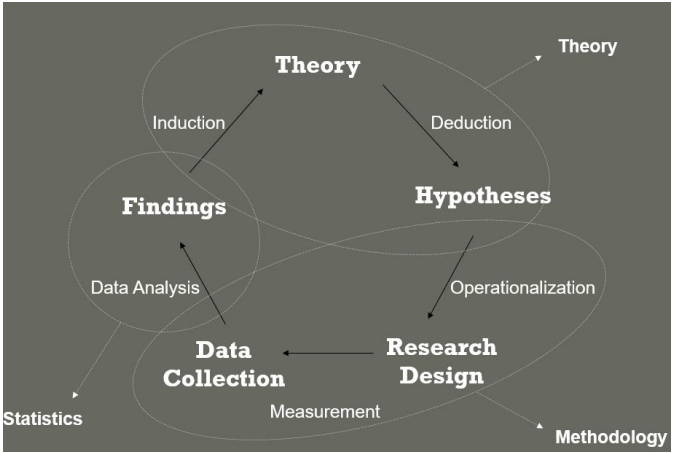
Lecture 01 - Introduction, Types of Variables, LoM, and Data Summarization/Visualization

Data Analysis (CJUS 6103)

Science, Methods, & Replication

- What is Science?
  - "The process of gathering and analyzing data in a systematic and controlled way using procedures that are generally accepted by others in the discipline."
- Methods? Replication?
  - Methods: Procedures used in processing and analyszing data.
  - Replication: Repeating methods with the same data in an effort to achive the same results.

Wheel of Science



Types of Statistical Research

- Evaluation
  - Examines program/law/law enforcement outcomes
- Exploratory
  - Examines unfamiliar topics lacking previous research
- Descriptive
  - Examines a particular phenomenon in a particular sample

Populations and Samples

- What is a Population?
  - 'The universe of people, objects, or locations that researchers wish to study.'
- What is a Sample?
  - 'A subset pulled from a population with the goal of ultimately using the people, objects, or places in the sample as a way to generalize to the population.'
    - What is Probability Sampling?

Types of Variables and Levels of Measurement

Outline

- Units of Analysis
- Variables
- Levels of Measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Levels of Measurement Examples

Units of Analysis

What is a unit of analysis?

- The objects or targets of a research study.
- Examples:
  - A researcher is examining homicide patterns over time by gathering yearly city-level data for the years 1980-1985.
    - What is the unit of analysis?
  - A researcher is aiming to evaluate the effectiveness of 'fighting back' during a violent incident as it relates to injuries.
    - What is the unit of analysis?

Variables

- What is a variable?
  - 'A characteristic that describes people, objects, or places and takes on multiple values in a sample or population.'
- What makes a constant different from a variable?
  - A constant only takes on one value, whereas a variable takes on multiple values.

Levels of Measurement

- Levels of Measurement and Computerization
  - Quantitative analyses require that data are computerized and numbers are assigned to all categories.
  - Level of measurement represents how these numbers are to be interpreted.
    - Are the numbers arbitrary, or do they have specific meanings?
  - Four Levels of Measurement
    - Nominal, Ordinal, Interval, Ratio (NOIR)

Nominal Variables

- What is a nominal variable?
  - A variable whose numeric values can only be interpreted as representing distinct categories (i.e., 1=red; 2=blue; ...).
- What are the properties of these categories?
  - Distinct
  - Mutually Exclusive
  - Exhaustive
- What are some example of nominal variables?
  - Religion
  - Marital status
  - Race

Ordinal Variables

- What is an ordinal variable?
  - A variable whose values can be rank-ordered in terms of 'more than' or 'less than'.
- Values have a logical order:
  - Likert Scales:
    - Opinions – (1) Strongly Disagree; (2) Disagree; (3) No Opinion; (4) Agree; (5) Strongly Agree
  - Year in School – (1) Freshman; (2) Sophomore; (3) Junior; (4) Senior

Interval Variables

- What is an interval level variable?
  - A variable whose distance between two adjacent values is fixed and known.
  - Allows more detailed descriptions of why one value is 'more than' or 'less than' the other.
  - Value of '0' is arbitrary, and does not signal that the phenomenon in question is entirely absent.
    - 0 degrees Fahrenheit is not an absence of temperature.
- What are some examples of interval variables?
  - Temperature
  - IQ Tests

Ratio Variables

- What is a ratio level variable?
  - A variable whose value of zero is meaningful (making multiplication and division possible).
  - i.e., a homicide rate of 10 per 100,000 is twice that of 5 per 100,000.
- What are some examples of ratio level variables?
  - Age
  - Years of education
  - Weight
  - Income

Review of Level of Measurement

Nominal	Ordinal	Interval	Ratio
Distinct			
Exclusive			
Exhaustive			
	Rank Order		
		Equal Intervals	
			True Zero

Levels of Measurement Examples

What level of measurement is applicable for each of the following?

- Property crime rate
- Crime type (violent, property, drug)
- Sentence length (in months)
- Fear of crime (1-10)
- Conviction status (convicted/not convicted)
- Crime seriousness (1-100)
- Number of arrests

Alternative Ways of Classifying Variables

- Qualitative (Nominal/Ordinal)
  - Variables tel us "what kind," "what group," or "what type."
  - Marital status, religion, race, liberalism, conservativism
- Quantitative (Interval/Ratio)
  - Variables tell us "how much," or "how many."
  - Hours worked a week, sentence length (in months), crime rates

Alternative Ways of Classifying Variables (cont.)

- Discrete
  - Values assume only a finite or countable number of alternatives
    - Number of crimes, family size, number of convictions
    - Cannot have part of any one of those units - e.g., no 1.5 crimes committed
- Continuous
  - Values can assume theoretically infinite number of values between any two points on a scale
    - Crime rates per 100,000, GPA, arrest rates
    - Can have 3.5 murders per 100,000 population or 13.33 arrests per 100 reported offenses

Alternative Ways of Classifying Variables (cont.)

Independent Variable (X) = "Cause," "Predictor," "Regressor," "Covariate," list goes on. . . .(sometimes referred to as controls, also)

Dependent Variable (Y) = "Effect," "Outcome," "Response," list goes on. . . (but not as far as for IVs)

Note: not so much a way of classifying measurement, but classifying what side of a regression equation the variable is on. Left-hand side variables are the DV, right-hand side variables are the IVs.

REMEMBER: An independent variable in one context can be the dependent variable in another. The side of the equation they are on depends entirely on the research question.

Examples

Nationwide, the average starting salary for entry-level police officers is about $24,000. You believe that the location of police departments in urban vs. rural areas influences starting salaries. In a random sample of 90 police departments, you find that the average starting salary is $25,000 in urban departments and $24,000 in rural departments.

- Unit of observation = Police department
- Population = All police departments in U.S.
- Sample = 90 police departments
- Dependent variable = Starting salary
  - Ratio level, continuous (also discrete)
- Independent variable = Location
  - Nominal

Examples (cont.)

Many sociologists argue that poverty is a cause of criminal behavior. To test this claim, you collect data from a random sample of 250 counties across the U.S. You obtain measures of the poverty rate (% population living below poverty line) and the number of offenses reported to the police.

- Unit of observation = County
- Population = All counties in U.S.
- Sample = 250 counties
- Dependent variable = Number of offenses reported to police
  - Ratio level, discrete
- Independent variable = Percent living below poverty line
  - Ratio level, continuous

Examples (cont.)

According to some theorists, teenagers who spend more time with their friends in unsupervised activities are more likely to engage in delinquent conduct. You question a random sample of 2,500 high-school youths about number of hours spent in unsupervised peer activities (e.g., cruising, shopping, movies) and frequency of delinquent behavior.

- Unit of observation = Individual
- Population = All high-school youth
- Sample = 2,500 high-school youths
- Dependent variable = Frequency of delinquent behavior
  - Ratio level, discrete
- Independent variable = Hours in unsupervised peer activities
  - Ratio level, continuous

Why Does Level of Measurement Matter?

Level of measurement determines what type of statistical test is appropriate.

Tests are chosen based upon the lecel of measurement of the IVs and DV.

- Examples
  - X = Sex, Y = Employment(Y/N) = Chi-Square Test
  - X = Sex, Y = Number of Arrests = t-Test
  - X = Crime Type, Y = Sentence Length (in months) = F-Test
  - X = Divorce Rate, Y = Homicide Rate = Correlation and Regression

Data Summarization

Section Outline

- Proportions & Percents
- Rates & Ratios
- Proportion/Percent Difference (Change)
- Frequency Distributions with...
  - Nominal data
  - Ordinal data
  - Interval/Ratio data

Summarizing Data

- Sample size (n)
  - Total number of observations in a sample
- Frequency (f)
  - Count or number of observations in a subset of the sample

Proportion and Percent

- Proportion (p)
  - Ratio of the numb er of observations in a subset of the sample to the total number of cases in the sample (relative frequency)
  - $p = \frac{\text{\# in subset of sample}}{\text{Total \# of cases in sample}} = \frac{f}{n}$
- Percent (%)
  - $\% = \frac{f}{n} * 100$

Ratios and Rates

- Ratio
  - Expresses the relationship between two values, indicating their relative sizes.
  - e.g., 2:1 indicates that the first value occurs twice as much as the second value
- Rate
  - Ratio of the number of occurrenses of an event to the population at risk for experiencing the event.
  - Rate $= \frac{\text{\# of occurrences of an event}}{\text{estimated population at risk}} * 10$ to some power

Rates (cont.)

- Birth Rates = X 10^3 (per 1,000 women of childbearing age)
- School expenditure = X 10^0 (per pupil)
- Crime rates = X 10^5 (per 100,000 population)
  - In 2016, there were 332,198 robberies reported to the police, and the estimated population in the U.S. that year was 323,127,513
  - 102.8 robberies per 100,000 population
  - $\frac{332198}{323127513} * 100000 = 102.8$

Rates (cont.)

Difference between a rate and a proportion?}

- Proportions cannot lie outside 0,1 interval.
  - Rates can be anywhere between 0 and ∞
- Denominator of a proportion is a fixed sample size (n)
  - Denominator of a rate is often an **estimated** population size

Calculating Crime Rates

Which region had the most robberies in 2016?

Region	Count	p
Northeast	53033	0.16
Midwest	64022	0.19
South	128842	0.39
West	86301	0.26
Total	332198	1.00

Why does the South have more robberies?

- More dangerous?
- More people?

Calculating Crime Rates (cont.)

Does the South have the highest crime rate?

Region	Count	Population	Rate per 100k
Northeast	53033	56209510	94.35
Midwest	64022	67941429	94.23
South	128842	122319574	105.33
West	86301	76657000	112.58

- West is actually more dangerous (by a hair)
  - Rates account for differences in the **eligible** population, or the **population at risk**

Proportional Difference

- Proportional Difference (Change)
  - Comparison of a single variable across two time periods, or simply to compare the relative magnitudes of two values.
- Typically multiplied by 100 to get a **percent difference**
- Percent change $= \frac{\text{Time 2 - Time 1}}{\text{Time 1}} * 100$ or $\frac{\text{Comparison - Baseline}}{\text{Baseline}} * 100$

Proportional Difference (cont.)

- Change in homicide rate
  - 2011 Rate = 4.7 per 100,000
  - 2016 Rate = 5.3 per 100,000
- How large was the increase in the homicide rate?
  - Percent Change = $\frac{\text{Time 2 - Time 1}}{\text{Time 1}} * 100 = \frac{5.3 - 4.7}{4.7} * 100 = 12.8$
- Homicide rate increased 12.8% from 2011 to 2016

Proportional Difference (cont.)

Homicide rate (per 100,000) by region:

Region (# of states)	2011	2016	% Change
New England (6)	2.6	2.0	-23.1
Middle Atlantic (3)	4.4	4.0	-9.1
East North Central (5)	4.9	6.4	30.6
West North Central (7)	3.4	4.3	26.5
South Atlantic (8)	5.4	6.4	18.5
East South Central (4)	5.7	7.3	28.1
West South Central (4)	5.4	6.3	16.7
Mountain (8)	4.4	4.7	6.8
Pacific (5)	4.1	4.4	7.3

Which region experienced the largest change?

Proportional Difference (cont.)

- Alternative way to compute percent difference
  - Subtract 1.0 from the ratio of the comparison (T2) to baseline (T1)

$$\text{Percent difference} = \left(\frac{\text{Time 2 - Time 1}}{\text{Time 1}}\right) * 100 = \left(\frac{\text{Time 2}}{\text{Time 1}} - \frac{\text{Time 1}}{\text{Time 1}}\right) * 100 = \left(\frac{\text{Time 2}}{\text{Time 1}} - 1\right) * 100$$

Proportional Difference (cont.)

- Proportional/Percent difference can be misleading with a low baseline.
  - A 100% increase from a baseline of 1 is not as notable as a 100% increase from a baseline of 50
- Useful to report both the baseline value as well as the proportion/percent difference (change)

Frequency Distributions

- A kind of data table that is a usefule way to summarize data.
  - Table should list categories and frequencies, at a minimum.
  - Often lists proportions and percents, also.
- Provides visualization of how cases are spread out across categories/values.

Frequency Distributions with Nominal Data

- Display f, n, p, %

Family Structure	f	p	%
Both Biological Parents	3350	.483	48.3
One biological/one step	1002	.145	14.5
Biological mom only	1972	.285	28.5
Biological dad only	233	.034	3.4
Other family member	372	.054	5.4
Total	6929	1.001	100.1

- HINT: Make sure the proportions add up to within rounding error of 1.0

Frequency Distributions with Nominal Data (cont.)

- Can be informative to compare 2+ groups

	White Youth			Black Youth		
Family Structure	f	p	%	f	p	%
Both bio parents	2743	.594	59.4	607	.263	26.3
One bio/one step	706	.153	15.3	296	.128	12.8
Bio mom only	864	.187	18.7	1108	.480	48.0
Bio dad only	172	.037	3.7	61	.026	2.6
Other family	136	.029	2.9	236	.102	10.2
Total	4621	1.000	100	2308	.999	99.9

- How does family structure differ?

Frequency Distributions with Ordinal Data

- Can add cumulatives (cf, cp, c%)
  - Cumulative frequency is the frequency of a row plus all preceding rows

Grades in 8th	f	p	%	cf	cp	c%
A's & B's	3186	.370	37.0	3186	.370	37.0
B's & C's	3238	.376	37.6	6424	.746	74.6
C's & D's	1850	.215	21.5	8274	.961	96.1
D's & F's	330	.038	3.8	8604	.999	99.9
Total	8604	.999	99.9			

- Notice that the cf of the last row is equal to n

Women Really Are Smarter

- Use % to determine if young males or females perform better academically.
  - What percentage get B's & C's or better?

	Males		Females	
Grades in 8th	%	c%	%	c%
A's & B's	29.3	29.3	45.2	45.2
B's & C's	38.9	68.2	36.3	81.5
C's & D's	26.8	95.0	16.0	97.5
D's & F's	5.0	100.0	2.6	100.1
Total	100.0		100.1	

Frequency Distributions with Discrete, Interval-Ratio Data

- Do delinquents spend their free time differently than non-delinquents?

# of Weekdays Do Homework?	Non-delinquents %	Non-delinquents c%	Delinquents %	Delinquents c%
0	8.4	8.4	14.8	14.8
1	3.3	11.7	5.1	19.9
2	7.5	19.2	10.4	30.3
3	17.5	36.7	21.0	51.3
4	24.4	61.1	19.8	71.1
5	38.8	99.9	29.0	100.1
Total	99.9		100.1	

Frequency Distributions with Discrete, Interval-Ratio Data

- Number of months sentenced for armed robbery (n=40)
  - 36 38 39 47 50 51 51 53
  - 55 55 56 57 60 62 63 64
  - 64 66 67 68 69 70 70 70
  - 71 75 78 79 80 80 81 83
  - 85 86 87 89 95 98 99 99

- What would be the problem with creating a frequency distribution with these data?

Simple Frequency Distribution of Sentence Length

Sentence Length	f	Sentence Length	f	Sentence Length	f
36	1	62	1	79	1
38	1	63	1	80	2
39	1	64	2	81	1
47	1	66	1	83	1
50	1	67	1	85	1
51	2	68	1	86	1
53	1	69	1	87	1
55	2	70	3	89	1
56	1	71	1	95	1
57	1	75	1	98	1
60	1	78	1	99	2

Lots of 1s in the frequency columns - makes it more difficult to understand how sentence length is distributed.

Grouped Frequency Distributions

- A grouped frequency distribution can come in handy for certain types of data.
  - Discrete, interval-ratio data with a large number of values.
  - Continuous, interval-ratio data.
- Groups values into 'classes' or intervals.
- Frequency distribution of these classes.
- Transforms quantitative data into qualitative (ordinal) data.

Steps in Creating a Grouped Frequency Distribution

- Arrange raw data in ascending order
- Choose the number of intervals (generally 5-10 will do)
- Determine the width of the intervals
  - Calculate the range of the data (99-36=63)
  - Divide by the # of desired intervals (63/6=10.5)
  - Round to a convenient interval width (10)
    - A multiple of five is usually the easiest

Steps in Creating a Grouped Frequency Distribution (cont.)

- Construct the interval limits
  - Choose the lower limit of the 1st interval (35)
    - Again, easier if the lower limit is a multiple of five
  - Add interval width to get the 1st interval (35-44)
    - You can't just add the interval width to the 1st lower limit; you must include the lower limit in your count
  - Construct non-overlapping intervals such that the 1st interval contains the smallest value and the final interval contains the largest value
- Tally the number of cases that fall into each interval
  - Make sure the frequencies add up to n

Steps in Creating a Grouped Frequency Distribution (cont.)

Sentence Length	f	p	%
35 - 44	3	.075	7.5
45 - 54	5	.125	12.5
55 - 64	9	.225	22.5
65 - 74	8	.200	20.0
75 - 84	7	.175	17.5
85 - 94	4	.100	10.0
95 - 104	4	.100	10.0
Total	40	1.0	100.0

Frequency Distributions with Continuous, Interval-Ratio Data

- State-level unemployment rates, 2016
  - 2.8 2.8 3.0 3.2 3.2 3.3 3.3 3.4 3.7 3.7
  - 3.8 3.9 3.9 4.0 4.0 4.1 4.1 4.2 4.3 4.4
  - 4.4 4.5 4.6 4.8 4.8 4.8 4.9 4.9 4.9 4.9
  - 4.9 5.0 5.0 5.1 5.1 5.3 5.3 5.3 5.4 5.4
  - 5.4 5.4 5.7 5.8 5.9 6.0 6.0 6.1 6.6 6.7

- $\sim$ Intervals; Range = 6.7-2.8=3.9
  - Width = 3.9/9 = .43 - $\sim$ .5
  - 1st interval: 2.5-2.9

Frequency Distribution for State-Level Unemployment

This one is for you to finish on your own (I'll provide the answer in the lecture notes).

Interval Limits	f	p	%
2.5-2.9			
Total			

Graphing Data

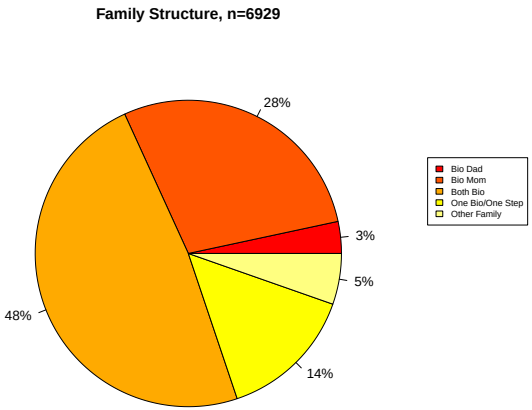
Outline

- Graphing Qualitative Data
  - Pie Charts
  - Bar Graphs
- Graphing Quantitative Data
  - Histogram
  - Polygon
- Other Useful Graphing Techniques
  - Time Series
  - Map

Graphing Qualitative Data

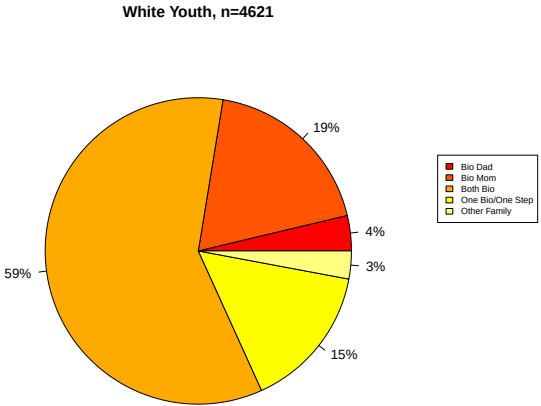
- Pie Charts
  - Each category receives a "slice" of the pie
  - Best if there are no more than five categories to make interpretation clear
  - Label categories; percents should sum to 100
- Bar Graphs
  - Each category gets a bar and a label
  - Use spaces between bars to imply distinct categories (especially for nominal data)
  - Y-axis xan measure f, p, or %
  - Put the frequency (or p or %) above each bar

Family Structure - Pie Chart



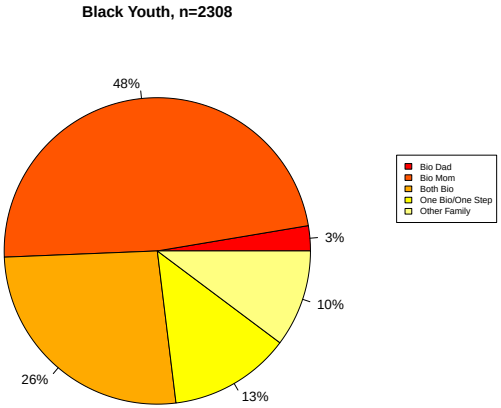
Family Structure by Race - Pie Charts

White Youth

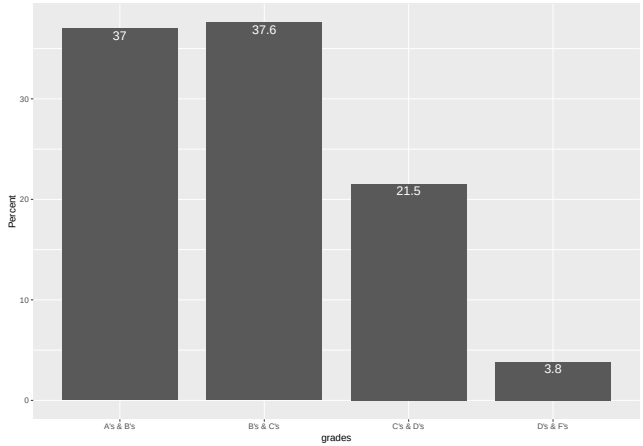


Family Structure by Race - Pie Charts

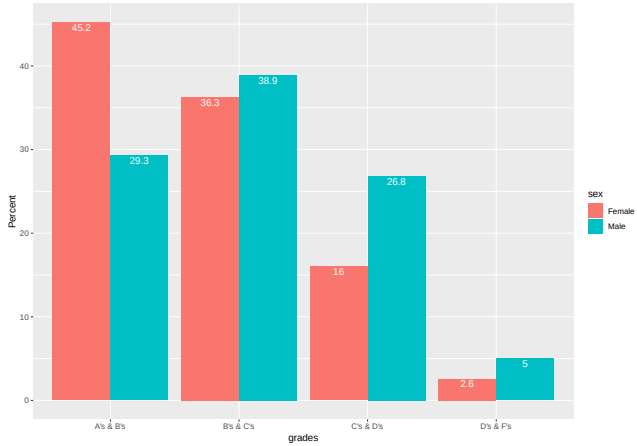
Black Youth



Scholastic Performance Bar Graph



Scholastic Performance Bar Graph

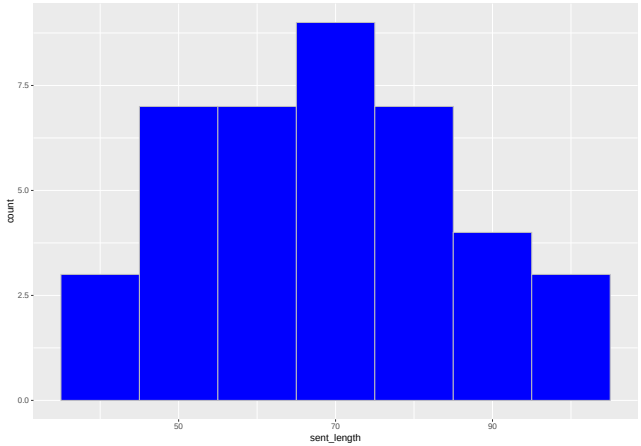


Graphing Quantitative Data

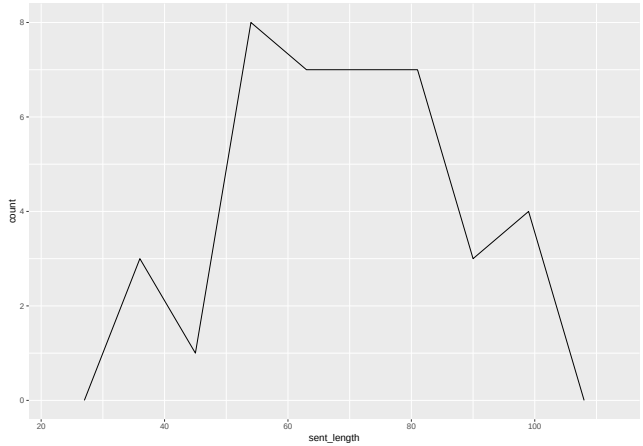
- Histogram
  - Bars should touch to imply that original data are quantitative in nature (interval-ratio)
  - Y-axis can be measured with f, p, or %
  - X-axis should be labeled with values or interval limits
- Polygon
  - Dot placed at midpoint of each interval, with a line to connect the dots together
  - Line should connect to the x-axis

Both types of graphs provide visualize evidence of skew in the distribution of values.

Robbery Sentence Length Histogram



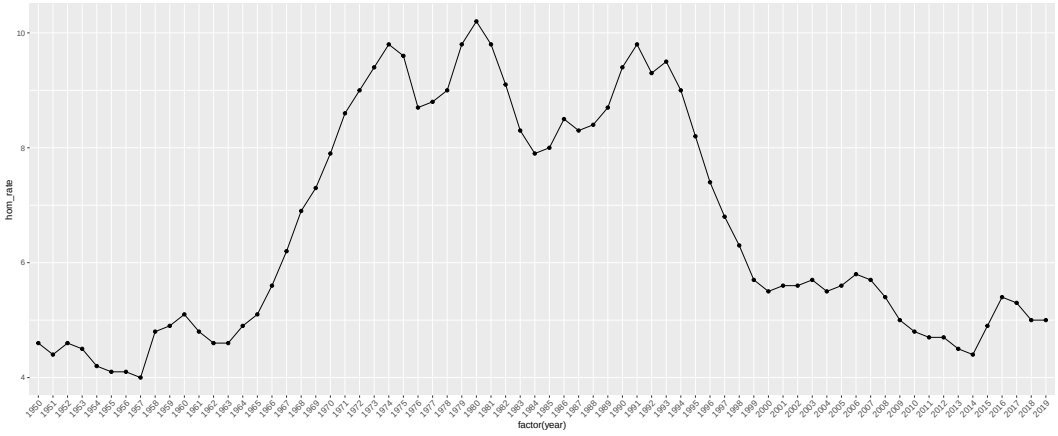
Robbery Sentence Length Polygon



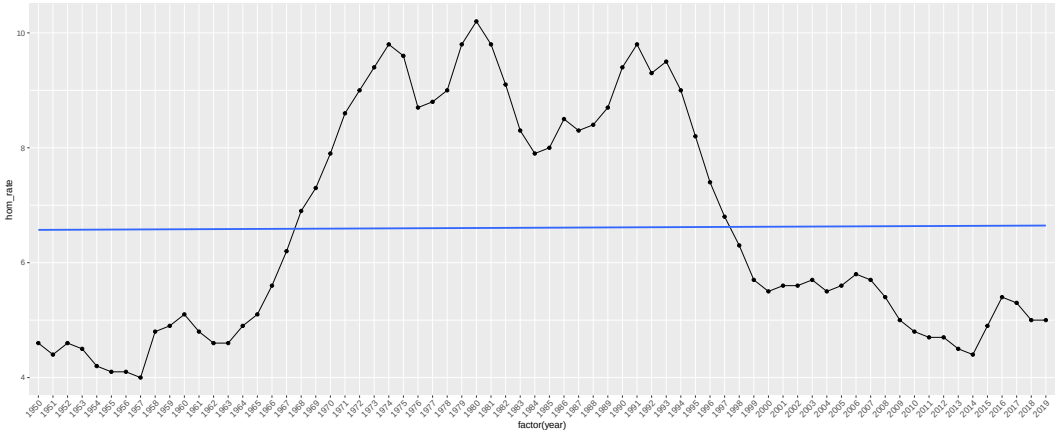
Other Useful Graphing Techniques

- Time Series
  - Line graph where X-axis represents time, often in years
  - Displays temporal trends
    - Can overlay a trend line to see how some characteristic increases or decreases over time.
- Map
  - Useful for displaying census-related data
  - Displays spatial trends

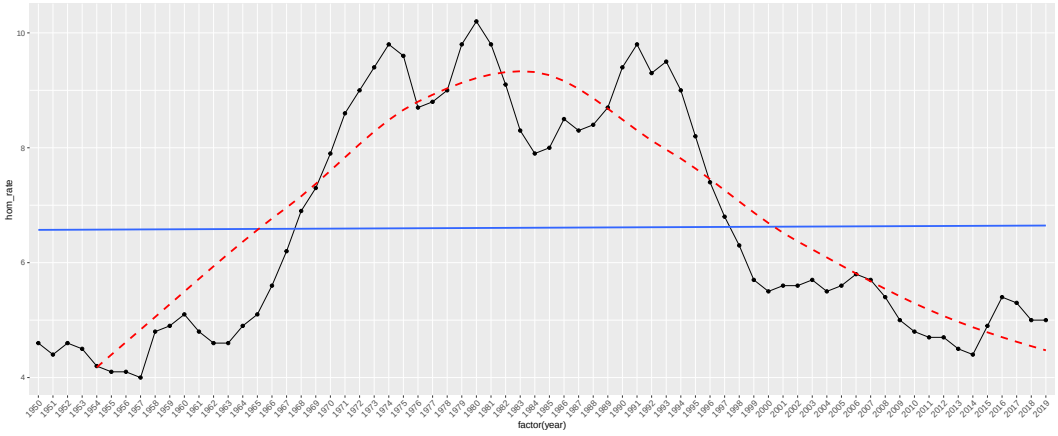
Homicide Rates, 1950 - 2019 Time Series



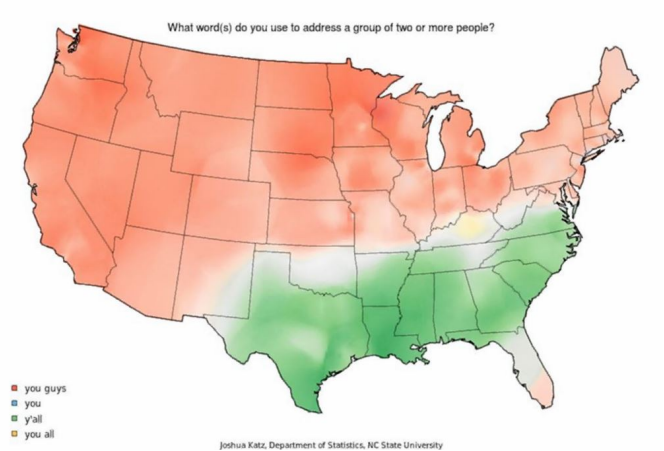
Homicide Rates, 1950 - 2019 Time Series with Trend Line



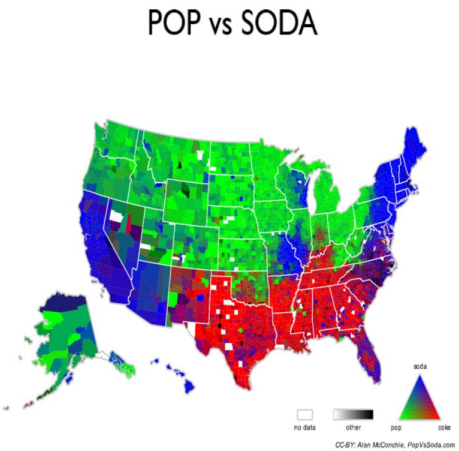
Homicide Rates, 1950 - 2019 Time Series with (Better) Trend Line



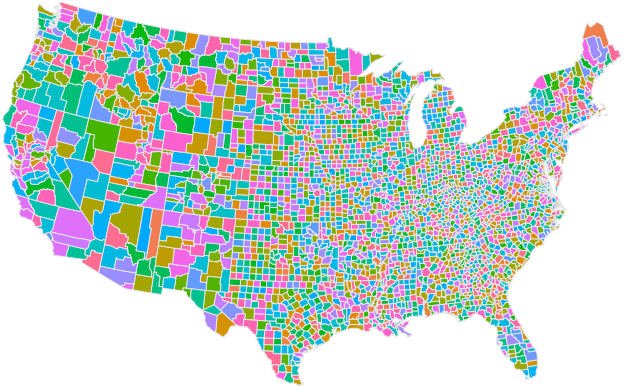
Y'All Use Y'All?



Pop Versus Soda



You Can Even Make a Map in R!



The End

Time for your Two Questions!