Lecture 01 - Basic Statistical Concepts, Descriptive Statistics, and Data Visualization

Data Analysis in CJ (CJUS 6103)

Basic Statistical Concepts

Outline

I. Categories of statistics.
II. Basic statistical definitions.
III. Variable measurement.
IV. Variable classification.
V. Summarizing Data
VI. Frequency Distributions
VII. Data Visualization

I. Categories of Statistics

Descriptive Statistics (Data Reduction)

The branch of statistics that is concerned with describing, displaying, and summarizing data. These are numerical and graphical methods that facilitate interpretation of a large amount of data. These techniques reduce a series of numbers down to a small number of more convenient and more easily communicated descriptive terms. In a sense, descriptive statistics provide a "snapshot" of a group of data.

Inferential statistics (decision making, hypothesis testing) – the branch of statistics that is concerned with making inferences from sample data to a population. In essence, statistical inference involves generalizing on the basis of limited information. As such, there is bound to be error in making these generalizations (measurement error, sampling error). Inferential statistics allow us to take into account this uncertainty and to make generalizations with a reasonable amount of confidence.

Let us consider an example. The university issues me a class roster. From this roster I can use counts, ratios, percentages, or averages to describe the characteristics of this class. (Gender, year in school, major, residency status, grade point average) This is the idea behind descriptive statistics. They provide a snapshot of the composition of this specific class. Now, what if I wanted to generalize this class information to the whole campus? Suppose that I want to know the gender composition and GPA of the average student at the University of North Carolina - Charlotte. Assuming that this class is representative of all classes here (which it probably is not), I can use inferential statistics to estimate these quantities for the entire student body.

II. Basic Statistical Definitions

Population

The largest set of cases or people in which a researcher is actually interested. For example, we may want to ask questions about the entire U.S. population, about all youths ages 12 to 16, about all single-parent families, about all persons arrested by the police, about all sentencing decisions made by judges, about all women, etc. There are several practical problems with collecting information on a whole population, including time, cost, and the fact that it is sometimes impossible to get information on an entire population.

Sample

A subset of the population that a researcher uses to make generalizations about the population. In essence, we want to use what we know about a sample to understand what we do not know about a population.

Random Selection (Probability Sampling)

A way of ensuring that a sample is representative of the population from which it is drawn. In a simple random sample, each element of the population has a known, non-zero, independent, and equal probability of being selected into the sample. There are other types of random sampling procedures that are commonly used, such as systematic random sampling, weighted sampling, and multistage cluster sampling. You should note that there are no statistical techniques that we can use to make a non-representative sample more representative of the population.

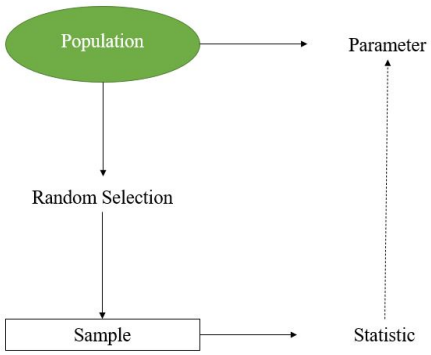
Parameter

An unknown (but knowable) characteristic of the population of interest. This characteristic is fixed; we assume that there is one true value, or one "right" answer. But, for a variety of reasons we do not know what the right answer is, so we use various statistical techniques to estimate the parameter.

Statistic

A characteristic derived from sample data that is an estimate of a population parameter. A statistic represents our "best guess" as to the value of an unknown parameter. It should be mentioned that a sample statistic can and will vary from one sample to the next. This means that there is error in estimating the true population parameter. This becomes an important idea when we move on to statistical inference.

Figure 1: Diagram of Statistical Inference



III. Variable Measurement

Unit of observation

The element that is being studied or observed, and thus the level at which data are collected. In much criminological research, the individual is the unit of observation. For example, are interested in studying whether youths who make poor grades in school are more likely to smoke cigarettes or skip classes. We want to know if participation in a drug treatment program reduces subsequent drug use. We want to know if gender, race, and class influence sentencing decisions. We want to know if citizens have favorable attitudes toward community policing initiatives. The list goes on. However, sometimes aggregate units are studied, such as the family, neighborhood, census tract, state, or country. For example, we want to know if increasing globalization is related to crime rates across countries. We want to know if neighborhood social disorganization causes violent crime. We want to now if greater per capita expenditure on policing has crime-control benefits in urban areas. So on and so forth.

Variable

An attribute or characteristic that can take on different values from one individual to another or from one point in time to another. An example of a variable in this classroom is sex (not yes or no, but male or female), since it can take on the value of male or female. In other words, sex varies from one person to another in this classroom. The same goes for year in school, which can take on the value of freshman, sophomore, junior, senior, 5th year senior, and career student.

Often, a variable is some phenomena that a researcher is trying to learn something about. For example, crime is a variable; it can take on the value of zero offenses, one offense, two offenses, and so on. As a researcher, I am interested in trying to explain why the number of crimes committed varies from one person to another.

Constant

An attribute that does not vary. For example, in this class, everyone is enrolled in statistics, so this would be a constant. Enthusiasm for statistics is also a constant, since everyone is equally excited to be taking this class. If I took a sample of male 21 year olds, gender and age would be constants. If I took a sample of inmates, having a prior arrest would be a constant, since everyone in the sample would have been arrested, by definition.

It is important to keep in mind that a phenomenon can be both a variable and a constant at different times, depending on the context of the research.

IV. Variable Classification

Classification #1: Levels of Measurement (see Figure 2)

Nominal

Values represent categories or qualities only. These values take on a limited number of manifestations that differ only in the labels assigned to them. They are distinct (easily separated), mutually exclusive (a case cannot be in more than one category), and exhaustive (there is a category for every case). Examples include gender, race, and political orientation:

Gender	Race	Political orientation
1 Male	1 White	1 Democrat
2 Female	2 Black	2 Republican
	3 Asian	3 Independent
	4 Other	

Although researchers often assign a different number to each category in order to distinguish them, these numbers do not mean that one category has more or less of "something" than another category. Thus, the numerical values are arbitrary.

Ordinal

Values represent categories that can be rank ordered in terms of "more than" or "less than." In other words, the categories have some relationship to each other. Although there is rank order (i.e., magnitude), we cannot know the magnitude of the difference among cases that fall into different categories. Examples include year in school, weekly income, and level of agreement:

Year in school	Socioeconomic status	Level of agreement
1 Freshman	1 Lower class	1 Strongly disagree
2 Sophomore	2 Middle class	2 Disagree
3 Junior	3 Upper class	3 Neutral
4 Senior		4 Agree
		5 Strongly agree

Although someone who "strongly agrees" has more agreement than someone who "agrees," it is not clear what the exact magnitude of this difference is, or that it is the same as the difference between people who "disagree" or "strongly disagree," or people who "agree" or are "neutral."

Interval

The distance between values is equal and known (i.e., equal intervals). This means that the magnitude of the attribute represented by a unit of measurement on the scale is the same regardless of where on the scale the unit falls. In other words, the difference between two adjacent values is the same at every point on the scale. Examples include temperature (with the exception of the Kelvin scale of temperature, which is measured at the ratio level), IQ, and GPA. The difference between 70F and 80F (10F) is the same as the difference between 90F and 100F. Importantly, the zero point on an interval scale is arbitrary; it has no real meaning, in that it does not imply a complete absence of the attribute being measured. A value of "0" does not imply that temperature does not exist or that a person has no IQ. Rather, it is an arbitrary designation that simply means that it is really cold or that someone is very unintelligent. Relatedly, a temperature of 80F does not mean that it is twice as warm as 40F, or an IQ of 120 does not mean that a person is twice as intelligent as someone with an IQ if 60. Thus, ratios of interval values are meaningless.

Ratio

The measurement scale has a true zero point, which implies a complete absence of the characteristic being measured. This scale is termed ratio because its properties allow ratio statements to be made about the attribute being measured. Examples include a person's weight, height, and income. In these examples, a value of "0" means that a person has no weight, no height, or receives no income. Moreover, someone who weighs 140 pounds has twice the weight as someone who weighs 70 pounds. Someone with $20 in their pocket has one-third as much as someone with $60.

Figure 2: Characteristics of Levels of Measurement

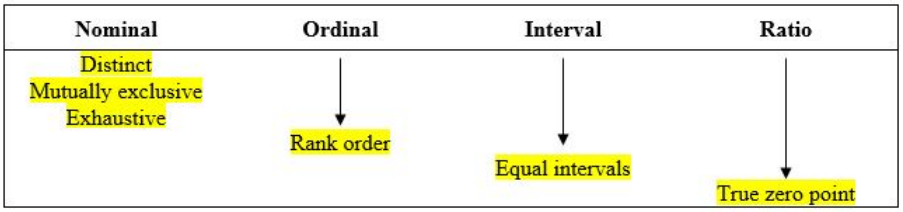
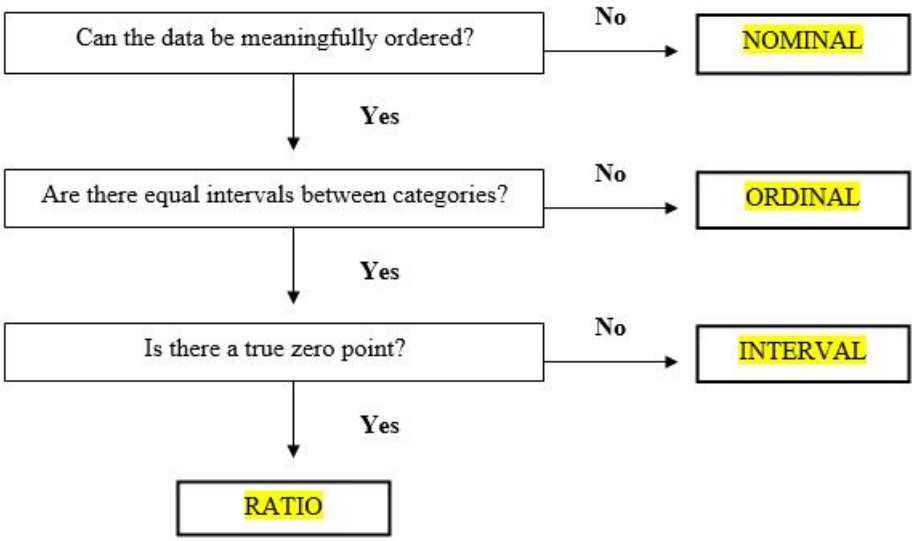


Figure 3: Determining the Level of Measurement



Consider the following variables that are commonly encountered in criminological research. For each, provide the level of measurement and give an explanation: property crime rate, crime type (violent, property, drug), sentence length (in months), fear of crime (1-10), conviction status (yes/no), crime seriousness (1-100), number of arrests.

If the difference between interval-level and ratio-level variables seems a bit confusing, it's because there really is some gray area. You should bear in mind that from a statistical standpoint, the distinction between the two is not really all that important. Often we are mostly concerned with whether or not a variable can be categorized as qualitative (nominal, ordinal) or quantitative (interval, ratio).

Classification #2: Qualitative vs. Quantitative

Qualitative (Categorical, Nominal/Ordinal) Variable

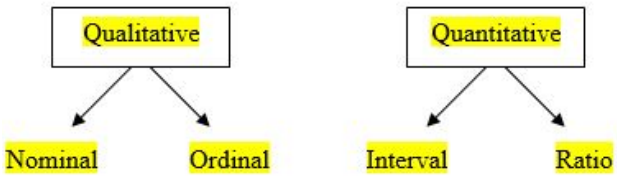
A variable in which the values represent qualities or categories only. They tell us what kind, what group, or what type a value is referring to. We cannot assign meaningful and precise numerical values to these categories.

Quantitative (Continuous, Interval-Ratio) Variable

A variable in which the values refer to quantities or numerical measurements. They tell us how much or how many of something a value has. In these cases, we assign precise and meaningful numerical values that can take on any of a number of possibilities.

Let's consider some examples. Work status (employed, not employed), gender (male, female), and family living arrangements (both biological parents, single parent, biological parent and step-parent, adoptive parents, foster parents) are qualitative variables. Work hours, age, and time spent on homework are quantitative variables.

Figure 4: Elaborating the Levels of Measurement



Classification #3: Causal Relationships Between Variables

Independent Variable (X)

The presumed cause. In the parlance of experimental studies, this is the variable that is manipulated or controlled by the experimenter. May also be referred to as a regressor, predictor, or right-hand side variables (because they are on the right-hand side of a regression equation).

Dependent Variable (Y)

The presumed effect or outcome. This is the variable that a researcher wishes to explain. It is the "dependent" variable because its value is assumed to "depend" on the independent variable(s). This variable is also known as an effect or a left-hand side variable (because it appears on the left-hand side of a regression equation).

We must keep in mind that defining a variable as independent or dependent relies solely on a researcher's assumptions about the underlying causal relationship between two variables. This may change according to the nature of the research problem. In other words, one researcher's independent variable is another's dependent variable. For example, consider the relationship between unemployment and crime. On one hand, there is reason to believe that being unemployed leads to greater involvement in criminal behavior due to a perceived lack of opportunity, financial strain, etc. In this case, unemployment is the independent variable, and crime is the dependent variable. On the other hand, there is equal reason to believe that criminal behavior that leads to arrest and conviction makes it more difficult to acquire a job, because a prison record serves as a stigma. In this case, crime is the independent variable and unemployment is the dependent variable.

As another example, consider the association between crime rates and the number of police. Is measured crime a cause or consequence of having more police?

Classification #4: Discrete vs. Continuous

Discrete Variable

A measure whose values can assume only a finite or countable number of alternatives. Examples include number of cigarettes smoked, number of criminal behaviors reported to the police, or the number of arrests an individual has (although.

Continuous Variable

A measure whose values can assume an infinite number of values between any two points on the scale. Since the values are infinitely divisible, there are an uncountable number of alternatives. Examples include crime rate per 100,000, grade point average, temperature, and height in meters.

The distinction between discrete and continuous variables applies to the realizations of the variable being measured and not to the scale used to measure the variable. Also, we usually only want to know if the dependent variable is discrete or continuous, because the statistical techniques that are appropriate for each are different. In practice, though many discrete variables are treated as continuous, such as hourly wages.

Examples

1) 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 = 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

2) 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

V. Summarizing Data (Basic Descriptives)

Sample size

The total number of observations in a sample (denoted n, sometimes N).

Frequency

The count or number of observations in a subset of the sample (denoted f).

Proportion/Percent

The ratio of the number of observations in a subset of the sample to the total number of cases in the sample. In other words, a proportion is a ratio of the frequency to the sample size. A proportion is often referred to as a relative frequency. A percent is simply a proportion times 100.

$$\text{Proportion (p)} = \frac{\text{\# in subset of sample}}{\text{Total \# of cases in sample}} = \frac{f}{n}$$

$$\text{Percent (\%)} = p * 100 = \left(\frac{f}{n}\right) * 100$$

Ratio

Expresses the relationship between two numbers, indicating their relative sizes. It is conventional to place the larger figure in the numerator.

Rate

The ratio of the number of occurrences of an event to the total number of possible occurrences. Rates are standardized using a base number that is the "estimated population" (like a proportion, but without using n), and then multiplied by a multiple of 10 to eliminate decimal points:

$$\text{Rate} = \frac{\text{\# of occurrences of event}}{\text{estimated population}} * 10^k = \frac{f}{Pop} * 10^k$$

Demographers often use 1,000 as the population unit, so that k = 3. For example, the birth rate is calculated as ratio of the total number of births to the total number of women of childbearing age. A birth rate of 150.5 means that there are 150.5 births per every 1,000 women of childbearing age. Economists typically express school expenditure per pupil, so that k = 0 (since $10^0 = 1$). Criminologists typically calculate crime rates per 100,000, so that k = 5. There is a practical reason that crime rates are computed per 100,000 population. Crime is such a rare phenomenon, that if we just calculated the count of crimes to the estimated population, we would find that in 2016 there were 0.001028 robberies per person in the U.S. It would be more informative to say that 102.8 robberies occurred per 100,000 inhabitants in the U.S. in 2016. The following table provides robbery counts and estimated population size for the four census regions in 2016.

When you calculate the rate of robbery per 100,000 in each region, it is evident that raw crime counts have one important weakness. To say that the West had 86301 robberies reported to police compared to 128842 in the South makes the West seem like a safer place to live. However, when we take into account different population sizes by computing a rate, we learn that the West has a higher rate of robbery (112.58) compared to the South (105.33).

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

Proportional (Percent) Change

A comparison of a single variable across two time periods/conditions. This allows us to track changes in some phenomenon over time (or conditions) and indicates the proportional increase or decrease in the magnitude of a variable between two time periods (or two conditions). It is common to multiply the proportional change by 100 to compute a percent change.

$$\text{Percent change} = \frac{\text{Time 2 - Time 1}}{\text{Time 1}} * 100 \text{ or } \frac{\text{Comparison - Baseline}}{\text{Baseline}} * 100$$

Consider an example using changes in crime rates. In 2011, there were 4.76 homicides per 100,000 in the U.S. By 2016, the homicide rate increased to 5.39 per 100,000. This table provides percent change in homicide rates by nine census regions from 2011 to 2016.

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

VI. Frequency Distributions

A frequency distribution is a useful way to summarize data. It is a table that lists all of the categories or scores of a variable and the frequency of each category, at a minimum, and often lists in addition the proportions and percents for each of the categories. A frequency distribution provides visualization of how cases are spread out across categories.

A frequency distribution is a particularly useful may to summarize categorical (nominal, ordinal) data. To do this, you must list all possible categories of the variable of interest. The following table provides data on family living arrangements for a nationally representative sample of 12- to 16-year-old adolescents.

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

Since this is a nominal-level variable, we may order the categories in any way that we see fit. One important thing to note is that the sum of the frequencies will always be equal to the sample size. By way of an example, let's consider the ratio of two-parent to single-parent families. I will use the frequencies to do so: (3350 + 1002) / (1972 + 233) = 4352 / 2205 = 1.97. Thus, there is a 2-to-1 ratio of two-parent to single-parent households.

It can also be informative to compare the frequency distribution of a single variable across two or more subgroups of a sample. The next table compares family living arrangements among white and black youths.

	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

We can use the information provided in this table to ask important questions about family structure. For example, we can use a ratio of proportions to determine that white youths are 2.26 (.594/.263) times more likely to live with both biological parents as black youths. That is, for every black youth that lives with both biological parents, there are 2.26 white youths that live with both biological parents. On the other hand, black youths are 2.57 (.480/.187) times more likely than white youths to live with only their biological mother, and 3.52 (.102/.029) times more likely to live with some other family member.

With ordinal data (or higher) in a frequency distribution, we must rank the categories according to their magnitudes. Also, since we have the property of rank order, we may add additional information to the frequency distribution. We refer to these as cumulative statistics, and we can compute a cumulative frequency (cf), proportion (cp), or percent (c%). The cumulative frequency of a category (or a particular score, in the case of interval-ratio data), for example, provides the number of observations that are in that category or in a lower category.

This table provides a frequency distribution of 8th grade scholastic performance.

The cumulative frequency for the first category is the frequency of that category. Then we simply add the next frequency to this value as we move to the second category. The cumulative frequency of the highest

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			

category must be equal to the sample size. The same goes for cumulative proportions and percents. For example, we can determine that half of the sample (50.8%) earned mostly B's or better during the 8th grade.

The following table provides scholastic performance by gender.

	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	

This table shows that females are 1.79 (19.1/10.7) times more likely to make mostly A's than males. If we consider mostly A's and a mixture of A's and B's, we can see that females are 1.55 ([19.1+26.1] / [10.7+18.5]) times more likely to make these grades than males. Conversely, males are 2.12 (1.7/0.8) times more likely to make below D's, and 1.92 ([1.7+3.3] / [0.8+1.8]) times more likely to make mostly D's or below.

With discrete interval-ratio data, we have a choice of constructing a simple "ungrouped" or a "grouped" frequency distribution. Thus far, we have been constructing ungrouped frequency distributions, in which all possible values or categories are listed, and their respective frequencies, etc. An ungrouped frequency distribution with interval-ratio data is useful when there are only a handful of different values that a variable may take on.

The next table provides information on adolescent time use broken out by self-report delinquent status.

	Non-delinquents		Delinquents	
# of Weekdays Do Homework?	%	c%	%	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	

The values that this variable may assume are integers between zero and five, to represent the number of weekdays that a youth engages in a particular activity. In general, it appears that delinquents spend less time doing homework and reading for pleasure than non-delinquents. You can see that non-delinquents are 1.34 (38.8/29.0) times more likely to do homework every weekday compared to delinquents, and 1.29 (19.8/15.3) times more likely to ready a book every weekday. There seem to be few differences in the distributions of watching TV across the two groups.

Discrete interval-ratio data that include a large number of values, or continuous interval-ratio data, are not as easily summarized in a simple (ungrouped) frequency distribution. Consider the following raw data on number of months sentenced for robbery (n = 40).

- 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

If I were to construct a simple frequency distribution, it would look something like this.

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

The problem is that this distribution is relatively flat and does not provide very much information. In other words, there is not much repetition of values in the data. In these instances, it is convenient to group the data into "classes" or intervals, and then to construct a frequency distribution using these classes. In essence, we are transforming quantitative data into qualitative data (specifically, ordinal data).

- 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
- 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

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

Let's work through an example with continuous interval-ratio data. The following data provides state-level unemployment data for 1998.

- 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

Interval Limits	f	p	%	cf	cp	c%
2.5 - 2.9	2	0.04	4.0	2	0.04	4.0
3.0 - 3.4	6	0.12	12.0	8	0.16	16.0
3.5 - 3.9	5	0.10	10.0	13	0.26	26.0
4.0 - 4.4	8	0.16	16.0	21	0.42	42.0
4.5 - 4.9	10	0.10	10.0	31	0.62	62.0
5.0 - 5.4	11	0.16	16.0	42	0.84	84.0
5.5 - 5.9	3	0.06	6.0	45	0.90	90.0
6.0 - 6.4	3	0.06	6.0	48	0.96	96.0
6.5 - 6.9	2	0.04	4.0	50	1.00	100.0

VII. Data Visulaization Techniques

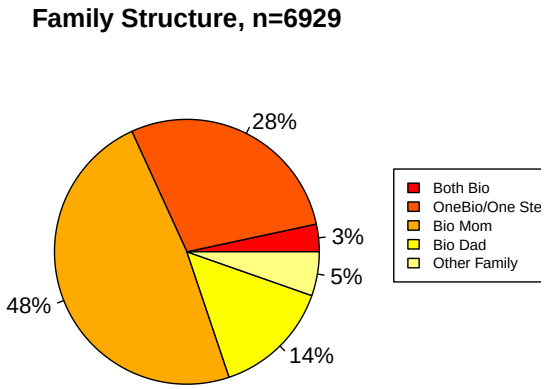
Qualitative Data (Nominal/Ordinal)

If you have qualitative (nominal, ordinal) data, there are two common ways to display them.

Pie Chart

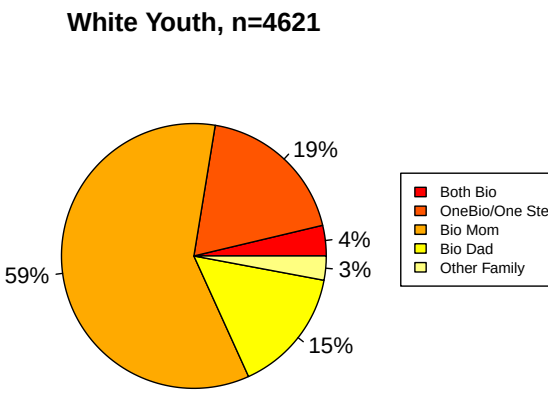
Each category receives a "slice" of the pie, where a slice represents the percent of the total sample. It is best if there are only five or fewer categories when displaying data with a pie chart; this makes interpretation much more clear. It is also necessary to label the categories and the percent of the sample represented in each category.

Family Structure - Pie Chart



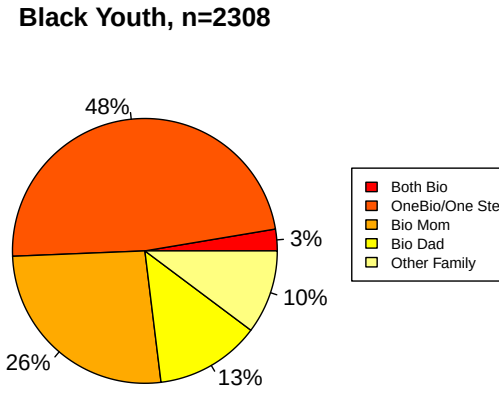
Family Structure by Race - Pie Charts

White Youth



Family Structure by Race - Pie Charts

Black Youth

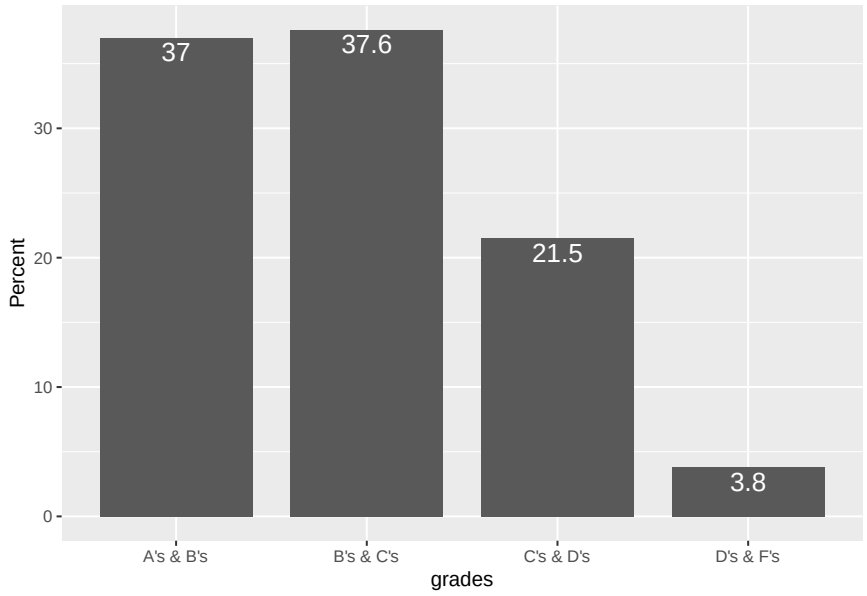


Bar graph

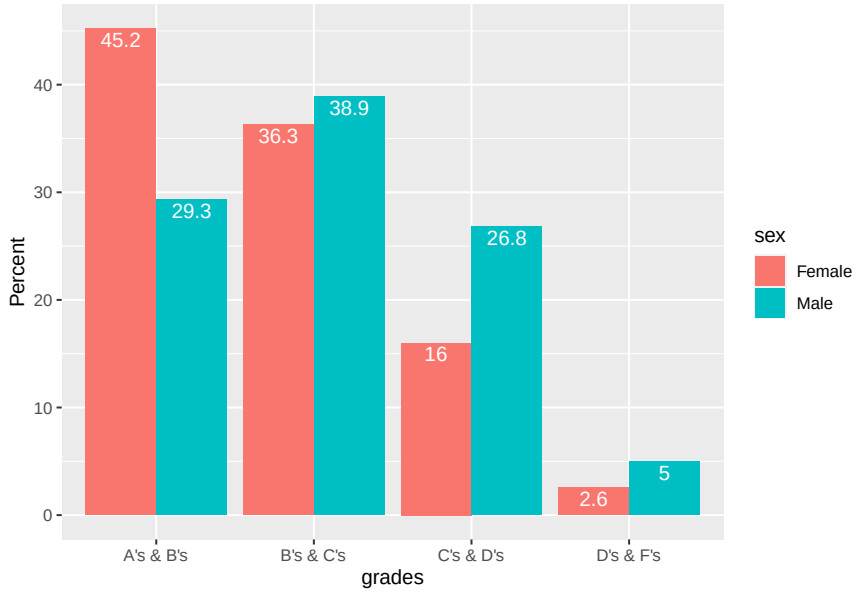
Each category gets a bar. The y-axis can represent frequencies, proportions, or percents. It is important that there be spaces between each of the bars, indicating that the categories are distinct. If not easily discerned, it is also useful to place the frequency (or proportion or percent) above the bar.

With quantitative (interval, ratio) data, there are three common forms of display. Note that grouped data, which are technically ordinal, are treated as quantitative for the purpose of graphing.

Scholastic Performance Bar Graph



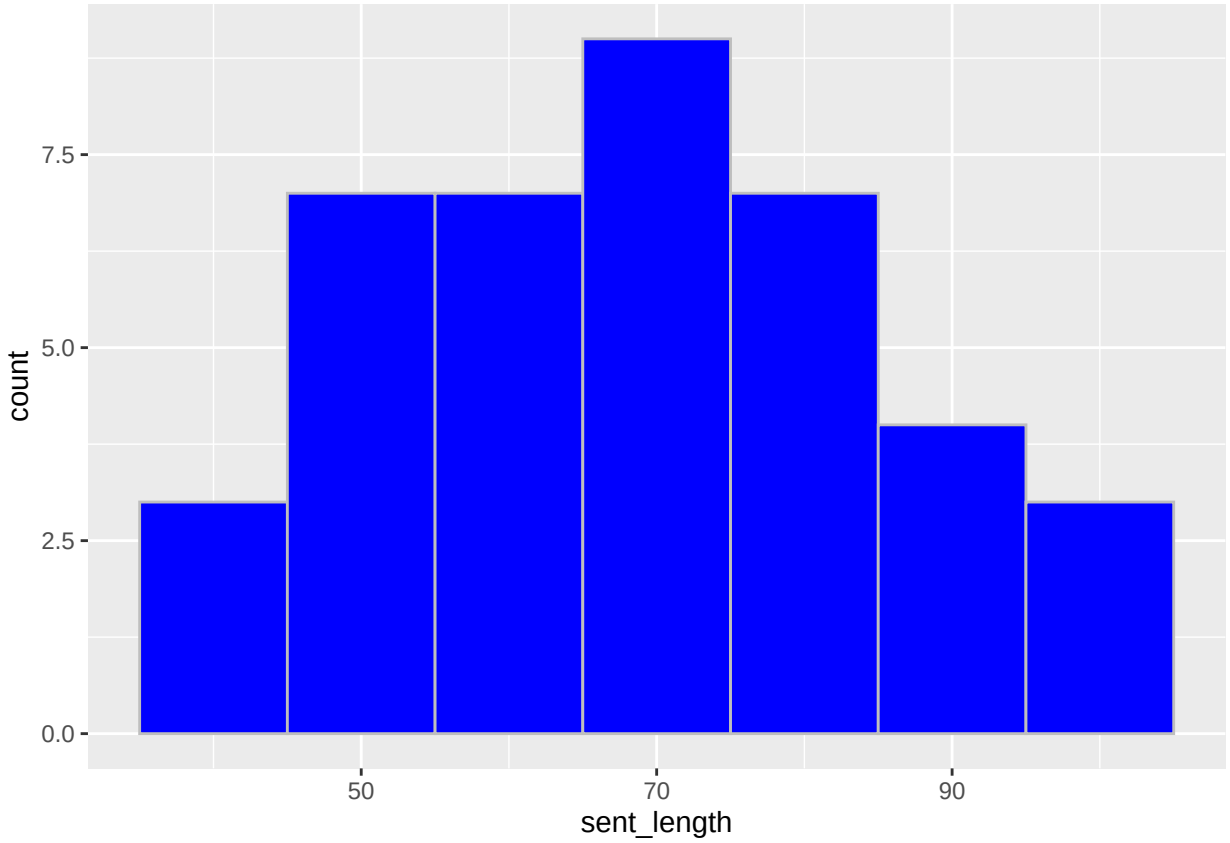
Scholastic Performance Bar Graph



Histogram

A histogram is very similar to a bar graph; the primary difference is that there are no spaces between bars. This is to indicate that the original data are quantitative in nature, rather than qualitative. The y-axis represents the frequency (or proportion or percent) of each interval, and the x-axis represents the interval limits (or for the sake of convenience, the interval number). A useful property of a histogram is that it gives a visual indication of the amount of skew in the data, or the degree to which the data depart from a normal distribution.

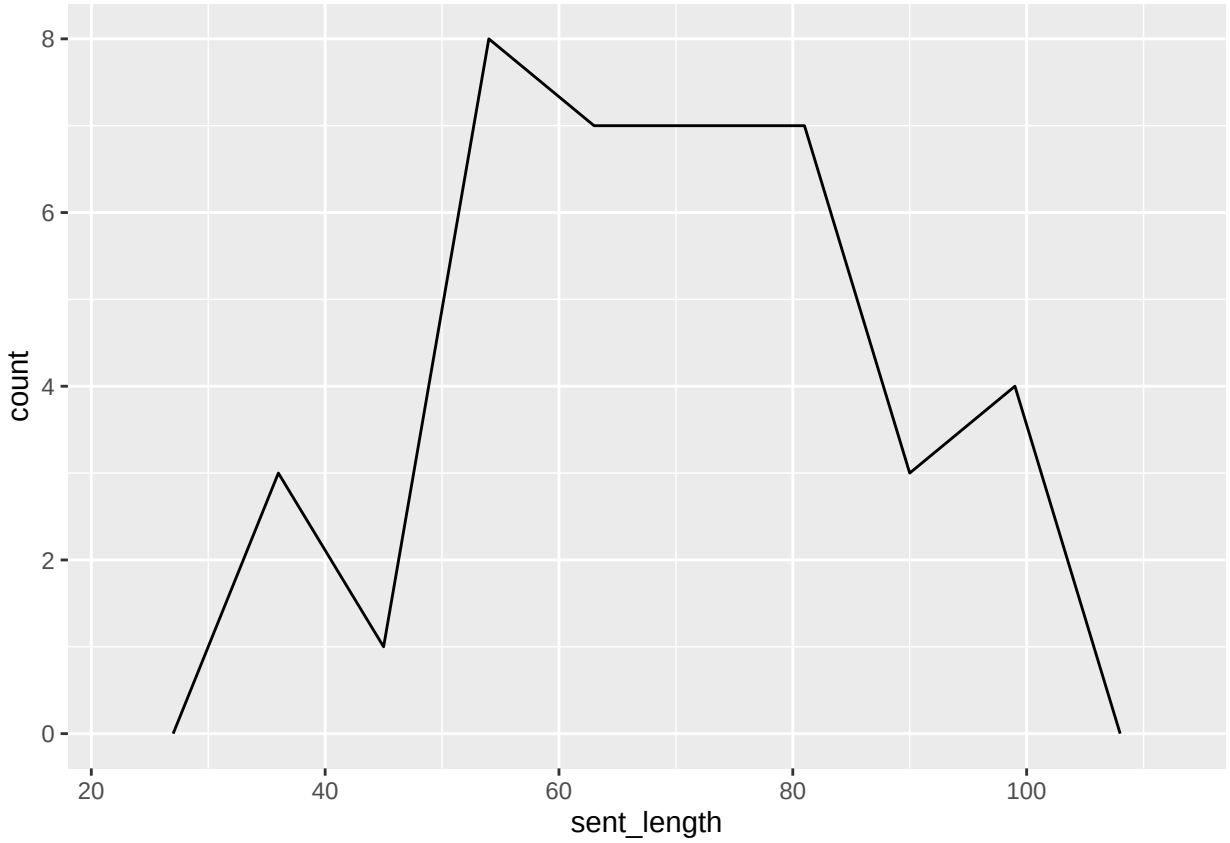
Robbery Sentence Length Histogram



Polygon

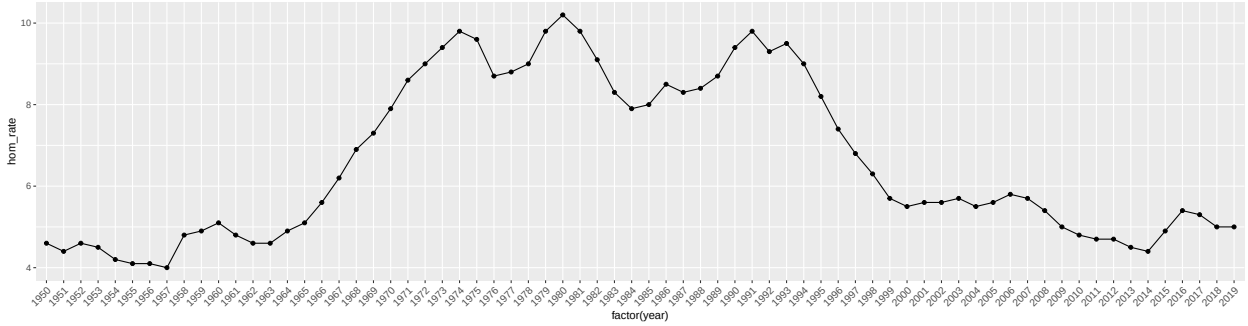
This is a line chart that is constructed much like a histogram, but instead of using bars to represent frequencies (or proportions or percents) a dot is placed at the midpoint of each interval, and a line to connect the dots together. The line should be connected to the ¬x-¬axis (since it is referred to as a "polygon").

Robbery Sentence Length Polygon

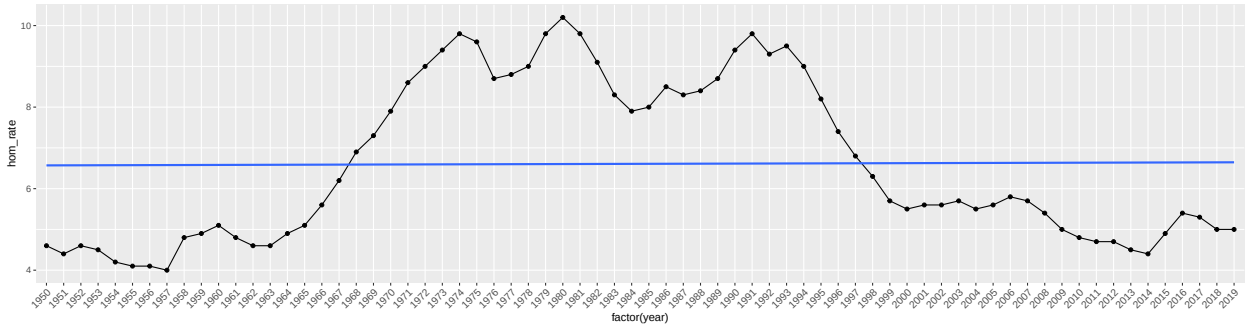


Time series

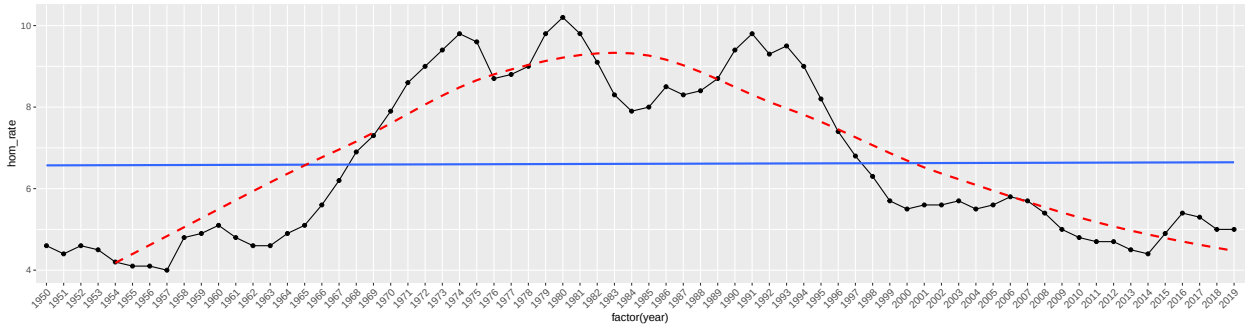
Data that are time series use a line graph. Consider the following data on U.S. homicide rates from 1950 to 2019 (n = 70).



Homicide Rates, 1950 - 2019 Time Series with Trend Line



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

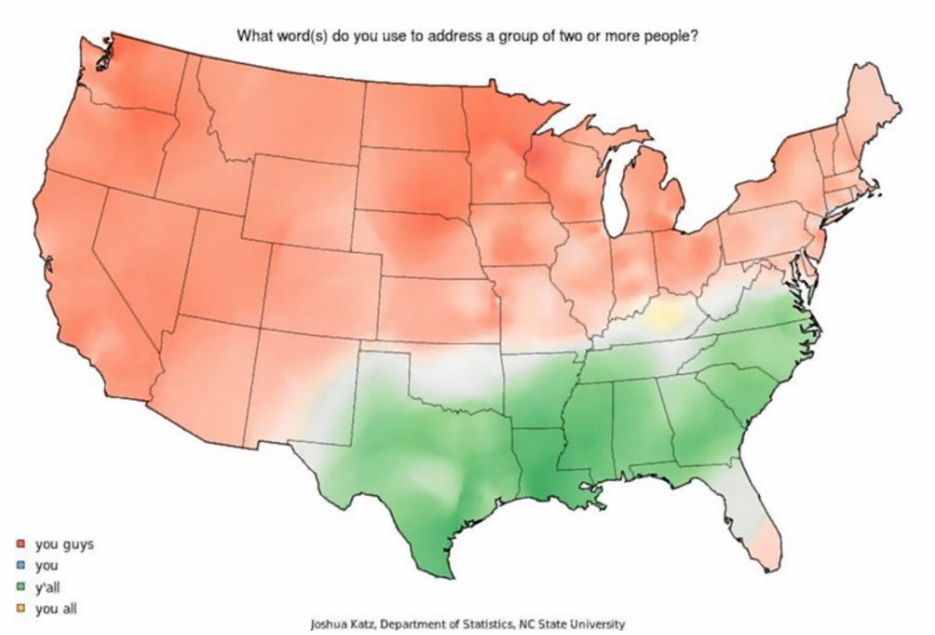


Here the x-axis represents time, measured in years. Time series plots are useful to look for trends in some phenomenon, either over the course of the entire time period under consideration, or during specific periods.

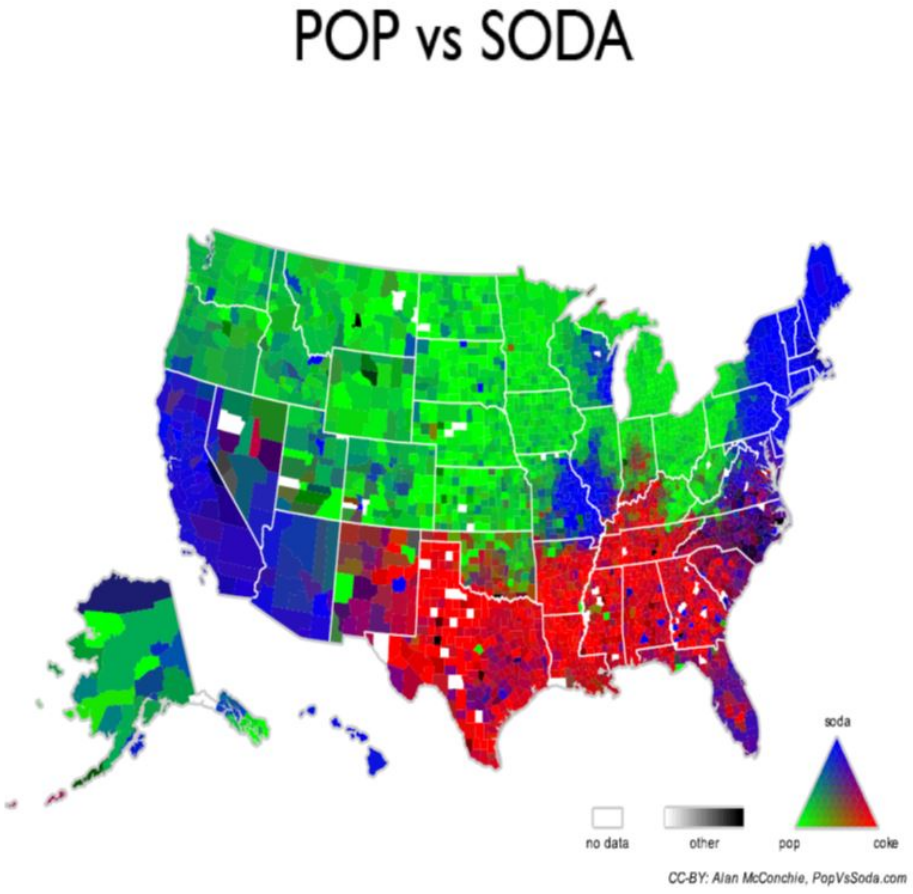
Map

This method is particularly useful for displaying census-related information. Whereas time series plots are useful for showing temporal trends, maps are useful for showing spatial trends.

Y'All Use Y'All?



Pop Versus Soda



You Can Even Make a Map in R!