GEOG 413/613

LECTURE 1

Housekeeping
Syllabus
Office Hours
Laptops

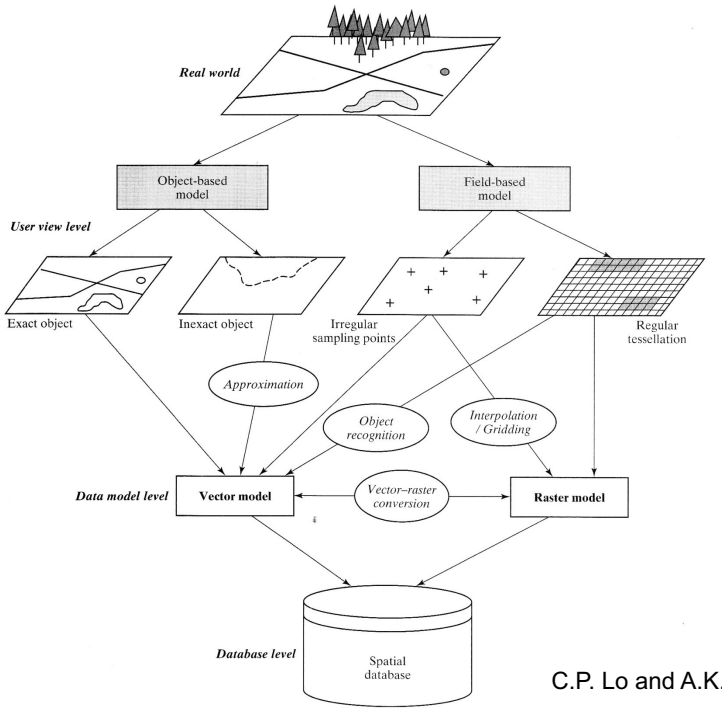
Data and Introductory Stats

Geospatial Data

- Real world features are represented in two basic forms
  - Objects
    - discrete, definite boundaries, persistent identity
    - E.G. Highways, buildings, parks, administrative regions
  - Phenomena
    - Distributed continuously over a large area
    - E.G. terrain elevation, temperature, rainfall, soil pH

Geospatial Data Models

- Objects
  - The Object-based model
    - Points, line, polygons, volumes
    - Persistent Identity
    - Identifiable boundary/spatial extent
    - Has attribute(s)
    - Representative/relevance of some entity
- Fields
  - The Field-based model
    - Grid
    - Tessellation of space
    - A single value for each unit of space
    - There is a value everywhere
    - Fields can be used for discrete phenomena (has implications for analysis)
      - Continuous – a function maps a smooth variable across space (e.g. elevation)
      - Discrete – space is mutually exclusive, cells in a one part are similar (e.g. land use)



C.P. Lo and A.K.W. Yeung

Geospatial Data Models

- Cognition of geographic space varies with scale
- The conceptualization of geographic space is influenced by the purpose as well as the methods of data collection
- How phenomena is represented determines what can be modeled and the information queried in a GIS

Geospatial Data Models

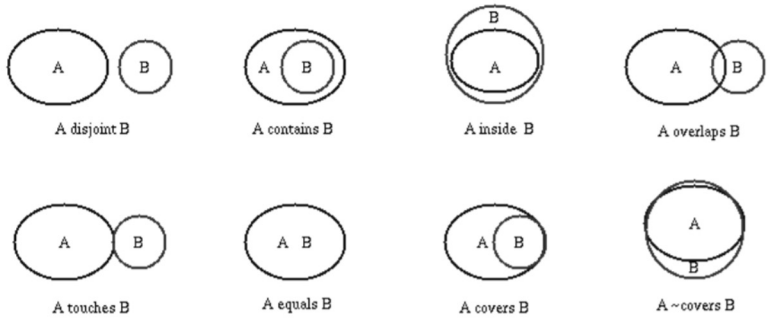
- Geographic ontology by definition encompasses the conceptualization of how the space-time relationship is specified.
  - Spatial Objects and their Relationships
    - Geographic Features
      - Crisp /Fuzzy Boundaries
      - Continuous fields
    - Relationships
      - Topological relationships (Egenhofer, Allen, Clementin etc)
      - Projective relationships
      - Space-time relationships

Geospatial Data Models

- Geographic ontology by definition encompasses the conceptualization of how the space-time is specified.
  - Spatial Objects and their Relationships
    - Semantics
      - Semantic Properties
      - Semantic Relationships

Geospatial Data Models

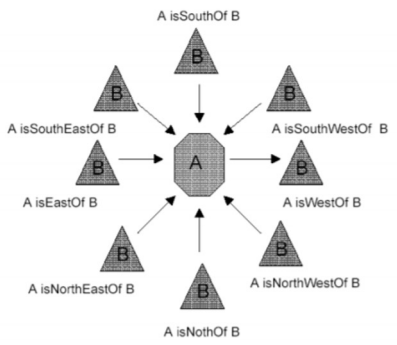
- Egenhofer Topological Relationships



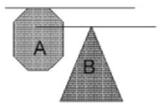
R. Laurini

Geospatial Data Models

- Projective Relationships



e.g. "*A isNorthOf B*"

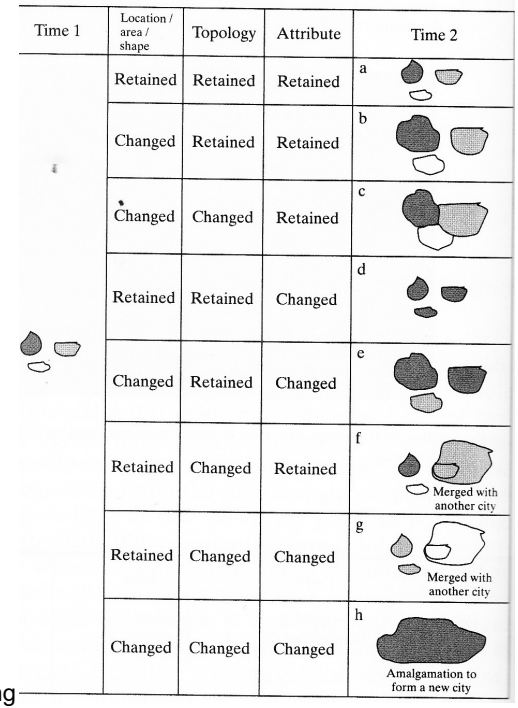


True if northern most point of A is north of the northern most point of B

R. Laurini

Geospatial Data Models

Space-time relationships

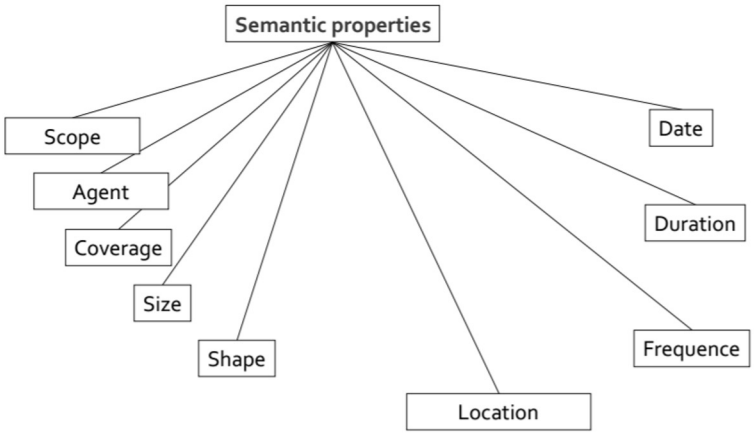


Time 1	Location / area / shape	Topology	Attribute	Time 2
	Retained	Retained	Retained	a
	Changed	Retained	Retained	b
	Changed	Changed	Retained	c
	Retained	Retained	Changed	d
	Changed	Retained	Changed	e
	Retained	Changed	Retained	f Merged with another city
	Retained	Changed	Changed	g Merged with another city
	Changed	Changed	Changed	h Amalgamation to form a new city

C.P. Lo and A.K.W. Yeung

Geospatial Data Models

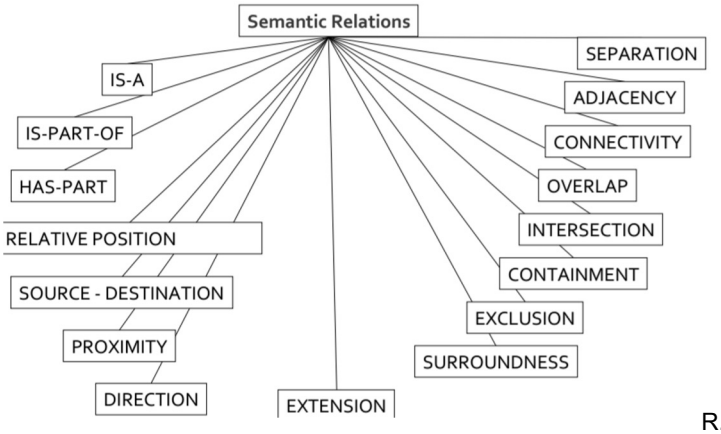
- Semantic Properties



R. Laurini

Geospatial Data Models

- Semantic Relationships



R. Laurini

Geospatial Data Models

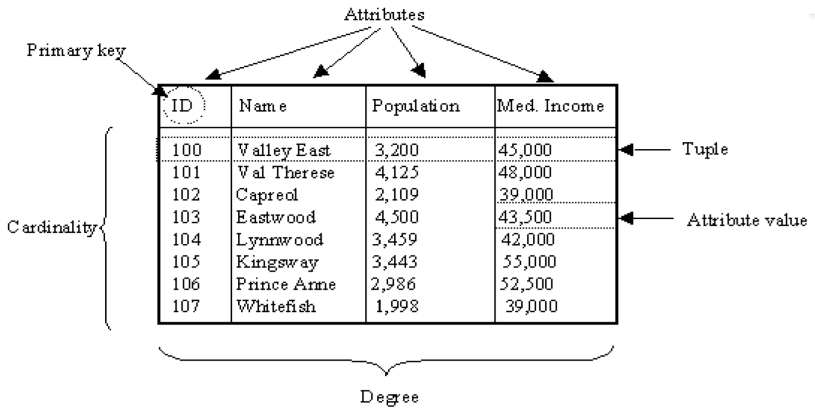
- Data Structures are specific formats for storing and organizing geographic data.
  - raster data structures represent geography as a finite grid of cells
  - vector data structures represent geography with coordinate lists (vector lines)
- Data Structures enable
  - Conversion between formats
  - Organization in a database
  - Modeling for analysis, visualization

Geospatial Data Models

- Ultimately the spatial data structures end up as *entities and their relationships* in a database hence the E/R model

Geospatial Data Models

Figure 13: Characteristics of a relational table

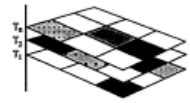


ID	Name	Population	Med. Income
100	Valley East	3,200	45,000
101	Val Therese	4,125	48,000
102	Capreol	2,109	39,000
103	Eastwood	4,500	43,500
104	Lynnwood	3,459	42,000
105	Kingsway	3,443	55,000
106	Prince Anne	2,986	52,500
107	Whitefish	1,998	39,000

A.K. Yeung 1998-10-10 u51-13

Geospatial Data Models

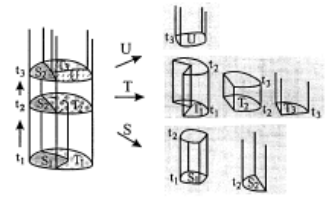
Representations of space-time information in a GIS (Yuan 1999)



a. Time-stamped layers (Armstrong 1988).

Poly id	T_1	T_2	T_3	T_4
1	Rural	Rural	Rural	Rural
2	Rural	Urban	Urban	Urban
3	Rural	Rural	Urban	Urban
4	Rural	Rural	Urban	Urban
5	Rural	Rural	Rural	Urban

b. Time-stamped attributes (columns): Space-Time Composites (Langran and Chrisman 1988).



ST-objects modeling regional change

Decomposition of ST-objects (U, T, and S) into 6 ST-atoms (U, T_1, T_2, T_3, S_1, and S_2).

Agriculture Urban Industry

c. Time-stamped space-time objects: the spatiotemporal object model (Worboys 1992).

Geospatial Data Models

- The raster and vector geospatial data models assume
  - Time is static
    - Suited to a temporal snapshot of geographic space
    - Time is another attribute of space
  - 2D but not 3D representation of space
    - 2.5D
    - True 3D would require x,y,z coordinates to display

Geospatial Data Models

- Cognition of geographic space varies with scale
- The conceptualization of geographic space is influenced by the purpose as well as the methods of data collection
- The choice of representation determines what can be modeled and the information queried in a GIS

Geospatial Data Models

- A case for a Unified Data Model
  - Raster, vector each has it's own relative merits
  - 3D representation
  - Spatial, Temporal & Semantic Domains

Geospatial Data Models

- The Geo-atom as an example of a unified data model
  - A geo-atom is defined as an association between a point location in space–time and a property. For example a weather station. It also acts an abstraction for all geographic information
  - A geo-field represents the variation of a phenomenon over space-time. It is an aggregation of geo-atoms, where each geo-atom defines the same set of properties.

Geospatial Data Models

- A geo-object represents an aggregation of geo-atoms which meet certain requirements, such as having specified values for certain properties.
- The field-object as a geo-object with internal heterogeneity conceptualized as a field. E.g. a wildfire or a storm may have a boundary and an internal structure defined by the variation of such field-like properties as ambient temperature or rainfall.

Geospatial Data Models

- Object fields here each geo-atom maps to a geo-object not a value. For example, mapping the visible area for every point on a topographic surface

Geospatial Data Models

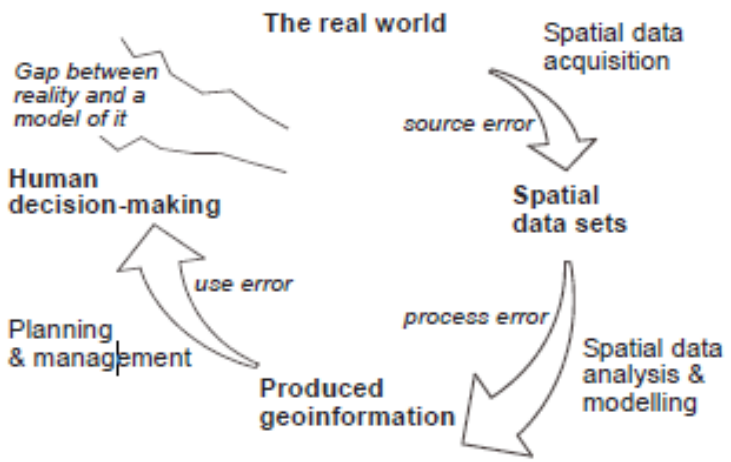
- The number of dimensions representing the space–time depends on the application
  - many GIS applications are adequate even though they ignore the temporal dimension and/or the third spatial dimension
- However, there is also room for simplifying representation and defining an agreed upon ontology

Geospatial Data Models

- Besides representation, important aspects to consider here data quality and error propagation
  - Data Quality
    - Accuracy
      - Positional
      - Attribute
      - Temporal
    - Precision
    - Completeness
    - Logical Consistency
    - Lineage

Geospatial Data Models

Error propagation in spatial data handling



Otto. Huisman & Rolf A . de By, 2009

Geospatial Data and the Geographer

- Geography attempts to address problems from a spatial and ecological perspectives
  - Spatial: patterns and processes
  - Ecological: relationships between living and nonliving entities in geographic space
- Therefore, the geographer asks questions about
  - Where; Why; How; What to do;
    - examples?

28

Statistics in Geography

- Stats are used to answer a variety questions
  - Describe and summarize data
  - Generalizations
  - Estimates for likelihood
  - Inferences
  - Differences between locations
  - Patters differ from what is expected

29

Spatial Data

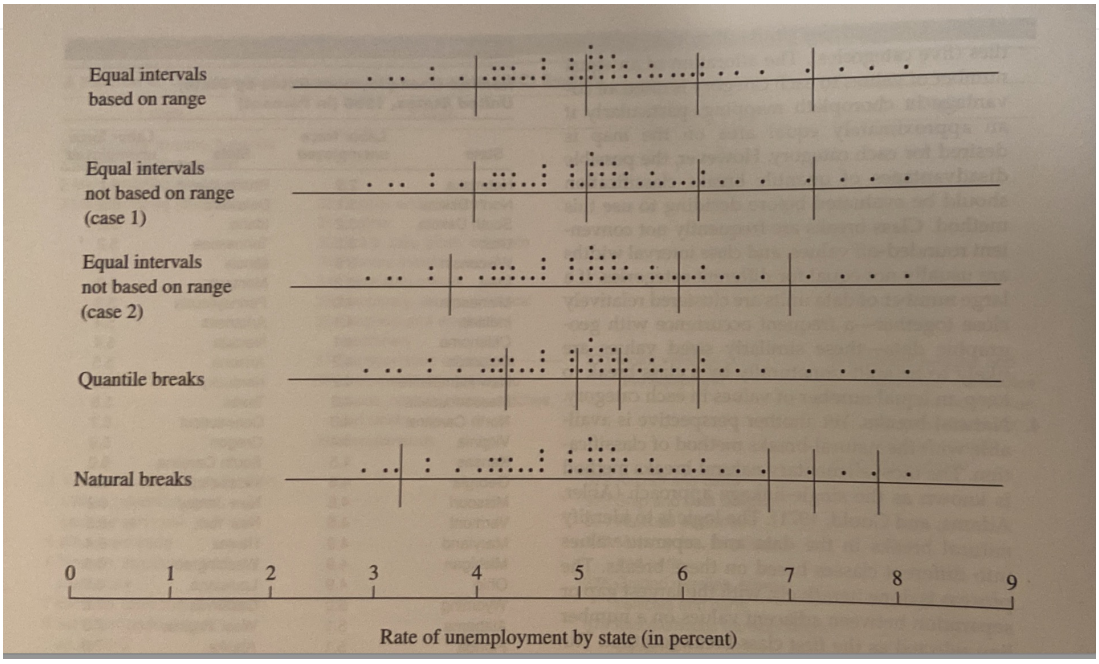
- Type
  - Primary/secondary
- Attribute/Variable
  - Continuous/discrete; Qualitative/quantitative
- Levels
  - Collection
    - Individual/Aggregated
  - Measurement
    - Nominal, Ordinal, Interval, Ratio

30

Spatial Data

- Measurement Concepts
  - Precision
  - Accuracy
  - Validity
  - Reliability
- Classification Methods
  - Equal intervals based on range
    - By dividing range (lowest - highest)
  - Equal intervals not based on range
    - Rounded off class breaks, arbitrary selection,
  - Quantile breaks
    - Commonly quartiles(4), quintiles(5)
  - Natural breaks
    - Natural separations between adjacent ranked values

31



McGrew and Monroe, 2000, An Introduction to Statistical Problem Solving in Geography. McGraw Hill Boston

32

Spatial Data

- Presentation
  - Histograms
  - Frequency tables
  - Scatter Plots
  - Line Graphs

33

Non-spatial Statistics

- Measures of Central Tendency
  - Mode: Most frequently occurring value
  - Median: middle value from a set of ranked observations
  - Mean
    - Arithmetic mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

X – arithmetic mean
n – number of observations
a_i – value of observation i

34

Non-spatial Statistics

- Measures of Dispersion
  - Deviation

$$d_i = (x_i - \bar{X})$$

  - Average Deviation

$$m = \frac{\sum |x_i - \bar{X}|}{n}$$

$|x_i - \bar{X}|$ => absolute value of the difference

35

Non-spatial Statistics

- Measures of Dispersion
  - Range
  - Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - X)^2}{n-1}}$$

standard deviation for sample (≈ n<30)

$$\delta = \sqrt{\frac{\sum_{i=1}^{n}(x_i - X)^2}{n}}$$

standard deviation for population (≈ n>30)

36

Non-spatial Statistics

- Measures of Dispersion
  - Variance

$$s^2 = \frac{\sum_{i=n}^{n}(x_i - X)^2}{n-1}$$

variance for sample (≈ n<30)

$$\delta^2 = \frac{\sum_{i=n}^{n}(x_i - X)^2}{n}$$

Variance for population (≈ n>30)

37

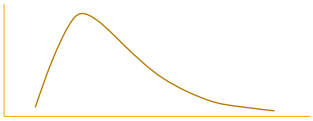
Non-spatial Statistics

- Measures of Dispersion
  - Coefficient of Variation (CV)
    - The standard deviation and the variance are absolute measures, i.e. their values are dependent on the magnitude of the units of measurement.
    - The coefficient of variation is a relative measure that addresses this

$$CV = \frac{s}{\bar{X}}$$

38

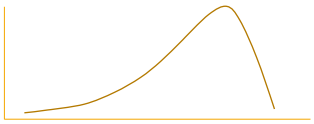
Non-spatial Statistics



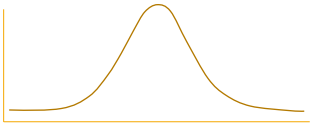
Positive Skewness, Skewness > 0

- Measures of Relative Position
  - Skewness

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^3}{ns^3}$$



Negative Skewness, Skewness < 0



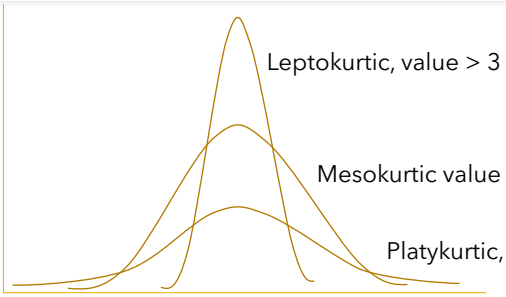
Normal, Skewness = 0

39

Non-spatial Statistics

- Measures of Relative Position
  - Kurtosis

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{ns^4}$$



40

Spatial Statistics

- Measures of Central Tendency
  - Mean Center
    - Arithmetic mean of a set of spatial objects (Centroid)
    - Mean Center $(\bar{x}, \bar{y})$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

41

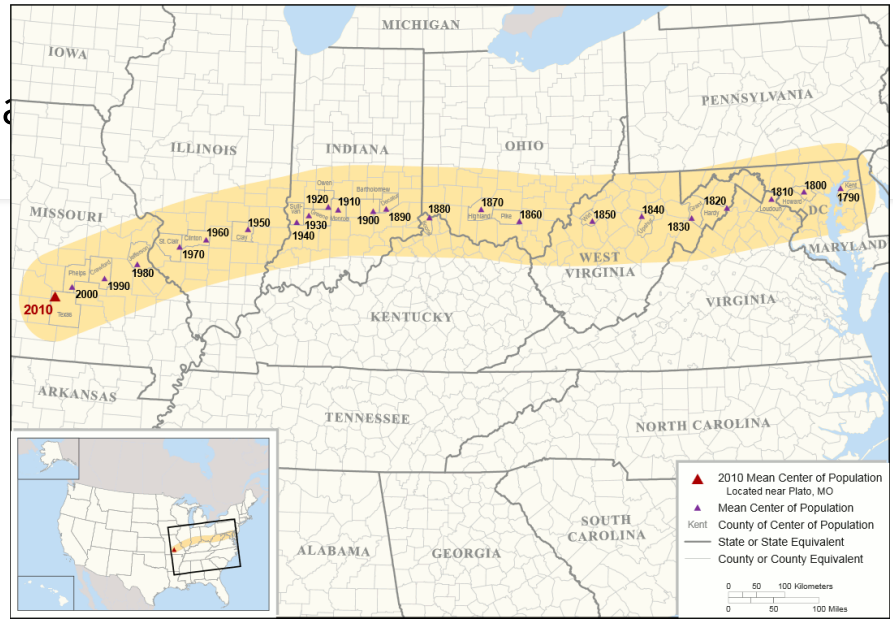
Spatial Statistics

- Measures of Central Tendency
  - Weighed Mean Center
    - Mean affected by a weight factor (e.g. frequency, population)
    - Represents the centre of gravity $(\bar{x}, \bar{y})$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=n}^{n} w_i} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i w_i}{\sum_{i=n}^{n} w_i}$$

42

Spa



Map showing changes to the **mean center of population** for the United States, 1790–2010 (US Census Bureau)

43

Spatial Statistics

- Measures of Central Tendency
  - Median Center/Euclidean Median
    - Center of minimum travel

44

Spatial Statistics

- Measures of Central Tendency
  - Manhattan Median
    - The point for which
      - half of the distribution is to the west the other half to the east (median of x coordinates)
      - And half to the north and the other half to the south (median of y coordinates)
    - The solution changes upon rotating the axes
    - For an even number of points, no exact solution

45

Spatial Statistics

- Measures of Dispersion
  - Standard Distance
    - Absolute spread of points around the mean center
    - Analogous to standard deviation

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} + \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}}$$

46

Spatial Statistics

- Measures of Dispersion
  - Relative Distance
    - Relative dispersion around a mean center
    - relates the standard distance to the size of the study area
    - RD = SD/r
      - RD - Relative Distance
      - SD - Standard Distance
      - r – radius of the cicle with the same area as the study area

47

Hypothesis Testing

- Multistep procedure that leads from a statement of hypothesis to a conclusive statement regarding the hypothesis
  - Conclusive statement is the decision
- The general goal is to make an ***inference*** about the magnitude of one or more population parameters based on sample statistics estimating those parameters

Parameter	Statistic
μ = population mean	x = sample mean
∂ = population standard deviation	s = sample standard deviation

48

Hypothesis Testing

- Steps
  1. State the null and alternate hypothesis
  2. Select appropriate statistical test
  3. Select level of significance
  4. Delineate regions of rejection and nonrejection of null hypothesis
  5. Calculate test statistic
  6. Make decision regarding null and alternate hypothesis

49

Two Complementary Hypotheses

H_0 = Null Hypothesis → There is **no significant** difference between two parameters

H_A = Alternate Hypothesis → There is **a significant** difference between two parameters

The aim of an inferential statistical test is to calculate probability that the null hypothesis is true. If this probability is acceptably low, then the null hypothesis can be rejected in favour of the alternative hypothesis. Thus, the sample results can be said to be significant.

50

Two Complementary Hypotheses

H_0: parameter$_1$ = parameter$_2$ (parameter$_2$ is the hypothesized parameter)

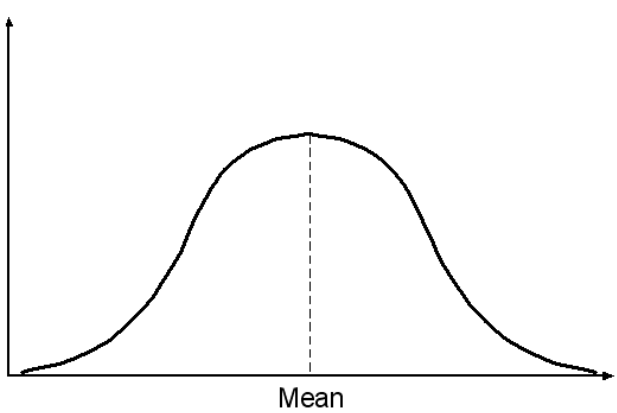
H_A: parameter$_1$ ≠ parameter$_2$ (two-tailed)

H_A: parameter$_1$ < parameter$_2$ (one-tailed)

H_A: parameter$_1$ > parameter$_2$ (one-tailed)

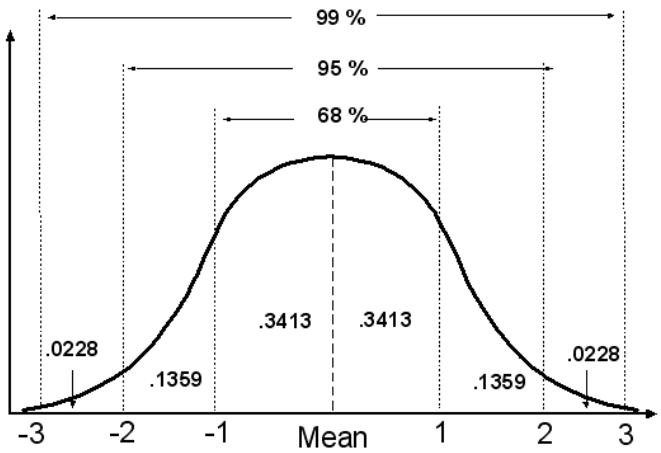
51

The frequency distribution can be used to test our hypothesis assuming that our data is normally distributed.



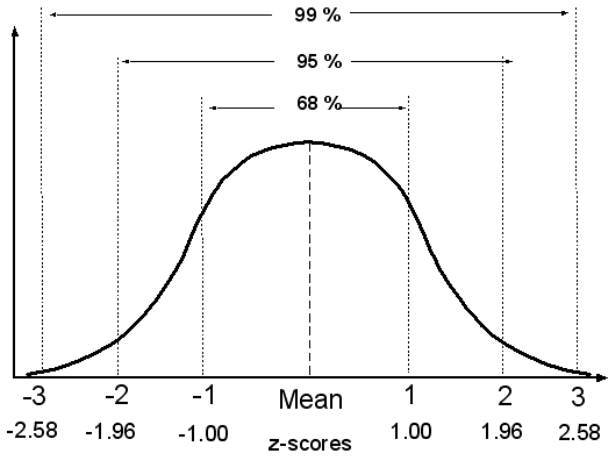
52

The frequency distribution can be divided into different sections, with each section containing a certain proportion of the data. Each section corresponds to a standard deviation of the data.



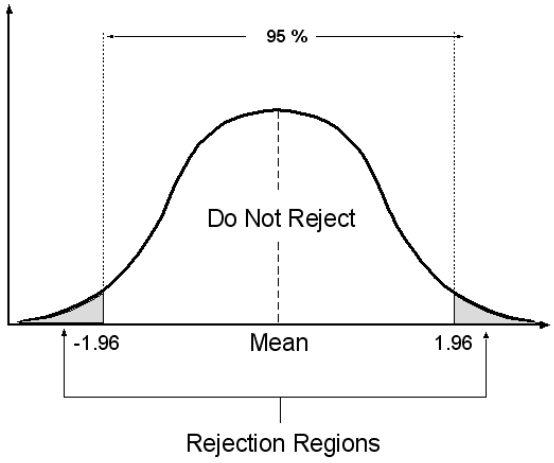
53

Each standard deviation corresponds to a particular z-score. The z-score values can be obtained in "The Normal Table" in most statistic text books.



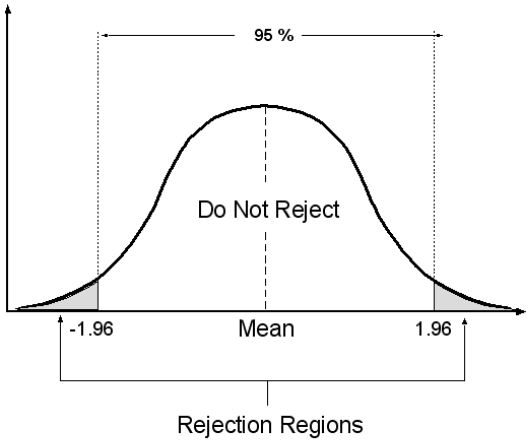
54

You can determine if a sample is significantly different than the mean by calculating the z-scores. If the z-score falls in the rejection region for a given level of confidence, the sample is significantly different from the mean.



55

Similarly, we can be 95% confident that a sample is different than random if the z-score is higher than 1.96 or lower than -1.96.



56

Spatial Data

- Some pitfalls of spatial analysis:
  - Spatial autocorrelation
    - Implies that you can't assume a phenomenon is distributed randomly
    - Understanding it's nature is of primary importance
  - The Modifiable Area Unit Problem
    - "a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomena resulting in the generation of artificial spatial patterns" (Heywood, 1988).

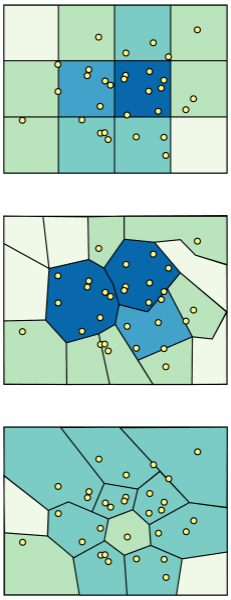
57

Spatial Data

- Some pitfalls of spatial analysis:
  - The Modifiable Area Unit Problem
    - Scale
    - Zones (see image)

"the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating" Openshaw, 1983

Thus, statistics should be interpreted and evaluated by acknowledging the particular boundary scheme used in the study or experiment (McGrew and Monroe 2000)



source: gispopsci.org

58

Let's assume a phenomenon such as average precipitation by zone

4.6	9.5	9.2	9.3
7.0	1.4	7.2	9.9
6.5	8.1	7.2	4.1
5.9	2.6	7.7	1.6

Mean: 6.3625
Standard Deviation: 2.7758

8.9	5.628
5.15	5.775

Mean: 6.3625
Standard Deviation: 1.7121

How do the statistics differ if the zones are configured differently?

59

Spatial Data

- Some pitfalls in spatial analysis:
  - Ecological Fallacy
    - Results Aggregated from data cannot be applied to individuals
  - Scale
    - Results depend on scale at which data was collected
  - Non-uniformity of space
    - Patterns can be random, clustered, uniform
  - Edge or Boundary Effects
    - No data beyond the study region
    - However, due to spatial autocorrelation, outside data could be affecting your study region

60

Further Reading

- Chor Pang Lo, Albert K. W. Yeung (2006) Concepts and Techniques of Geographic Information Systems (2nd Edition) . Chapter 3

- M.F. Goodchild, M. Yuan, T Cova (2007) Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science* 21(3) 239–260