sqrt{N}} \leq \mu \leq \frac{X + t_{N-1}\cdot s}{\sqrt{N}} \]

  - 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 n 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?

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

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

\[
6 - 2.01 \times 0.07 \leq \mu \leq 6 + 2.01 \times 0.07
\]

\[
5.86 \leq \mu \leq 6.14
\]

**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?

$$
\bar{X} - t_{\nu, 0.05} \cdot \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{\nu, 0.05} \cdot \frac{s}{\sqrt{N}}
$$
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?

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

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

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

**Conclusion**: You have 95% confidence to say that the PH value is between [5.39, 6.61], when only 5 samples were collected.
So far, all we discussed are general statistics. How about spatial statistics?
Spatial analysis

- Spatial descriptive analysis
  - Point pattern analysis
  - Spatial clustering analysis
  - Spatial associations

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

What:
- You have detected clusters using a spatial clustering method (e.g., traffic congestion, hot locations visited by many tourists ...)

Why:
- But what reasons to see such a pattern? (e.g., there is a baseball game? Or the location is at a famous landmark?...