• **An Example:** Consider a list of 10 hypothetical sample values:

```
2 2 9 8 7 9 5 6 8 3
```

• **Relative frequency table:**

\[
p_k = \frac{\# \text{ of data in } k\text{-th class}}{\text{total } \# \text{ of data}}
\]

```
\begin{array}{cccccccccc}
  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 \\
\end{array}
```

• **Please note:** Histogram shape depends on number and width of classes; rule of thumb for number of classes: \(5 \times \log_{10}(\# \text{ of data})\) 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

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_k$</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- 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:
  \[ \frac{x_{\text{max}} + x_{\text{min}}}{2} \]
- **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: \( x_{\text{max}} - x_{\text{min}} \)
- **interquartile 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**: \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 \)
  - **standard deviation**: square root of variance \( s \) or \( \sigma \)
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 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)
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'
Assuming the null hypothesis is true, the p-value is the probability a test statistic 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 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: Examples

Example 1

<table>
<thead>
<tr>
<th>200</th>
<th>100</th>
<th>400</th>
<th>10</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>300</td>
<td>300</td>
<td>50</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>500</td>
<td>400</td>
<td>100</td>
<td>60</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Base Population  University degree (count)

<table>
<thead>
<tr>
<th>5%</th>
<th>20%</th>
<th>3%</th>
<th>6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>3%</td>
<td>2%</td>
<td>8%</td>
</tr>
<tr>
<td>12%</td>
<td>3%</td>
<td>20%</td>
<td>9%</td>
</tr>
</tbody>
</table>

a - scale effect

<table>
<thead>
<tr>
<th>10%</th>
<th>7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>8%</td>
<td>16%</td>
</tr>
</tbody>
</table>

b - zoning effect

HOW TO STEAL AN ELECTION

50 PRECINCTS
60% BLUE
40% RED

5 DISTRICTS
5 BLUE
0 RED
BLUE WINS

5 DISTRICTS
3 RED
2 BLUE
RED WINS
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

<table>
<thead>
<tr>
<th></th>
<th>87</th>
<th>95</th>
<th>72</th>
<th>37</th>
<th>44</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>41</td>
<td>56</td>
<td>49</td>
<td>44</td>
<td>25</td>
<td>55</td>
</tr>
<tr>
<td>40</td>
<td>55</td>
<td>55</td>
<td>38</td>
<td>34</td>
<td>44</td>
<td>24</td>
</tr>
<tr>
<td>50</td>
<td>38</td>
<td>26</td>
<td>35</td>
<td>38</td>
<td>24</td>
<td>35</td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>56</td>
<td>57</td>
<td>67</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>49</td>
<td>44</td>
<td>51</td>
<td>67</td>
<td>17</td>
<td>37</td>
<td>59</td>
</tr>
<tr>
<td>55</td>
<td>25</td>
<td>33</td>
<td>59</td>
<td>54</td>
<td>68</td>
<td>72</td>
</tr>
</tbody>
</table>

Table: \( \rho(v_1, v_2) = 0.83 \)
Scale effect

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

Example

<table>
<thead>
<tr>
<th>spatial aggregation strategy # 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.0</td>
</tr>
<tr>
<td>35.0</td>
</tr>
<tr>
<td>54.5</td>
</tr>
<tr>
<td>35.5</td>
</tr>
<tr>
<td>34.0</td>
</tr>
<tr>
<td>13.0</td>
</tr>
</tbody>
</table>

| 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: $\rho(v_1, v_2) = 0.90$
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

<table>
<thead>
<tr>
<th></th>
<th>63.5</th>
<th>75</th>
<th>63.5</th>
<th>37.5</th>
<th>66</th>
<th>29.0</th>
<th>61.0</th>
<th>67.5</th>
<th>67.0</th>
<th>37.5</th>
<th>71.0</th>
<th>26.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.5</td>
<td>43</td>
<td>31.5</td>
<td>34.5</td>
<td>23</td>
<td>21</td>
<td></td>
<td>20.0</td>
<td>41.0</td>
<td>35.0</td>
<td>32.5</td>
<td>26.5</td>
<td>21.5</td>
</tr>
<tr>
<td>52.0</td>
<td>34.5</td>
<td>42</td>
<td>49.5</td>
<td>38.0</td>
<td>45.5</td>
<td></td>
<td>48.0</td>
<td>43.5</td>
<td>49.0</td>
<td>45.0</td>
<td>28.5</td>
<td>51.5</td>
</tr>
</tbody>
</table>

Table: $\rho(v_1, v_2) = 0.94$
Zoning effect: another example

**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

<table>
<thead>
<tr>
<th>91.0</th>
<th>47.5</th>
<th>35.5</th>
<th>73.5</th>
<th>55.0</th>
<th>33.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.0</td>
<td>46.5</td>
<td>40.0</td>
<td>27.5</td>
<td>42.5</td>
<td>49.0</td>
</tr>
<tr>
<td>54.5</td>
<td>46.5</td>
<td>30.5</td>
<td>57.0</td>
<td>47.5</td>
<td>32.0</td>
</tr>
<tr>
<td>35.5</td>
<td>59.0</td>
<td>32.5</td>
<td>35.5</td>
<td>52.0</td>
<td>42.0</td>
</tr>
<tr>
<td>34.0</td>
<td>61.0</td>
<td>31.0</td>
<td>44.0</td>
<td>53.5</td>
<td>29.5</td>
</tr>
<tr>
<td>13.0</td>
<td>27.0</td>
<td>56.5</td>
<td>18.5</td>
<td>35.0</td>
<td>45.0</td>
</tr>
</tbody>
</table>

**Table**: $\rho(v_1, v_2) = 0.90$
Ecological Fallacy (II)

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

Example

\[ \rho(v_1, v_2) = 0.21 \]
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, programmable
Acknowledgement

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