

# *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](mailto:guofeng.cao@ttu.edu)

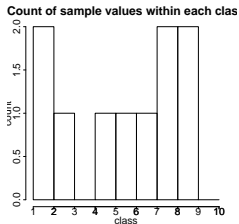
Spring 2018



# Histogram

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

2	2	9	8	7	9	5	6	8	3
---	---	---	---	---	---	---	---	---	---



- Relative frequency table:

$$p_k = \# \text{ of data in } k\text{-th class} / (\text{total } \# \text{ of data})$$

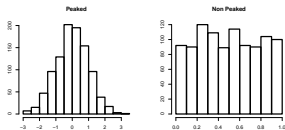
k	1	2	3	4	5	6	7	8	9
$p_k$	0.2	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.0

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

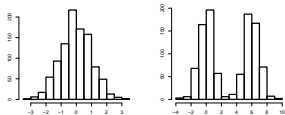


# Histogram Shape Characteristics

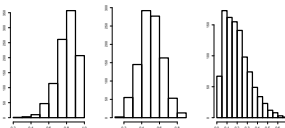
- Peaked or not



- Numbers of peaks



- Symmetric or not



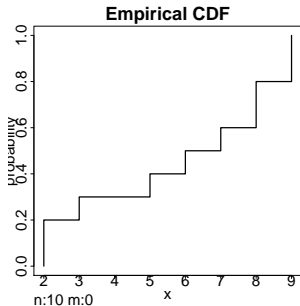


# Cumulative Histogram

- 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



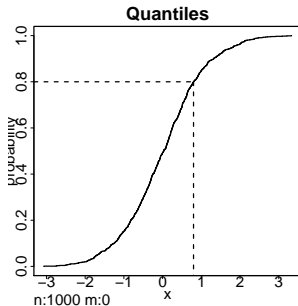
- Proportion of sample values less than, or equal to, any given cutoff value
- Probability that any random sample is no greater than and given cutoff value



# 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:  $x_{0.01}, x_{0.02}, \dots, x_{0.99}$
  - Deciles:  $x_{0.10}, x_{0.20}, \dots, x_{0.90}$
- 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 occurring values in data sets
- median: datum value that divides data set into halves; also defined as 50-th percentiles:  $x_{0.5}$
- mean: arithmetic 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 estimation 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: average 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 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:  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$
  - Note: variance is expressed in squared data units
- standard deviation: square root of variance  $s$  or  $\sigma$ 
  - 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 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

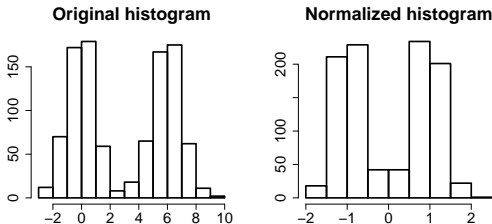




# Normalizing Data

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

- Normalizing procedure:
  1. compute mean  $m$  and standard deviation  $s$  of data set A
  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 distribution does not change (e.g., modes remain the same)

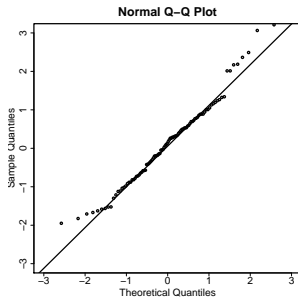




# Quantile-Quantile (Q-Q) Plots

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



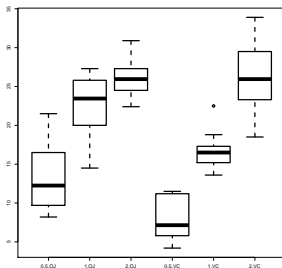
- Interpretation: straight plot aligned with  $45^\circ$  line implies two similar distribution



## 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 often 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

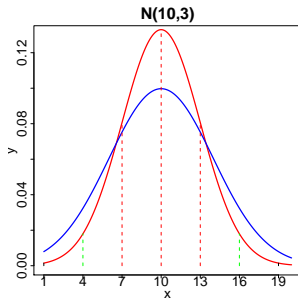


- Points outside of range are usually taken as outliers



# Commonly Used Probability Distributions

- Gaussian (or normal) distribution



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



# Covariance and Correlation Coefficient

---

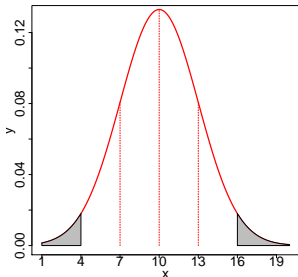
Suppose  $X$  and  $Y$  are two random variables for a random experiment

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

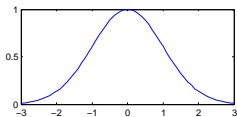
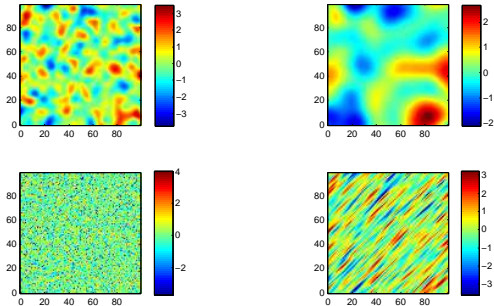
- problem close related to the MAUP
- relationships established at a specific level of aggregation (e.g., census tracts) do not hold at more detailed levels (e.g., individuals)





# 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

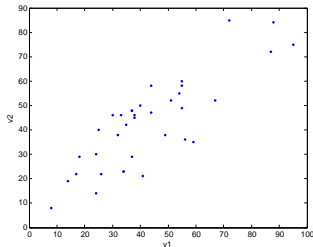
Analytical results depending on the size of units used (generally, bigger units lead to stronger correlation)

## 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

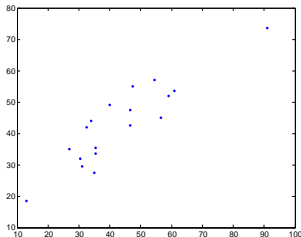
Analytical results depending on the size of units used (generally, bigger units lead to stronger correlation)

## 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

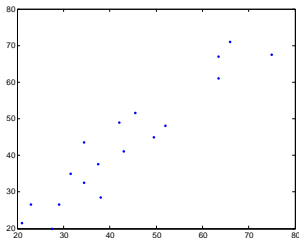
Analytical results depending on how the study area is divided up, even at the same scale

## 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

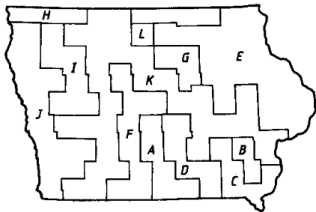


Figure 2a. Zoning system that minimises the regression slope coefficient  
(-24,  $r = -.25$ )

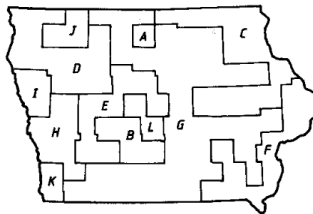


Figure 2b. Zoning system that maximises the regression slope coefficient  
(12,  $r = .87$ )

*Figure:* Image Courtesy of OpenShaw



# Ecological Fallacy (I)

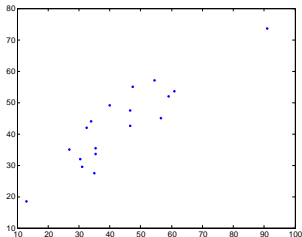
- relationships established at a specific level of aggregation do not hold at more detailed levels

## Example

Table: spatial aggregation strategy # 1

91.0	47.5	35.5	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:  $\rho(v1, v2) = 0.90$







## Ecological Fallacy (II)

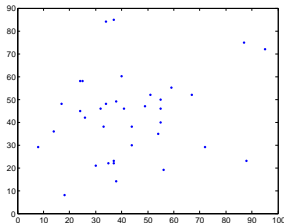
- relationships established at a specific level of aggregation do not hold at more detailed levels

### Example

*Table:* spatial variable #1 versus spatial variable #2

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

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





## 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