GIST 4302/5302: Spatial Analysis and Modeling

Basics of Statistics

Guofeng Cao

http://www.spatial.ttu.edu

Department of Geosciences
Texas Tech University

guofeng.cao@ttu.edu

Fall 2016

Outline of This Week

- Review basics of statistics and probability
- Learning pitfalls of spatial data

Basic Concepts in Statistics

Population vs. Samples

- Population: total set of elements/measurements that could be (hypothetically) observed in a study, e.g., all U.S. college students
- Sample: subset of elements/measurements from population, e.g., surveyed college students in Texas Tech

Population Parameters vs. Sample Statistics

- Parameters: summary measures that describe a population variable, e.g., average age of college students in Texas Tech.
- Statistics: summary measures that describe a sample variable, e.g., average age of surveyed Tech students

Statistical Procedures I

Statistical Sampling

- procedure of getting a representative sample of a population, e.g., a random visit of all U.S. colleges
- random sample: sample in which every individual in population has same chance of being included
- preferential sampling: sample in which certain individuals in population has higher chance of being included
- Law of large numbers and central limit theorem
  - Sample average should be close to the expected value given a large number of trials
  - Sample mean approaches the normal distribution *(under regular conditions)*

Statistical Procedures II

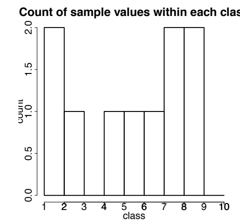
Statistics

- Descriptive statistics
  - procedure of determining sample statistics, e.g., determination of the average student age of all randomly visited colleges
- Statistical inference
  - procedure of making statements regarding population parameters from sample statistics, e.g., average student age of all randomly visited colleges = average age of college students in the U.S.?
- Statistical estimate
  - best (educated) guess about the value of a population parameter
- Hypothesis testing
  - procedure of determining whether sample data support a hypothesis that specifies the value (or range of values) of a certain population parameter

Histogram

- An Example: Consider a list of 10 hypothetical sample values:

2	2	9	8	7	9	5	6	8	3



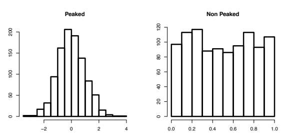
- Relative frequency table:
  $p_k = \#$ of data in k-th class/(total $\#$ of data)

k	1	2	3	4	5	6	7	8	9
p_k	0.2	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.0

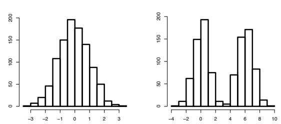
- *Please note:* Histogram shape depends on number and width of classes; rule of thumb for number of classes: $5 * log_{10}(\#$ of data$)$ and use non-overlapping equal intervals

Histogram Shape Characteristics

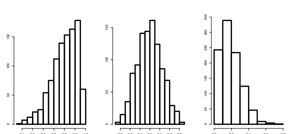
- Peaked or not



- Numbers of peaks



- Symmetric or not

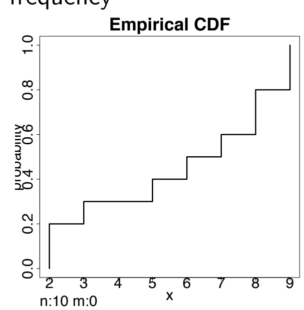


Cumulative Histogram

- Ranked sampled data and their relative frequency

k	1	2	3	4	5	6	7	8	9
p_k	0.2	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.0

- Cumulative relative frequency

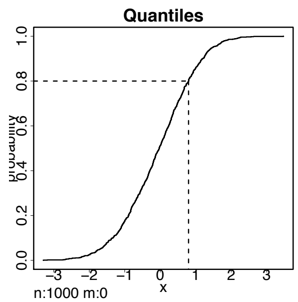


  - Proportion of sample values less than, or equal to, any given cutoff value
  - Probability that any random sample is no greater than and given cutoff value

Quantiles

Definition:

- datum value x_p corresponding to specific cumulative relative frequency value p



- Commonly used quantiles:
  - min: $x_{0.0}$, lower quantiles: $x_{0.25}$, median: $x_{0.50}$, upper quantile: $x_{0.75}$, max: $x_{1.00}$
  - Percentiles: $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$
  - Deciles: $x_{0.10}, x_{0.20}, \ldots, x_{0.90}$
- Quantiles are not sensitive to extreme values (outliers)

Measure of Central Tendency

- mid-range: arithmetic average of highest and lowest values: $\frac{x_{max}+x_{min}}{2}$
- mode: most frequently occuring values in data sets
- median: datum value that divides data set into halves; also defined as 50-th percentiles: $x_{0.5}$
- mean: arithmatic average of values in data set
  - sample mean: $m = \bar{x} = \frac{1}{n}\sum_{x=1}^{n} x_i$
  - population mean: $\mu = \frac{1}{N}\sum_{x=1}^{N} x_i$
  - sample mean is an esimation of population mean
- *Note:* Most appropriate measure of central tendency depends on distribution shapes

Measure of Dispersion I

- range: difference between highest and lowest values: $x_{max} - x_{min}$
- interquantile range (IQR): difference between upper and lower quantiles: $x_{0.75} - x_{0.25}$
- mean absolute derivation from mean: averange absolute difference between each datum value and the mean: $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
- median absolute derivation from median: median absolute difference between each datum value and the median: $|x_i - x_{0.5}|_{0.5}$
- variance: average squared difference between any datum values and the mean:
  - sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$
  - population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
  - sample variances is an estimate of the population variance

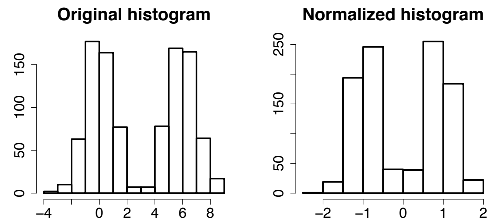
Measure of Dispersion II

- variance:
  - alternative definition: difference between average squred data and the mean squared
  - sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \cdot m^2$
  - population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$
  - Note: variance is expressed in squared data units
- standard deviation: square root of variance s or σ
  - unit of standard deviation is same as the data
- coefficient of variation: ratio of standard deviation and the mean
  - sample coefficient: $\frac{s}{m}$
  - population coefficient: $\frac{\sigma}{\mu}$
  - coefficient of variation is unitless
- choose alternative measures of dispersion:
  - any summary statistic involving squared values is senstive to outliers
  - any summary statistic based on quantiles is robost to outliers
  - coefficient of variation: very use for comparing spread of different data sets

Normalizing Data

Normalizing data to zero mean and unit variance allows more meaningful comparison of different data sets

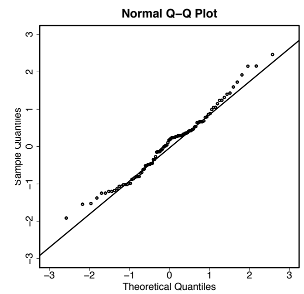
- Normalizing procedure:
  1. compute mean m and standard deviation s of data setA
  2. subtract the mean from each datum: $x_i - m$
  3. divide by the standard deviation: $z_i = \frac{x_i - m}{s}$
- normlized data are unit free; shape of distibution does not change (e.g., modes remain the same)



Quantile-Quantile (Q-Q) Plots

Graph for comparing the shapes of distribution

- Normalizing procedure:
  1. rank both data sets from smallest to largest values
  2. compute quantiles of each data set
  3. cross-plot each quantile pair

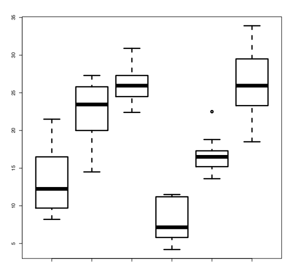


- Interpretation: straight plot aligned with $45°$ line implies two similar distribution

Boxplot

Graph for describing the the degree of dispersion and skewness and identify outliers

- Non-parametric
- 25%, 50%, *and* 75% percentiles
- end of the hinge (whisker) could mean differently; most ofen represent the lowest datum within 1.5 IQR of the lower quantile, and the highest datum still within 1.5 IQR of the upper quantile



- Points outside of range are usually taken as <u>outliers</u>

Statistical Experiment and Events

- Statistical experiment
  - process in which one outcome from a set of possible outcomes occurs (also known as random trial), e.g., sampling n data from a population is a collection of n statistical experiments
- Elementary outcome
  - the outcome E of a statistical experiment, e.g., age of a single student in GIST 4302, or rain on a particular day
- Event
  - a collection of k elementary outcomes $A = \{E_1, E_2, \ldots, E_k\}$ of interests, e.g., all male GIST 4302 students
- Random variables
  - Don't have single, fixed values; it can take on a set of possible different values, each with an associated probability

Probability II

- Relative frequency definition:
  - if a statistical experiment is repeated N times, and event A occurs in n of these trials, then the probability for A to occur is:
    $P(A) = Prob\{A\} = \frac{n}{N}$, as N tends to infinity
  - e.g., the probability for a wet-day event in Lubbock can be seen as the proportion of wet-days in a very long precipitation record
- Axioms of probability:
  - probabilities are necessarily non-negative: $P(A) \geq 0$
  - the sample space S will certainly occur: $P(S) = 1$
  - for two mutually execluive events A and B:
    $P(A \cup B) = P(A) + P(B)$
- Elementary probability theorem
  - the impossible event $\emptyset$ has zero probability of occurrence: $P(\emptyset) = 0$
  - the probability of the complement event $\bar{A}$ of an event A is:
    $P(\bar{A}) = 1 - P(A)$
  - the probability of an event A to occur cannot be greater than one:
    $P(A) \leq 1$
  - the probability of either two events A and B to occur is:
    $P(A \cup b) = P(A) + P(B) - P(A \cap B)$

Probability III

- Example: wet or dry day in the 10 days of period, wet day event $A = 1$, dry day event $A = 0$

day i	1	2	3	4	5	6	7	8	9	10
event a_i	1	1	0	0	1	1	0	0	0	0

$$P\{A = 1\} = \frac{4}{10} = 0.4$$

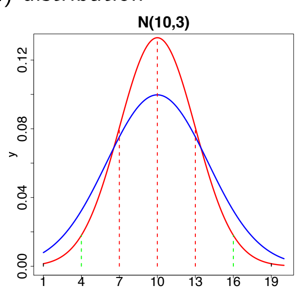
- Example: precipitation x in the 10 days of period, an binary event $a_i = 1$ if associated $x_i \leq 4$ otherwise $a_i = 0$

day i	1	2	3	4	5	6	7	8	9	10
precip i	6	0	3	0	0	2	1	5	6	5
event a_i	0	1	1	1	1	1	1	0	0	0

$$P\{a = 1\} = \frac{6}{10} = 0.6$$

Commonly Used Probability Distributions

- Gaussian (or normal) distribution



- The shapes are controlled by mean (μ) and variance (σ^2)
- Three sigma rule ($68 - 95 - 99.7$ rule)

Conditional Probability

- Example: Consider $n = 10$ days of precipitation (X) and daily max temperature (Y) records for a certain place:

day i	1	2	3	4	5	6	7	8	9	10
precip x_i	6	0	3	0	0	2	1	5	6	5
temp y_i	15	18	20	19	22	24	21	21	20	18

- Convert the previous records to binary events by $a_i = 1$ if $x_i > 0$, 0 if not and $b_i = 1$ if $y_i > 20$, 0 if not

day i	1	2	3	4	5	6	7	8	9	10
precip x_i	1	0	1	0	0	1	1	1	1	1
temp y_i	0	0	0	0	1	1	1	1	0	0

- Joint probability: $P(A, B) = \frac{1}{n} \sum_{i=1}^{10} a_i b_i = \frac{3}{10} = 0.3$
- Conditional probability: $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{0.3}{0.4} = 0.75$

Independence

Independence events

- two events A and B are independent iff: $P(A|B) = P(A)$
- knowledge of event B will not affect the probability of event A to occur
- in previous example, $P(A|B) = 0.75 \neq 0.7 = P(A)$

Alternatively

- two events A and B are independent iff: $P(A, B) = P(A)P(B)$
- joint probability $P(A, B)$ equals the product of individual occurence probability $P(A)P(B)$
- in previous example, $P(A, B) = 0.3 \neq 0.7 * 0.4 = P(A)P(B)$

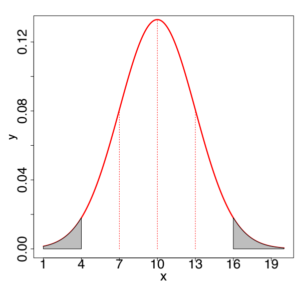
Covariance and Correlation Coefficient

Suppose X and Y are two random variables for a random experiment

- the *covariance* of X and Y measures how much these two random variables are related
  - $cov(X, Y) = E[(X - E(X)(Y - E(Y)))]$
- The *correlation coefficient* of X and Y a normalized version of covariance
  - $cor(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
- $cov(X, Y) = 0$ means X and Y are 'unrelated'

p-value

- Assuming the null hypothesis is true, the p-value is the probability a test statistics at least as extreme as the one that was actually observed



Spatial Versus Non-Spatial Statistics

Classical statistics

- samples assumed realizations of independent and identically distributed random variables (iid)
- most hypothesis testing procedures call for samples from iid random variables
- problems with inference and hypothesis testing in a spatial setting

Spatial statistics

- multivariate statistics in a spatial/temporal context: each observation is viewed as a realization from a different random variable, but such random variables are auto-correlated in space and/or time
- each sample is not an independent piece of information, because precisely it is redundant with other samples (due to the corresponding random variables being auto-correlated)
- auto- and cross-correlation (in space and/or time) is explicitly accounted for to establish confidence intervals for hypothesis testing

Some Issues Specific to Spatial Data Analysis

Spatial dependency

- values that are closer in space tend to be more similar than values that are further apart (Tobler's first law of Geography)
- redundancy in sample data = classical statistical hypothesis testing procedures not applicable
- positive, zero, and negative spatial correlation or dependency

The modified areal unit problem (MAUP)

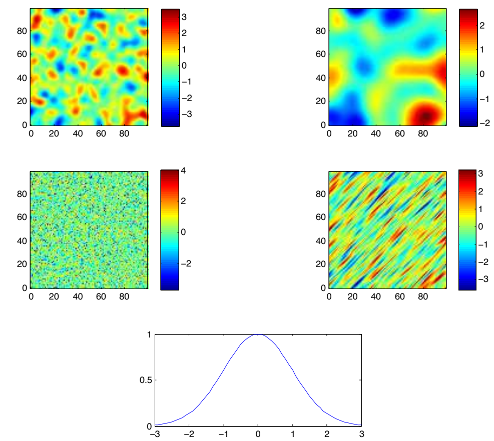
- spatial aggregations display different spatial characteristics and relationships than original (non-averaged) values
- scale and zoning (aggregation) effects

Ecological fallacy

- problem close related to the MAUP
- relationships established at a specific level of aggregation (e.g., census tracts) do not hold at more detailed levels (e.g., individuals)

Spatial Dependency (I)

- often termed as spatial similarity, spatial correlation and spatial pattern, spatial pattern, spatial texture ...
- Examples of synthetic maps with same histogram:



Spatial Dependency (II)

Spatial statistics

- inference of spatial dependency is the core of spatial statistics
  - spatial interpolation, e.g., kriging family of methods
  - spatial point pattern analysis
  - spatial areal units (regular or irregular)
- often extended into a spatio-temporal domain to investigate the dynamic phenomena and processes, e.g., land use and land cover changes

The Modified Areal Unit Problem

The same basic data yield different results when aggregated in different ways

- First studied by Gehlke and Biehl (1934)
- Applies where data are aggregated to areal units which could take many forms, e.g., postcode sectors, congressional district, local government units and grid squares.
- Affects many types of spatial analysis, including clustering, correlation and regression analysis.
- Example: *Gerrymandering* of congressional districts (Bush vs. Gore, Lincoln vs. Douglas)
- Two aspects of this problem: scale effect and zoning (aggregation) effect

The Modified Areal Unit Problem: Scale Effect (1)

Scale effect

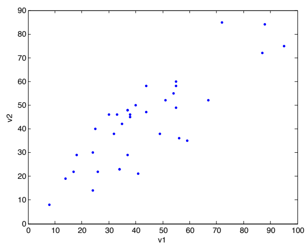
Analytical results depending on the size of units used (generally, bigger units lead to stronger correlation)

Example

Table: spatial variable #1 versus spatial variable #2

87	95	72	37	44	24	72	75	85	29	58	30
40	55	55	38	88	34	50	60	49	46	84	23
41	30	26	35	38	24	21	46	22	42	45	14
14	56	37	34	08	18	19	36	48	23	8	29
49	44	51	67	17	37	38	47	52	52	22	48
55	25	33	32	59	54	58	40	46	38	35	55

Table: $\rho(v1, v2) = 0.83$



The Modified Areal Unit Problem: Scale Effect (2)

Scale effect

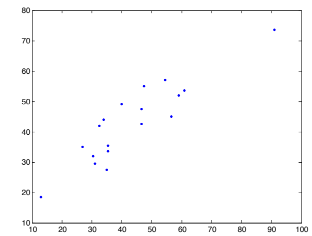
Analytical results depending on the size of units used (generally, bigger units lead to stronger correlation)

Example

Table: spatial aggregation strategy # 1

91.0	47.5	35.5	73.5	55.0	33.5
35.0	46.5	40.0	27.5	42.5	49.0
54.5	46.5	30.5	57.0	47.5	32.0
35.5	59.0	32.5	35.5	52.0	42.0
34.0	61.0	31.0	44.0	53.5	29.5
13.0	27.0	56.5	18.5	35.0	45.0

Table: $\rho(v1, v2) = 0.90$



The Modified Areal Unit Problem: Zoning Effect

Zoning effect

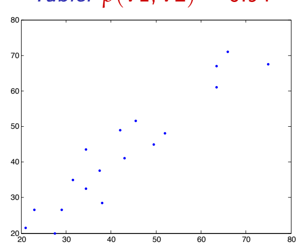
Analytical results depending on how the study area is divided up, even at the same scale

Example

Table: spatial aggregation strategy #2

63.5	75	63.5	37.5	66	29.0	61.0	67.5	67.0	37.5	71.0	26.5
27.5	43	31.5	34.5	23	21	20.0	41.0	35.0	32.5	26.5	21.5
52.0	34.5	42	49.5	38.0	45.5	48.0	43.5	49.0	45.0	28.5	51.5

Table: $\rho(v1, v2) = 0.94$



The Modified Areal Unit Problem: Zoning Effect

Zoning effect: another example

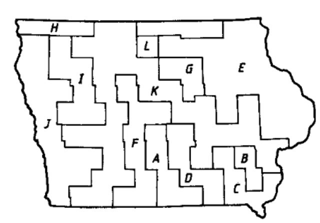


Figure 2a. Zoning system that minimises the regression slope coefficient (-24, r = -.25)

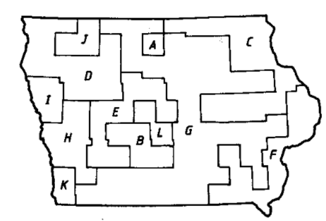


Figure 2b. Zoning system that maximises the regression slope coefficient (12, r = .87)

Figure: Image Courtesy of OpenShaw

Ecological Fallacy (I)

- relationships established at a specific level of aggregation do not hold at more detailed levels

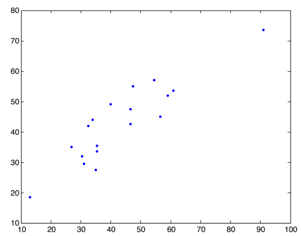
Example

Table: spatial aggregation strategy # 1

91.0	47.5	35.5
35.0	46.5	40.0
54.5	46.5	30.5
35.5	59.0	32.5
34.0	61.0	31.0
13.0	27.0	56.5

73.5	55.0	33.5
27.5	42.5	49.0
57.0	47.5	32.0
35.5	52.0	42.0
44.0	53.5	29.5
18.5	35.0	45.0

Table: $\rho(v1, v2) = 0.90$



Ecological Fallacy (II)

- relationships established at a specific level of aggregation do not hold at more detailed levels

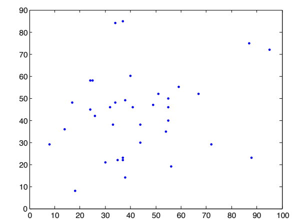
Example

Table: spatial variable #1 versus spatial variable #2

95	87	37	72	24	44
55	40	38	55	34	88
30	41	35	26	24	38
56	14	34	37	18	08
44	49	67	51	37	17
25	55	32	33	54	59

72	75	85	29	58	30
50	60	49	46	84	23
21	46	22	42	45	14
19	36	48	23	8	29
38	47	52	52	22	48
58	40	46	38	35	55

Table: $\rho(v1, v2) = 0.21$



Stages in Spatial Statistical Analysis

Exploratory analysis

- explore spatial data using cartographic (or other visual) representations
- statistical analysis for detecting possible sub-populations, outliers, trends, relationships with neighboring values or other spatial variables

Modeling or confirmatory analysis

- establish parametric or non-parametric model(s) characterizing attribute spatial distribution
- *estimate* model parameters from data; evaluate their statistical significance; *predict* attribute values at other locations and/or future time instants

Notes

- boundaries between above stages not always clear-cut, and it is often an iterative process

Software for Statistical Analysis of Spatial Data

GIS-based packages

- ESRI's Spatial Analyst, Geostatistical Analyst, Spatial Statistics
- opt for "close" or "loose" coupling with specialized external packages when specific functionalities are missing from a GIS

Statistical packages

- R packages, Matlab (new class will be available this Fall!)
- GeoDa/PySAL
- versatile in modeling, programable

Acknowledgement

- Some slides of the the materials are based on Dr. Phaedon Kyrikidis's classes in University of California, Santa Barbara