Spatial Analysis and Modeling (GIST 4302/5302)

Guofeng Cao
Department of Geosciences
Texas Tech University

Outline of This Week

- Last week, we learned:
  - Data representation: Object vs. Field-based approaches
  - Common spatial operations
- This week, we will :
  - Review statistics and probability
  - Learning pitfalls of spatial data

Basic Definitions I

- **Population** vs. **sample**
  - 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., college students in Texas Tech

Basic Definitions II

- **Population parameters vs. sample statistics**
  - *Parameters:* summary measures that describe a population variable, e.g., average age of college students in the U.S.
  - *Statistics:* summary measures that describe a sample variable, e.g., average age of college students in *western* U.S.

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

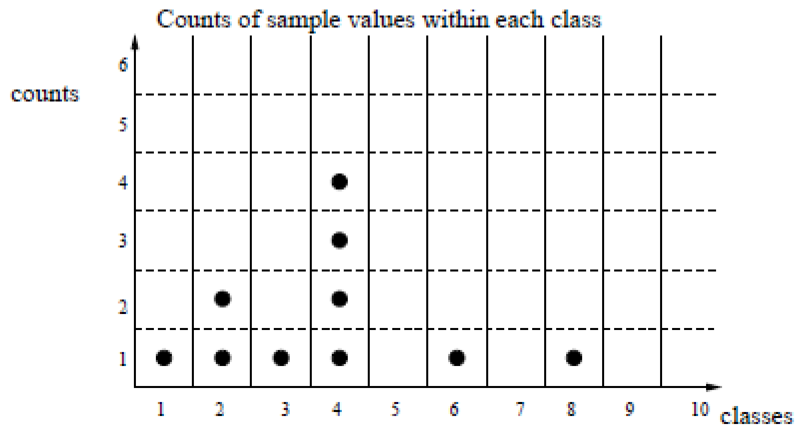
Statistical Procedures II

- **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 Example

Setting: Consider 10 hypothetical sample values:

4	1	3	8	4	4	2	4	6	2



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	0.1	0.2	0.1	0.4	0.0	0.1	0.0	0.1	0.0

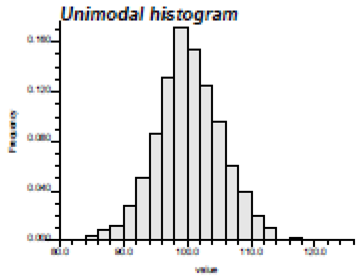
histogram shape depends on number and width of classes

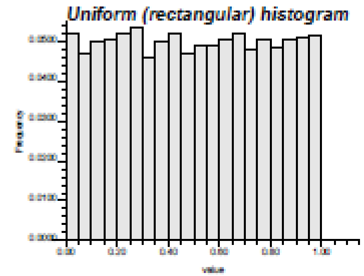
use non-overlapping equal intervals with simple bounds

rule of thumb for number of classes: $5 \times log_{10}$(*#of data*)

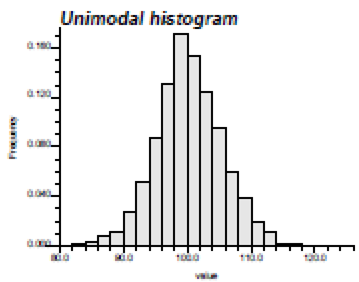
Histogram Shape Characteristics

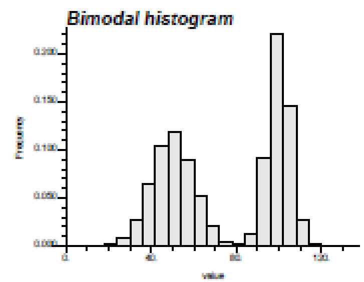
Peaked or not:



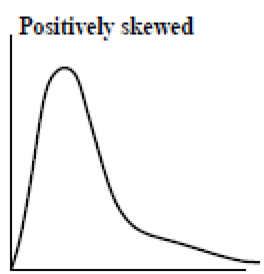


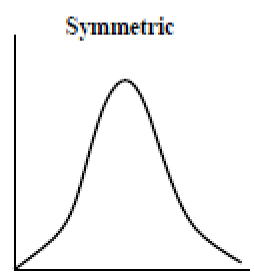
Number of peaks:

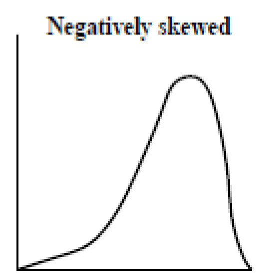




Symmetric or not:





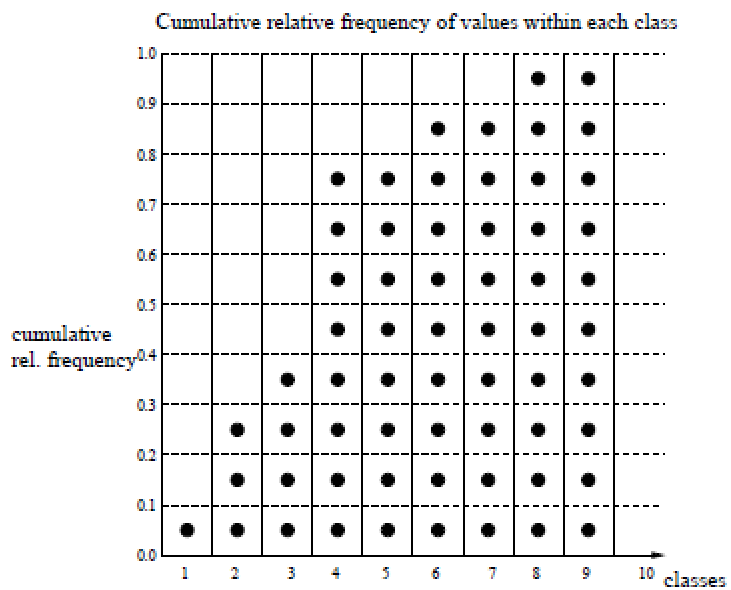


Cumulative Histogram Example

RANKED sample data and their relative frequency:

k	1	2	3	4	5	6	7	8	9
p_k	0.1	0.2	0.1	0.4	0.0	0.1	0.0	0.1	0.0

Cumulative relative frequency:



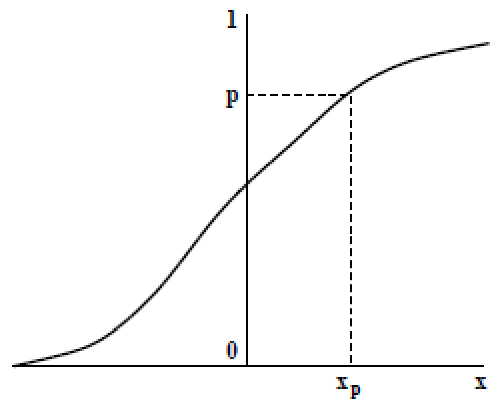
proportion of sample values less than, or equal to, any given cutoff

probability that any sample chosen at random be no greater than any given cutoff

Quantiles

Definition:

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



Useful quantiles:

min: $x_{0.0}$, lower quartile: $x_{0.25}$, median: $x_{0.5}$, upper quartile: $x_{0.75}$, max: $x_{1.0}$

e.g., upper quartile is the number (in data units) with 75% of data being less than or equal to this value

Percentiles: $x_{0.01}, x_{0.02}, \ldots, x_{0.98}, x_{0.99}$

Deciles: $x_{0.1}, x_{0.2}, \ldots, x_{0.8}, x_{0.9}$

Quantiles are not sensitive to extreme values (outliers)

Measures of Central Tendency

Mid-range:

- arithmetic average of highest and lowest data: $\frac{x_{max}+x_{min}}{2}$

Mode:

- most frequently occurring value in data set

Median:

- datum value that divides data set into two halves;
  also defined as 50-th percentile: $x_{0.5}$

Mean:

- arithmetic average of data set
- **sample mean**: $\bar{x}$ or $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **population mean**: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

Expressed in data units

Also, $m = \hat{\mu}$: the sample mean is an estimate of the population mean

Most appropriate measure of central tendency depends on distribution shape

Measures of Dispersion (1)

Range:

- difference between highest and lowest data: $x_{max} - x_{min}$

Interquartile range:

- difference between upper and lower quartiles: $x_{0.75} - x_{0.25}$

Mean absolute deviation from mean:

- average absolute difference between each datum and the mean:
  $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

Median absolute deviation from median:

- median absolute difference between each datum and the median:
  $median|x_i - x_{0.5}|$

Variance:

- average squared difference between any datum and the mean
- **sample variance**: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$
- **population variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

  $s^2 = \hat{\sigma}^2$: *the sample variance is an estimate of the population variance*

Measures of Dispersion (2)

Variance:

- alternative definition: difference between average squared 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**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$

Variance is expressed in data units SQUARED

Coefficient of variation:

- ratio of standard deviation and the mean
- **sample coefficient**: $\frac{s}{m}$ $s = \sqrt{s^2}$: sample std deviation
- **population coefficient**: $\frac{\sigma}{\mu}$ $\sigma = \sqrt{\sigma^2}$: population std deviation

The coefficient of variation, and std deviation are UNIT-LESS

Choosing alternative measures of dispersion:

- any summary statistic involving squared values is sensitive to outliers
- any summary statistic based on quantiles is robust to outliers
- coefficient of variation: very useful for comparing spread of different data sets

Normalizing Data

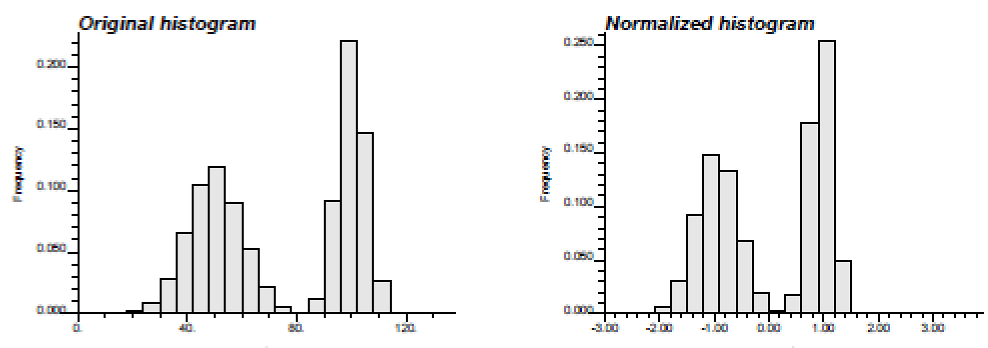
Normalizing data to zero mean and unit variance
allows more meaningful comparison of different data sets

Normalization procedure:

1. compute mean m and standard deviation s of data set
2. subtract the mean from each datum: $x_i - m$
3. divide by the standard deviation: $z_i = \dfrac{x_i - m}{s}$

Normalized data are unit free;
shape of distribution does not change (e.g., modes remain the same)

Example:



Normalized datum z_i is nothing else than the distance of the original datum x_i to the mean in terms of standard deviation units

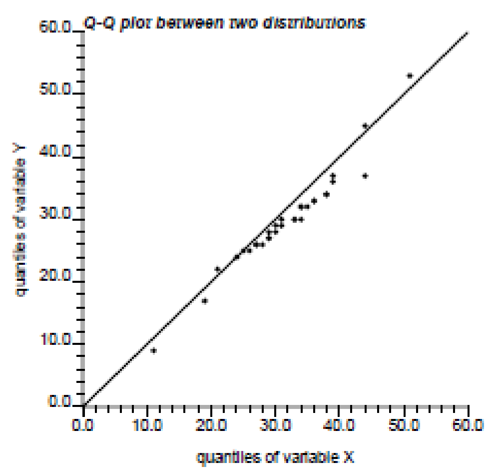
Quantile-Quantile (Q-Q) Plots

Graph for comparing the shapes of two distributions

Procedure:

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

Example:



Interpretation:

- straight plot alinged with 45° line implies two similar distribution shapes

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={E1, E2,..., Ek} 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.

Relationships Between Events

Complementary event:

- set $\overline{A}$ (not A) of elementary outcomes not in an event space A
- e.g., a dry-day event is the complementary of a wet-day event

Intersection of events:

- set $A \cap B$ (A and B) of elementary outcomes that belong to both events A and B of a sample space S
- e.g., a wet day with both liquid and frozen precipitation is the intersection of two events: (i) a wet day with liquid precipitation, and (ii) a wet day with frozen precipitation

Mutually exclusive events:

- events A and B defined on same sample space S and have no elementary outcomes in common; in this case: $A \cap B = \emptyset$ (null event)
- e.g., a wet day and a dry day are two mutually exclusive events

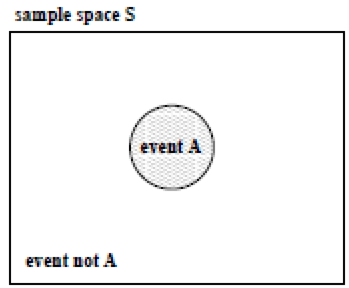
Union of events:

- set $A \cup B$ (A or B) of all elementary outcomes that belong to *at least one* of two events A and B, both defined over same sample space S
- e.g., union of liquid and frozen precipitation = wet-day event

Depicting Events via Venn Diagrams

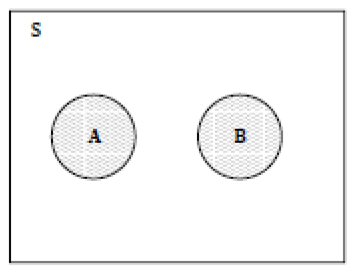
Venn Diagrams:

- pictorial representation of sample spaces (S) and events (A)

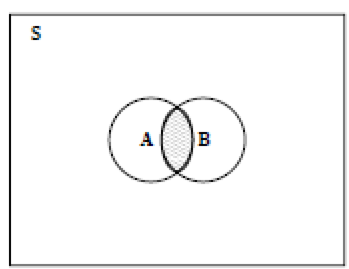


Examples:

- union $A \cup B$ of two events A and B



- intersection $A \cap B$ of two events A and B



Probability (1)

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:

$$P(A) = Prob\{A\} = \frac{n}{N}, \quad \text{as } N \text{ tends to infinity}$$

- e.g., the probability for a wet-day event over a region can be seen as the proportion of wet-days in a very large precipitation record

Axioms of probability:

- probabilities are necessarily non-negative: $P(A) \geq 0$

  e.g., the probability for a wet-day event is always zero or positive

- the sample space S will certainly occur: $P(S) = 1$

  the probability of all outcomes of a random experiment add up to 1;
  e.g., the probability of a wet-day event and that of a dry-day event is one:
  it will either rain or not

- for two mutually exclusive events A and B: $P(A \cup B) = P(A) + P(B)$

  probability that either A or B occur is equal to the probability of A to occur plus that of B to occur

Probability (2)

Elementary probability theorems:

- the impossible event $\emptyset$ has zero probability of occurrence: $P(\emptyset) = 0$

- the probability of the complement $\bar{A}$ of an event A to occur is:
  $P(\bar{A}) = 1 - P(A)$

  e.g., if the probability for a wet-day event is 0.4,
  then the probability for a dry-day event is: 1 - 0.4 = 0.6

- the probability of any event A to occur cannot be greater than one:
  $P(A) \leq 1$

- the probability of either two events A and B (not necessarily mutually exclusive) to occur is: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  $P(A \cap B)$ *or* $P(A,B)$ *= joint probability of* A *and* B *occurring simultaneously;*
  e.g., probability for either liquid or frozen precipitation =
  probability of liquid precipitation
  + that of frozen precipitation
  - that of both liquid and frozen precipitation

Probability Calculation

Example: $n = 10$ outcomes of a binary event A, e.g., 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	1	1	1	0	1	0	0

Mean of zeros and ones: $\frac{1}{n}\sum_{i=1}^{n} a_i = \frac{6}{10} = 0.6$

average of binary events a_i = probability for event A to occur = $Prob\{A = 1\}$

proportion of wet days in record = probability for wet-day event

Example: $n = 10$ outcomes x_i of a variable X, e.g., precipitation (in mm/day), and associated binary event a_i indicating values ≤ 4 ($a_i = 1$ if $x_i \leq 4$, 0 if not):

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

Mean of zeros and ones: $\frac{1}{n}\sum_{i=1}^{n} a_i = \frac{4}{10} = 0.4$

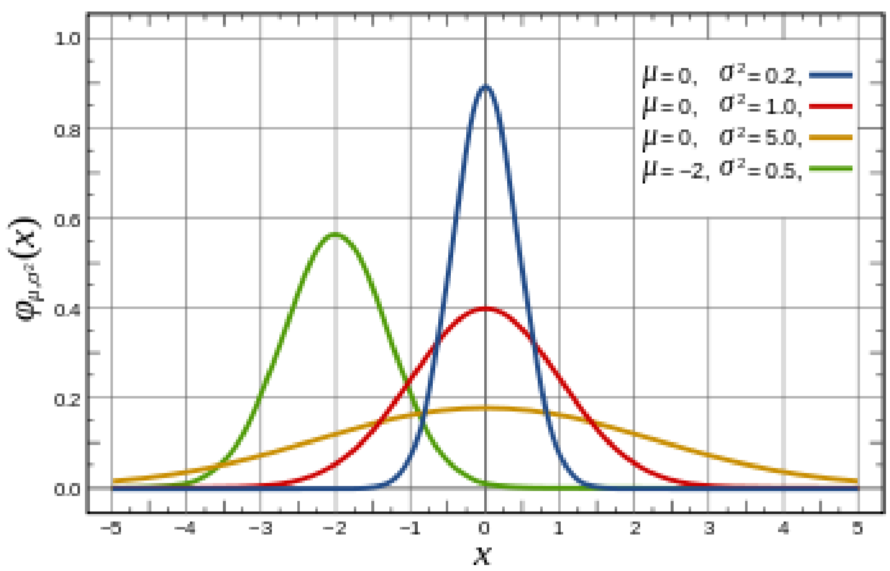
average of binary events a_i = probability for event A to occur =
$Prob\{A = 1\} = Prob\{X \leq 4\}$

proportion of days with precip no greater than $4mm/day$ in record

Frequently Used Probability Distribution

- Gaussian (Normal) distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Conditional Probability (1)

Definition:

- probability of event A to occur *given* that event B has occurred:

$$\boxed{P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A \cap B)}{P(B)}}$$

Interpretation:

- conditional probability $P(A|B)$ = ratio of probability that both events occur simultaneously $P(A,B)$, to probability $P(B)$ of conditioning event:

$$\text{cond. probability} = \frac{\text{joint probability}}{\text{probability of conditioning event}}$$

Example:

- in a weather forecasting context:

what is the probability of precipitation today,
given that temperature is lower than some value?

Conditional Probability Calculation

Example: $n = 10$ outcomes of two variables X and Y, e.g., precipitation X and temperature Y:

day i	1	2	3	4	5	6	7	8	9	10
x_i	0	3	5	0	0	4	8	5	0	0
y_i	15	40	56	25	15	45	60	50	30	10

Binary events: ($a_i = 1$, 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
a_i	0	1	1	0	0	1	1	1	0	0
b_i	0	1	1	1	0	1	1	1	1	0

Joint probability:

$$P(A,B) = \frac{1}{n}\sum_{i=1}^{n} a_i \cdot b_i = \frac{5}{10} = 0.5$$

average of product of indicators $a_i \cdot b_i$ = proportion of **joint** *events*

Conditional probability :

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{0.5}{0.7} = 0.71$$

Conditional Probability and Independence

Independent events:

- two events A and B are independent iff: $P(A|B) = P(A)$
- knowledge of conditioning event B does not alter the probability of event A to occur
- in our previous example: $P(A|B) = 0.71 \neq 0.5 = P(A)$

Alternatively:

- two events A and B are independent iff: $P(A,B) = P(A) \cdot P(B)$
- joint probability $P(A,B)$ of two events = product of individual occurrence probabilities $P(A)$ and $P(B)$
- in our previous example:
  $P(A,B) = 0.5 \neq (0.5 \cdot 0.7) = 0.35 = P(A) \cdot P(B)$

Covariance and Correlation Coefficient

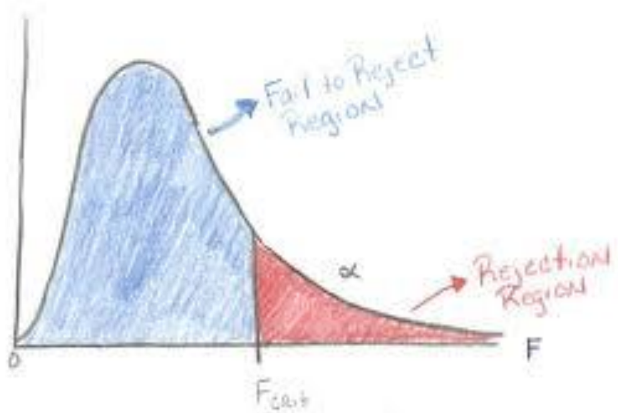
- Suppose that *X* and *Y* are random variables for a random experiment.
- The *covariance* of *X* and *Y* is defined by
  - $\mathrm{cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$
- The *correlation* of *X* and *Y* is defined by (normalized covariance)

$$\mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

- Cov(X,Y) = 0 ->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

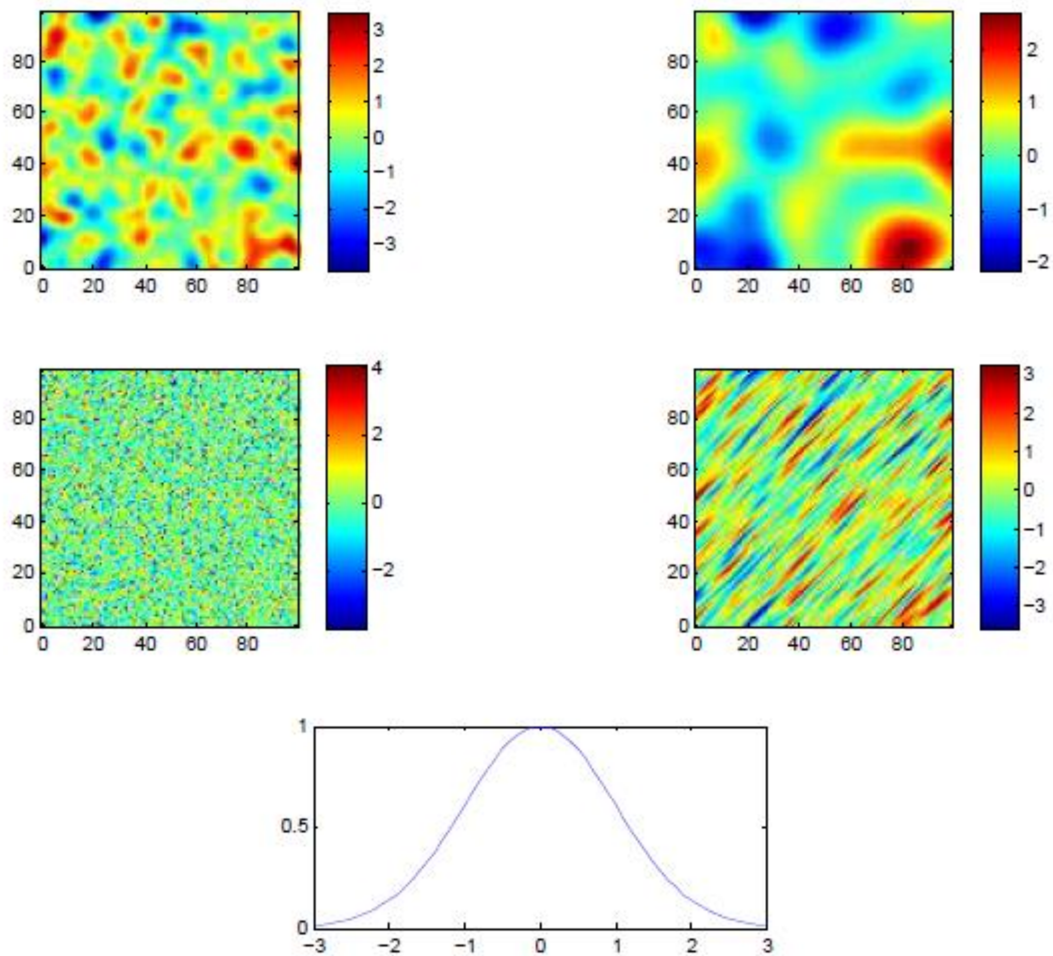


Pitfalls of Spatial Data

- Spatial effects
  - Spatial correlation: redundancy in sample data = classical statistical hypothesis testing procedures not applicable
  - Spatial heterogeneity
- The modified areal unit problem (MAUP)
  - spatial averages display different spatial characteristics and relationships than original (non-averaged) values
  - aggregation and zoning effects
- Ecological Fallacy
  - relationships established at a specific level of aggregation (e.g., census tracts) do not hold at more detailed levels (e.g., individuals)
  - Occasionally, it holds, e.g. tobacco vs lung cancer
- Scale effects
- Non-uniformity of space and edge effects

Spatial Effects

- Spatial patterns can make HUGE differences



The Modified Areal Unit Problem (MAUP)

- 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, and even Presidential election results, Gore vs Bush
  - Two aspects of this problem: scale effect and zoning (aggregation) effect

MAUP: Scale Effect (I)

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

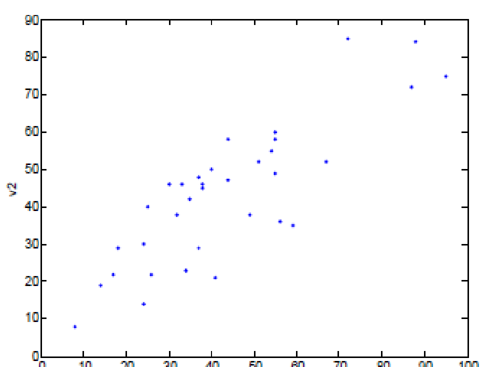
Example

spatial variable #1 versus spatial variable #2

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

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

$\rho(v1, v2) = 0.83$



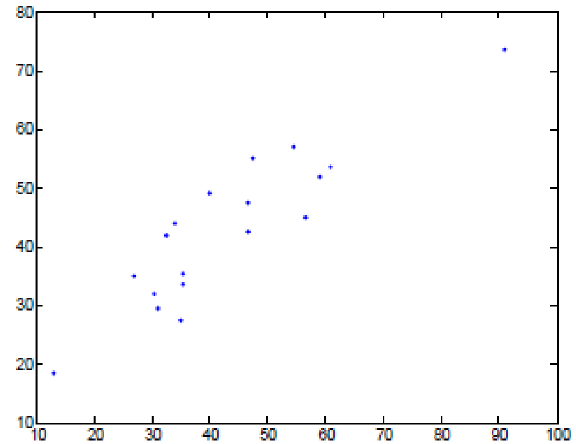
MAUP: Scale Effect (II)

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

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

$\rho(v1, v2) = 0.90$



MAUP: Zone Effect (1)

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

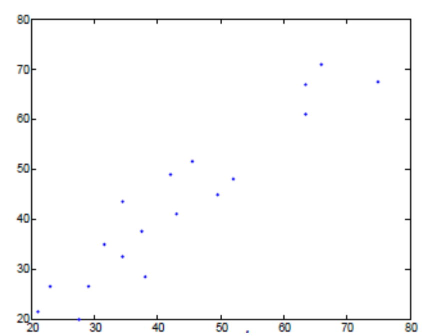
Example

spatial aggregation strategy #2

63.5	75	63.5	37.5	66	29.0
27.5	43	31.5	34.5	23	21
52.0	34.5	42	49.5	38.0	45.5

61.0	67.5	67.0	37.5	71.0	26.5
20.0	41.0	35.0	32.5	26.5	21.5
48.0	43.5	49.0	45.0	28.5	51.5

$\rho(v1, v2) = 0.94$



Ecology Fallacy (I)

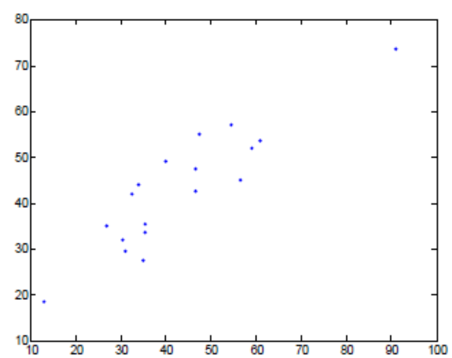
- relationships established at a specific level of aggregation do not hold at more detailed levels

Example

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

$\rho(v1, v2) = 0.90$



Ecology Fallacy (II)

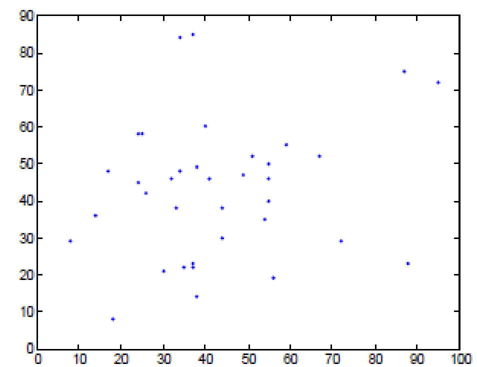
- relationships established at a specific level of aggregation do not hold at more detailed levels *Example*

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

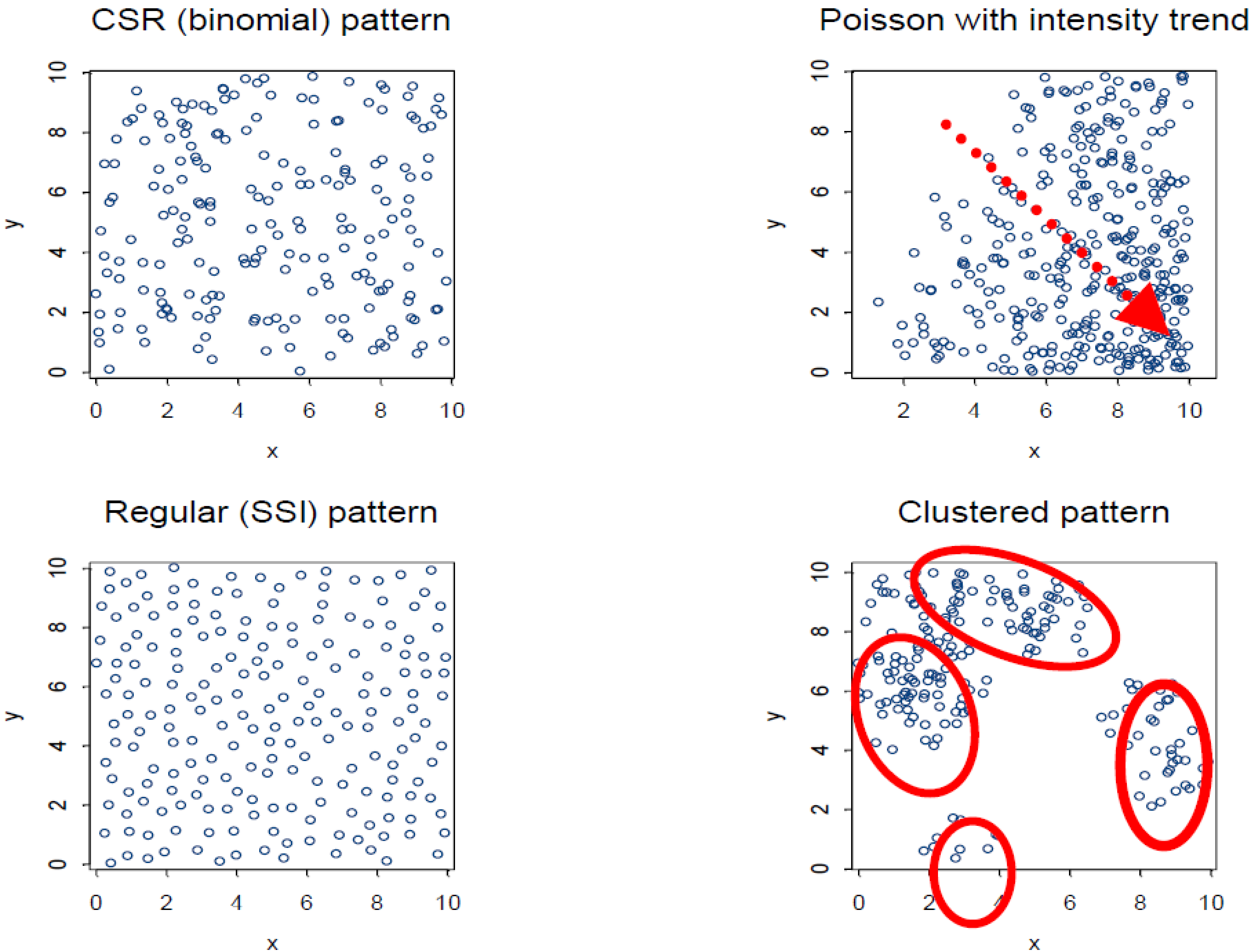
72	75	85	29	58	30
50	60	49	46	84	23
21	46	22	42	45	14
19	36	48	23	8	29
38	47	52	52	22	48
58	40	46	38	35	55

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



Next Topic

- Point pattern analysis
  - Point pattern descriptors
  - Point pattern analysis:
    - Density and distance measures (or first order vs. second order)
    - Hypothesis testing of clustering pattern


CSR (binomial) pattern
Poisson with intensity trend
Regular (SSI) pattern
Clustered pattern
x
y
0
2
4
6
8
10

- To be continued