GEOG 204

LECTURE 8
Quantitative Methods

1

House Keeping Stuff

Student Projects

- 5 or 6 topics
  - Choose 2
- If you have a specific idea that relates to your involvement in the community or studies, please talk to your TA. These are welcome.

Final Exam

- Cumulative
- Same structure as the midterm

2

2

Quantitative Methods in Geography

- Statistics
  - Collection, classification, presentation and analysis of numerical data
  - Drawing valid conclusions and making reasonable decision the basis of analysis
- For geographers
  - Describe and summarize spatial data
  - Generalizations about complex spatial patters
  - Assess whether pattern matches expected patterns
  - Making inferences about a population from collected data sample

3

3

Levels or Scales of Measurement

- Nominal
  - Categorical data e.g. land use type, religious affiliation
- Ordinal
  - Ranked data , e.g. main, secondary, minor roads
- Interval:
  - Interval between any two units can be measured on scale. Zero value is assigned arbitrarily e.g. Celsius and Fahrenheit scales (80°F is not twice as hot as 40°F)
- Ratio:
  - interval data with an absolute zero value

4

4

Summary of Levels of Measurement

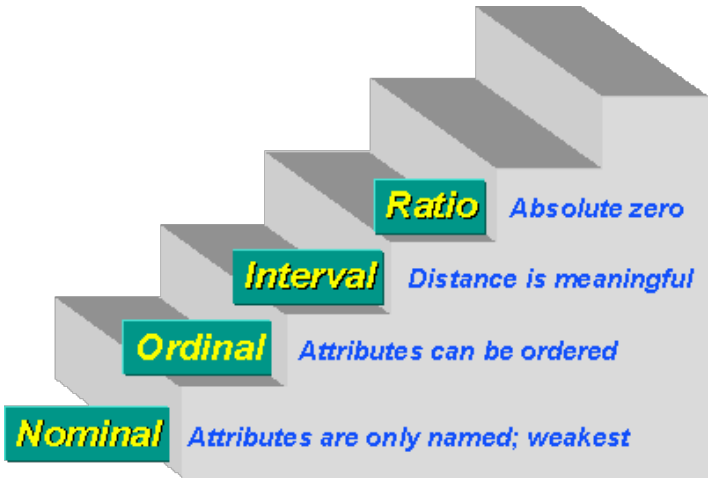
Level of measurement	Brief description
Nominal	Each value or unit of data is assigned to one of at least two categories or qualitative classes; no assumptions are made about relationships between categories—only that they are "different."
Ordinal	Values themselves are placed in some rank order.
Strongly ordered	Each value or unit of data is given a particular position in a rank-order sequence; that is, each value is assigned its own particular rank.
Weakly ordered	Each value or unit of data is assigned to a category, and the categories are then rank ordered.
Interval	Each value or unit of data is placed on a measurement scale, and the interval between any two units of data on this scale can be measured; origin or zero starting point is assigned arbitrarily (i.e., origin does not have a "natural" or "real" meaning).
Ratio	Each value or unit of data is placed on a measurement scale, and the interval between any two units of data on this scale can be measured; origin or zero starting point is "natural" or non-arbitrary, making it possible to determine the ratio between values.

Source: J. Chapman, Jr. McGrew. An Introduction to Statistical Problem Solving in Geography

OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

5

5



- **GIS Attribute field types:**
  - Interval/ Ratio:
    - **Integer (long/short)**
    - **Short:** - 32,000 to +32,000
    - **Long:** - billions to + billions
  - Real (float/double):
    - **Float:** 'precision'- field length; scale - number of decimals
    - **Double:** 15 digits
  - Nominal:
    - **Text, (includes code number)**
  - Ordinal:
    - **Date, Text, (short integer)**

6

Population and Sample

- In collecting data, it is often impossible to or impractical to observe the entire group
  - Population: the entire group
  - Sample: a small part of the group
- A population can be finite or infinite
  - Finite: the bounds of the population are known. e.g. birds in an enclosure
  - Infinite: the bounds of the population are unknown e.g. number of birds in Northern BC

7

7

Descriptive and Inferential Statistics

- If a sample is representative of the population, important conclusions can be made about a population
  - Inferential statistics are used to make generalization about the population based on the information based on the sample
  - Because the inference cannot be absolutely certain, the language of probability used to state the conclusions
- Descriptive statistics are used to describe and analyze a dataset (group) without making generalization about the population it may represent
  - Usually, a single measure or statistic

8

8

Discrete and Continuous Variables

- A variable is a property or a characteristic of each a given phenomenon or object that can be measured
  - The resulting measurement or code is called a ***data value***
  - a variable which can theoretically fall between two values is called a **continuous variable.**
    - Has infinite number of possible values
    - E.g. height can be 178,178.78,179 centimeters
  - a variable which can be determined by counting is a called **discrete variable**
    - Number of children in a household can be 0, 1, 2, 3… (not 2.5 or 2.2)

9

9

Descriptive Statistics

- Descriptive statistics provide concise, easily understood characteristics of a particular dataset
  - Measures of Central Tendency
  - Measures of Dispersion and Variability
  - Measures of Shape and Relative Position

10

10

Measures of Central Tendency

- Measures of central tendency represent the center or a typical value of a frequency distribution
  - Mode
  - Median
  - Mean

11

11

Measures of Central Tendency

- Mean
  - The arithmetic mean or average

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

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

12

12

Measures of Central Tendency

- Weighted Mean
  - Mean calculated for grouped data
    - Grouped data: Class intervals and Frequencies

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

X_w – *arithmetic mean*
f_i – *frequncy of class interval*
n – *number of observations*
x_i – *mid point of class interval i*

13

13

Class interval j	Class midpoint X_j	Class frequency f_j	$X_j f_j$
25–29.99	27.5	4	110.0
30–34.99	32.5	5	162.5
35–39.99	37.5	12	450.0
40–44.99	42.5	9	382.5
45–49.99	47.5	5	237.5
50–54.99	52.5	4	210.0
55–59.99	57.5	1	57.5
Total		40	1610.0

$$\bar{X}_w = \frac{\Sigma X_j f_j}{n} = \frac{1610.0}{40} = 40.25$$

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

14

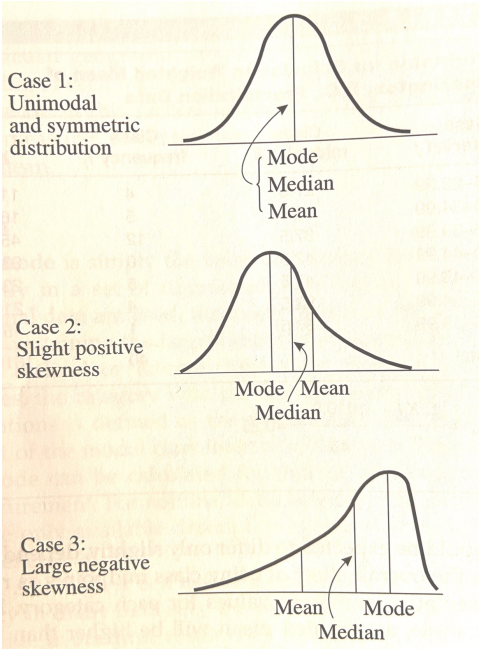
14

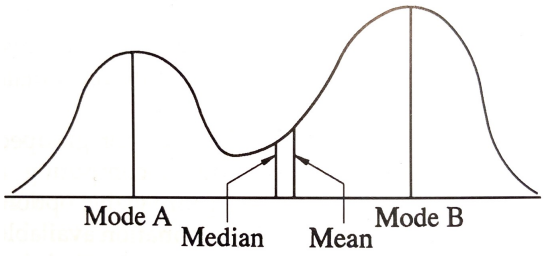
Selecting the Appropriate Measure of central Tendency

- Depends on the characteristics of the dataset
  - Unimodal and symmetric: all three centers will be similarly located and are all equally effective
  - If the frequency distribution has a degree of skewness, the measures of the center will be positioned at the different places

15

15





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

16

16

Selecting the Appropriate Measure of central Tendency

- Depends on the characteristics of the dataset
  - If the frequency distribution is bimodal (two modes) or multimodal (more than 2 modes), the mean and median will not provide meaning descriptions
  - If the frequency distribution contains one or more extreme values (**outliers**), the mean will be heavily influenced by these values
  - The existence of extreme values is indicated by skenwness

17

17

Measures of Dispersion and Variability

- Measures provide an indication of the spread or variability in the data
  - Range
    - Difference between highest and lowest
    - If grouped data: the difference between the upper value of the highest class and the low value of the lowest class
  - Deviation
    - The difference between each value and the mean

$$d_i = X_i - \bar{X}$$

18

18

Measures of Dispersion and Variability

- Average or Mean Deviation
  - The difference between each value and the mean

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

Where $|X_i - \bar{X}|$ is the absolute value of the deviation

19

19

Measures of Dispersion and Variability

- The deviation and its properties
  - The sum of all the deviations about the mean is always zero
    - The absolute value of the deviation makes all the negative deviations positive (see average deviation)
  - The sum of all the squared deviation about the mean is less than the sum of all the squared deviations about any other number

$$X: \min \sum (X_i - \bar{X})^2$$

20

20

Measures of Dispersion and Variability

- Standard Deviation

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

standard deviation for sample (≈ n<30)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$$

standard deviation for population (≈ N>30)

- Note that if n >30 the difference between (n-1) and (n) is small

21

21

Measures of Dispersion and Variability

- Variance

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

$$\sigma^2 = \frac{\sum_{i=n}^{n}(x_i - \mu)^2}{N}$$

22

22

Measures of Dispersion and Variability

- 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

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

23

23

Correlation

- Often one has to investigate if there is a relationship between two or more variables
  - Relationships total
    - Total household income and total monthly rent
    - Total population and the number of retail stores
    - Crime locations and the distance from police stations
- Correlation is a statistical method used to determine if a relationship between variables

24

24

Correlation

- Any two variables can be correlated, and the strength and direction of relationship determined.
  - caution must be used when evaluating or interpreting correlations.
  - A relationship does not necessarily imply the existence of a causal relationship.

25

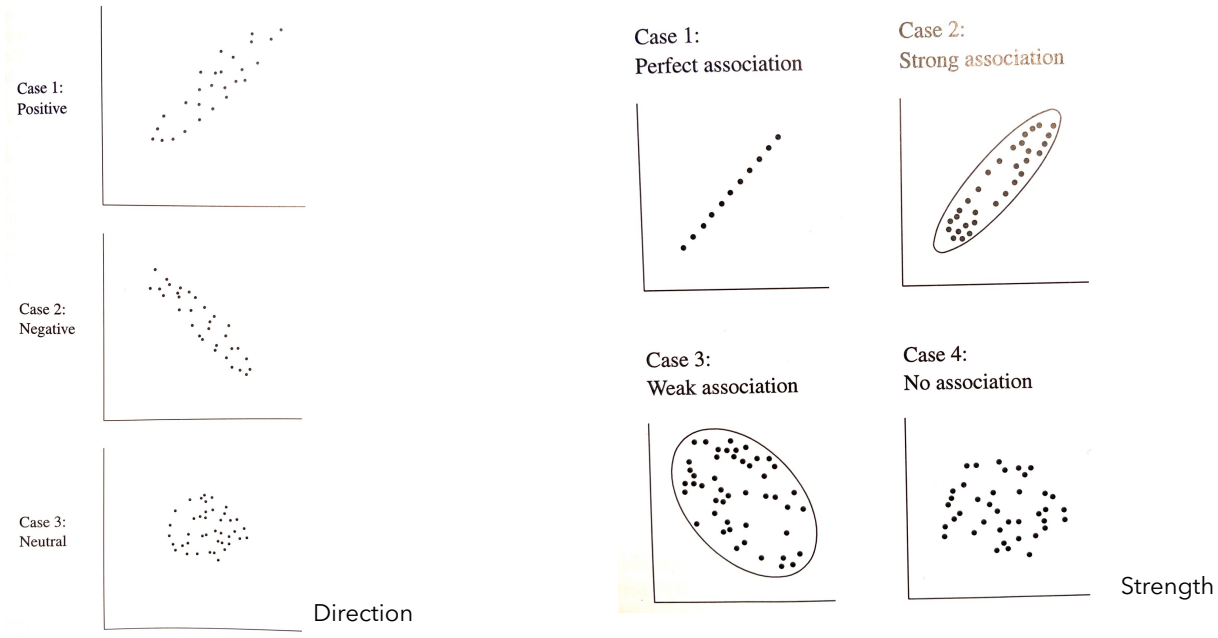
25

Correlation

- Scatterplots are a common tool used to portray the relationship or association between variables
- Scatterplots provide visual information about
  - Strength of relationship
  - Direction of relationship

26

26



27

Scatterplots

- Direction
  - Positive relationship
    - Increasing values in one variable correspond to increasing values in another variable
    - Decreasing values in one variable correspond to decreasing values in another variable
  - Negative (inverse) relationship
    - Increasing values in one variable correspond to decreasing values in another variable
    - Decreasing values in one variable correspond to increasing values in another variable
- Strength of relationship
  - Determined by the amount of spread in a scatterplot

28

28

State	MM	Celsius
Alabama	1480	17.1
Alaska	572	-3
Arizona	345	15.7
Arkansas	1284	15.8
California	563	15.2
Colorado	405	7.3
Connecticut	1279	9.4
Delaware	1160	12.9
Florida	1385	21.5
Georgia	1287	17.5
Hawaii	1618	21.1
Idaho	481	6.9
Illinois	996	11
Indiana	1060	10.9
Iowa	864	8.8
Kansas	733	12.4
Kentucky	1242	13.1
Louisiana	1528	19.1
Maine	1072	5

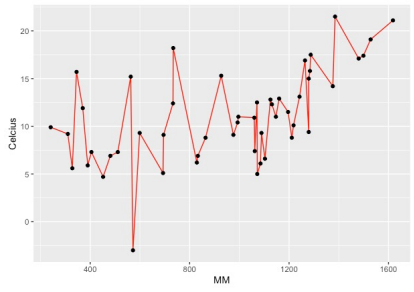
State	MM	Celsius
Maryland	1131	12.3
Massachusetts	1211	8.8
Michigan	833	6.9
Minnesota	693	5.1
Mississippi	1499	17.4
Missouri	1071	12.5
Montana	390	5.9
Nebraska	599	9.3
Nevada	241	9.9
New Hampshire	1103	6.6
New Jersey	1196	11.5
New Mexico	370	11.9
New York	1062	7.4
North Carolina	1279	15
North Dakota	452	4.7
Ohio	993	10.4
Oklahoma	927	15.3
Oregon	695	9.1
Pennsylvania	1089	9.3
Rhode Island	1218	10.1

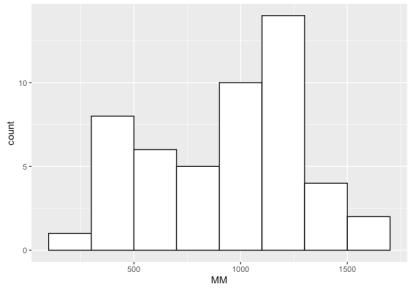
State	MM	Celsius
South Carolina	1264	16.9
South Dakota	511	7.3
Tennessee	1376	14.2
Texas	734	18.2
Utah	310	9.2
Vermont	1085	6.1
Virginia	1125	12.8
Washington	976	9.1
West Virginia	1147	11
Wisconsin	829	6.2
Wyoming	328	5.6

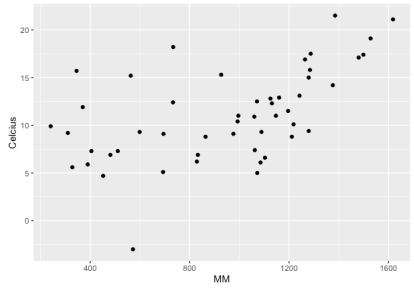
US. Annual Precipitation and Temperature

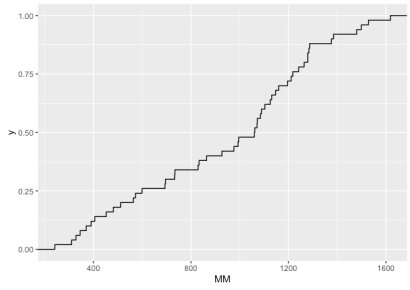
29

29









GRAPHING

30

30

Hypothesis Testing

- A hypothesis is an informed explanation, prediction, or supposition about something
  - Informed because there is evidence (but limited)
  - It has to be testable to so that one can conclude if its true or not
  - It has to be falsifiable (a possibility exists that it can be proven false)
- Bad hypothesis examples
  - Water is better than fire: not testable, criterion/variable is undefined
  - One can live to be 300 years old: not false falsifiable because no one gets to live that long

31

31

Hypothesis Testing

- A procedure that leads to a conclusive statement regarding a specified hypothesis
- The general goal is to make an ***inference*** about the magnitude of one or more population parameters based on sample data

32

32

Hypothesis Testing

- Two complimentary 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

H_0: *Sample Statistic = Population Parameter*
H_A: *Sample Statistic ≠ Population Parameter*

33

33

Hypothesis Testing

- The goal is to calculate probability that the null hypothesis is true
- If this probability is acceptably low
  - (meaning that it is likely not true)
  - then the null hypothesis is rejected in favour of the alternate hypothesis.
  - The sample results can be said to be significantly different

34

34

Hypothesis Testing

- Steps in hypothesis testing
  1. State the null and alternate hypotheses
  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 hypotheses

35

35

State the null and alternate hypotheses

- Consider the formulation of a hypotheses concerning the mean of the population

$$H_0: \mu = \mu_H$$

$H_A: \mu \neq \mu_H$ (nondirectional)
$H_A: \mu < \mu_H$ (directional)
$H_A: \mu > \mu_H$ (directional)

36

36