GEOG 413/613

LECTURE 2

1

Housekeeping
Syllabus
Office Hours
Laptops

Data and Introductory Stats

2

Geospatial Data

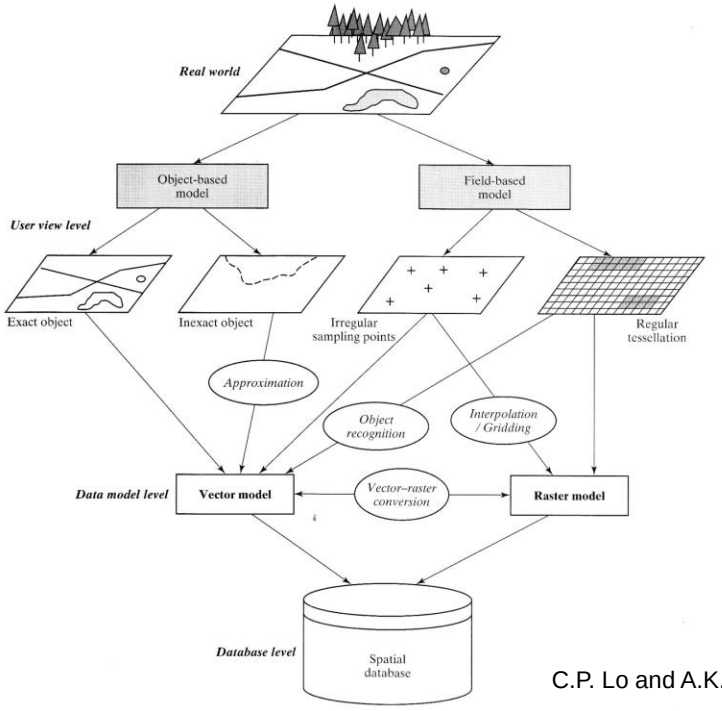
- Real world features exist in two basic forms
  - Objects
    - discrete, definite boundaries, persistent identity
    - E.G. Highways, buildings, parks, administrative regions
  - Phenomena
    - Distributed continuously over a large area
    - E.G. terrain elevation, temperature, rainfall, soil pH

3

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)

4



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

5

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

6

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
- There for the geographer asks questions about
  - Where: examples?
  - Why: examples?
  - What to do: examples?

7

7

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

8

8

Spatial Data

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

9

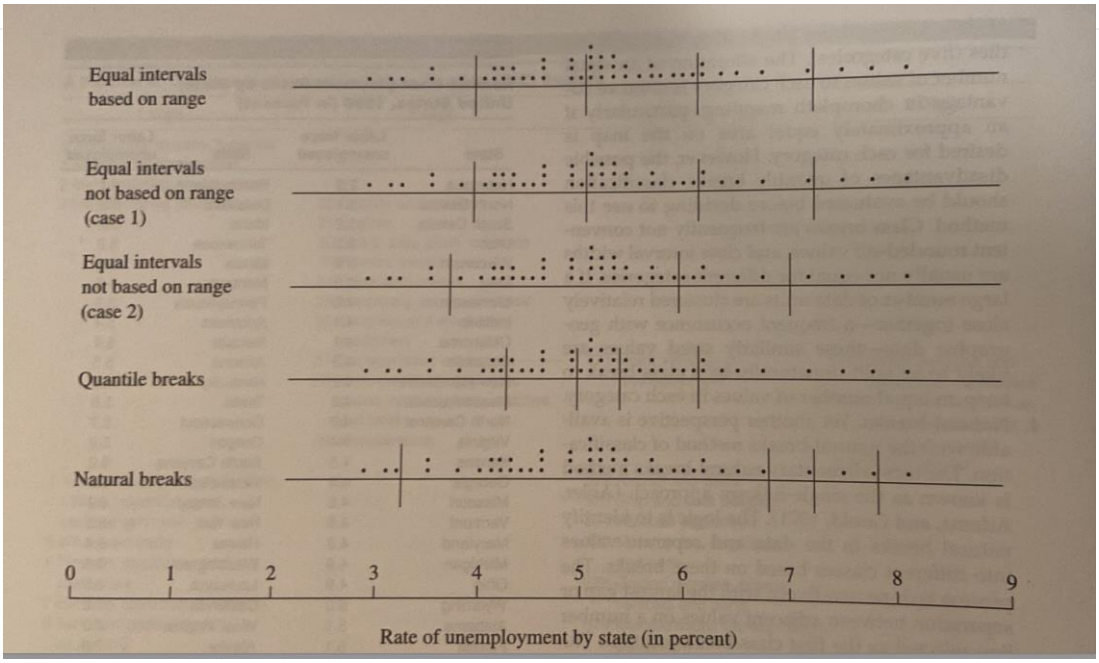
9

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

10

10



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

11

11

Spatial Data

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

12

12

Non-spatial Statistics

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

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

$\bar{X}$ – arithmetic mean
n – number of observations
a_i – value of observation i

13

13

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

14

14

Non-spatial Statistics

- Measures of Dispersion
  - Range
  - Standard Deviation

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

standard deviation for sample (≈ n<30)

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

standard deviation for population (≈ n>30)

15

15

Non-spatial Statistics

- Measures of Dispersion
  - Variance

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

variance for sample (≈ n<30)

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

Variance for population (≈ n>30)

16

16

Non-spatial Statistics

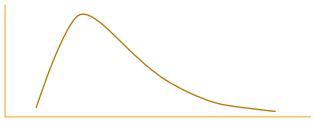
- Measures of Dispersion
  - Coefficient of Variation (CV)
    - The standard deviation and the variance are absolute measures, i.e. their values are dependent on the magnitude of the units of measurement.
    - The coefficient of variation is a relative measure that addresses this

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

17

17

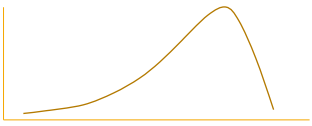
Non-spatial Statistics



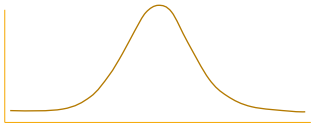
Positive Skewness, Skewness > 0

- 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

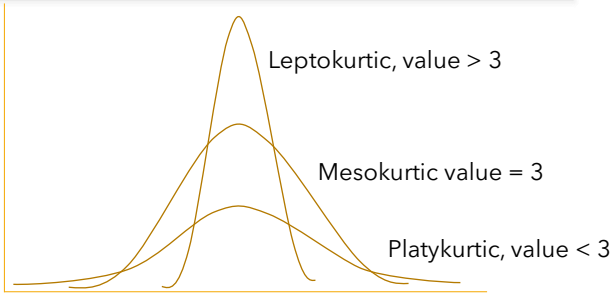
18

18

Non-spatial Statistics

- Measures of Relative Position
  - Kurtosis

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



19

19

Spatial Statistics

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

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

20

20

Spatial Statistics

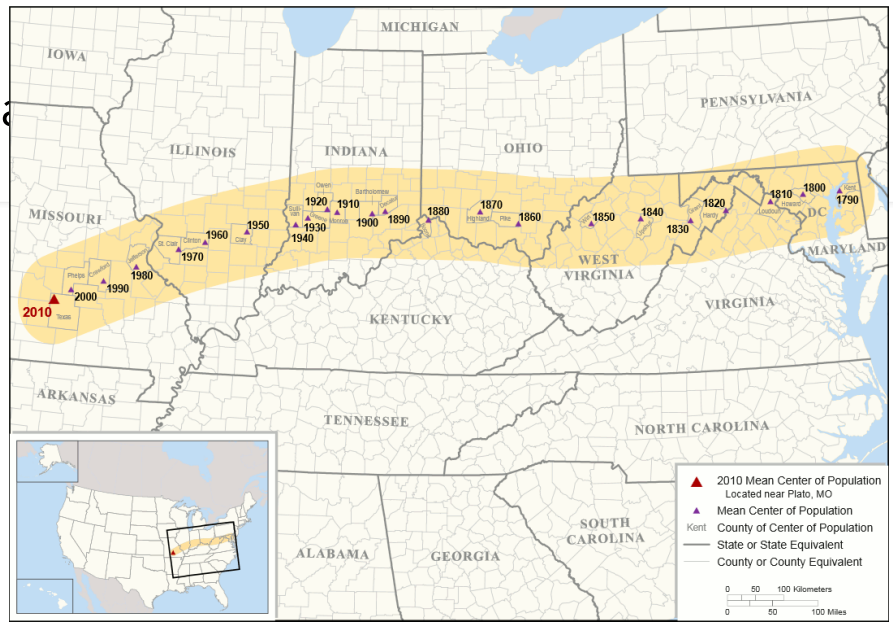
- Measures of Central Tendency
  - Weighed Mean Center
    - Mean affected by a weight factor (e.g. frequency, population)
    - Represents the centre of gravity $(\bar{x}, \bar{y})$

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

21

21

Spa



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

22

22

Spatial Statistics

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

23

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

24

Spatial Statistics

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

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

25

25

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

26

26

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	$\bar{x}$ = sample mean
∂ = population standard deviation	s = sample standard deviation

27

27

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

28

28

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.

29

29

Two Complementary Hypotheses

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

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

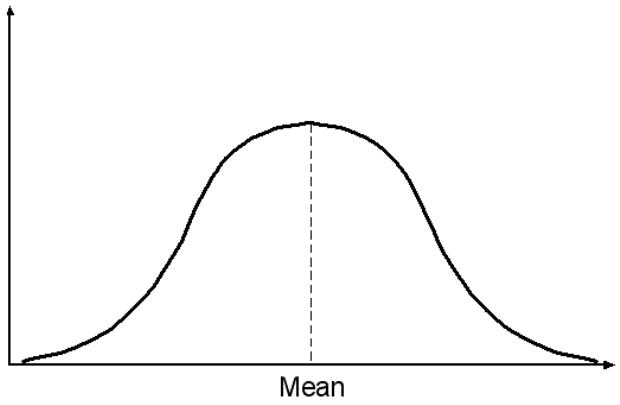
H_A: parameter$_1$ < parameter$_2$ (one-tailed)

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

30

30

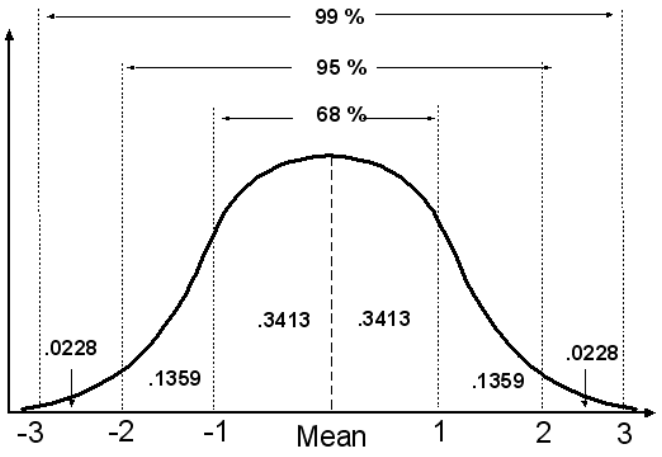
The frequency distribution can be used to test our hypothesis assuming that our data is normally distributed.



31

31

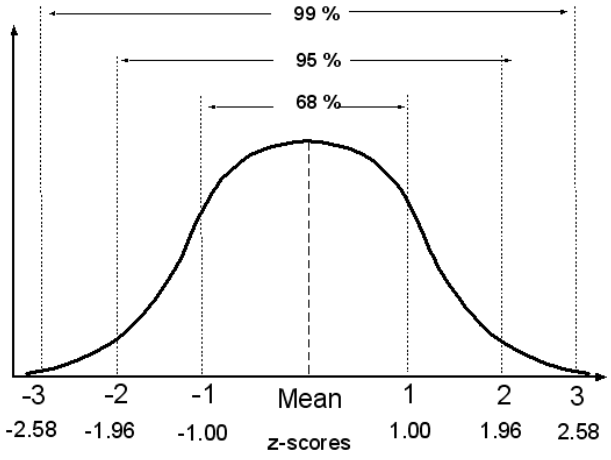
The frequency distribution can be divided into different sections, with each section containing a certain proportion of the data. Each section corresponds to a standard deviation of the data.



32

32

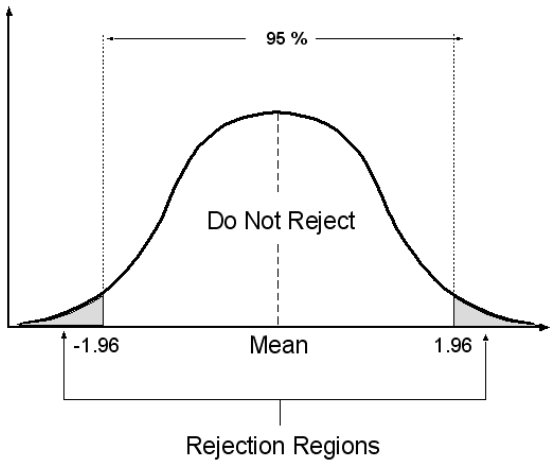
Each standard deviation corresponds to a particular z-score. The z-score values can be obtained in “The Normal Table” in most statistic text books.



33

33

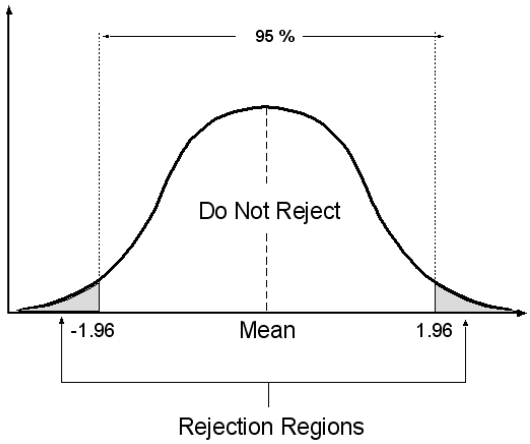
You can determine if a sample is significantly different than the mean by calculating the z-scores. If the z-score falls in the rejection region for a given level of confidence, the sample is significantly different from the mean.



34

34

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.



35

35

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).

36

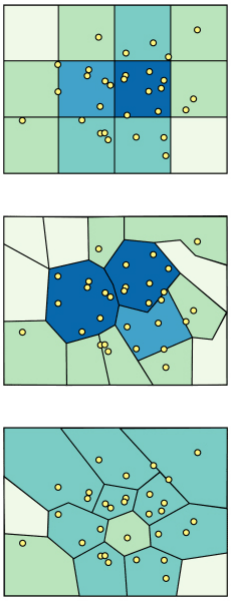
36

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

37

37

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?

38

38

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

39

39

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

40