STA291
Fall 2009

LECTURE 5
10 SEPTEMBER 2009

Itinerary

- Graphical Techniques for Interval Data (mostly review)
- Describing the Relationship Between Two Variables
- Art and Science of Graphical Presentations

Review: Graphical/Tabular Descriptive Statistics

- Summarize data
- Condense the information from the dataset
- Always useful: Frequency distribution
- Interval data: Histogram (Stem-and-Leaf?)
- Nominal/Ordinal data: Bar chart, Pie chart

Data Table: Murder Rates

Alabama	11.6	Alaska	9
Arizona	8.6	Arkansas	10.2
California	13.1	Colorado	5.8
Connecticut	6.3	Delaware	5
D.C.	78.5	Florida	8.9
Georgia	11.4	Hawaii	3.8
…		…	

- Difficult to see the "big picture" from these numbers
- Try to condense the data…

Frequency Distribution

- A listing of intervals of possible values for a variable
- And a tabulation of the number of observations in each interval.

Murder Rate	Frequency
0 – 2.9	5
3 – 5.9	16
6 – 8.9	12
9 – 11.9	12
12 – 14.9	4
15 – 17.9	0
18 – 20.9	1
> 21	1
Total	51

Frequency Distribution

- Use intervals of same length (wherever possible)
- Intervals must be mutually exclusive: Any observation must fall into one and only one interval
- Rule of thumb:

  If you have *n observations*, *the number of* intervals should be about $\sqrt{n}$

Frequency, Relative Frequency, and Percentage Distribution

Murder Rate	Frequency	Relative Frequency	Percentage
0 – 2.9	5	.10 (= 5 / 51)	10 (= .10 * 100%)
3 – 5.9	16	.31 (= 16 / 51)	31 (= .31 * 100%)
6 – 8.9	12	.24	24
9 – 11.9	12	.24	24
12 – 14.9	4	.08	8
15 – 17.9	0	0	0
18 – 20.9	1	.02	2
> 21	1	.02	2
Total	51	1	100

Frequency Distributions

- Notice that we had to group the observations into intervals because the variable is measured on a continuous scale

- For discrete data, grouping may not be necessary (except when there are many categories)

Frequency and Cumulative Frequency

- Class Cumulative Frequency: Number of observations that fall in the class and in smaller classes

- Class Relative Cumulative Frequency: Proportion of observations that fall in the class and in smaller classes

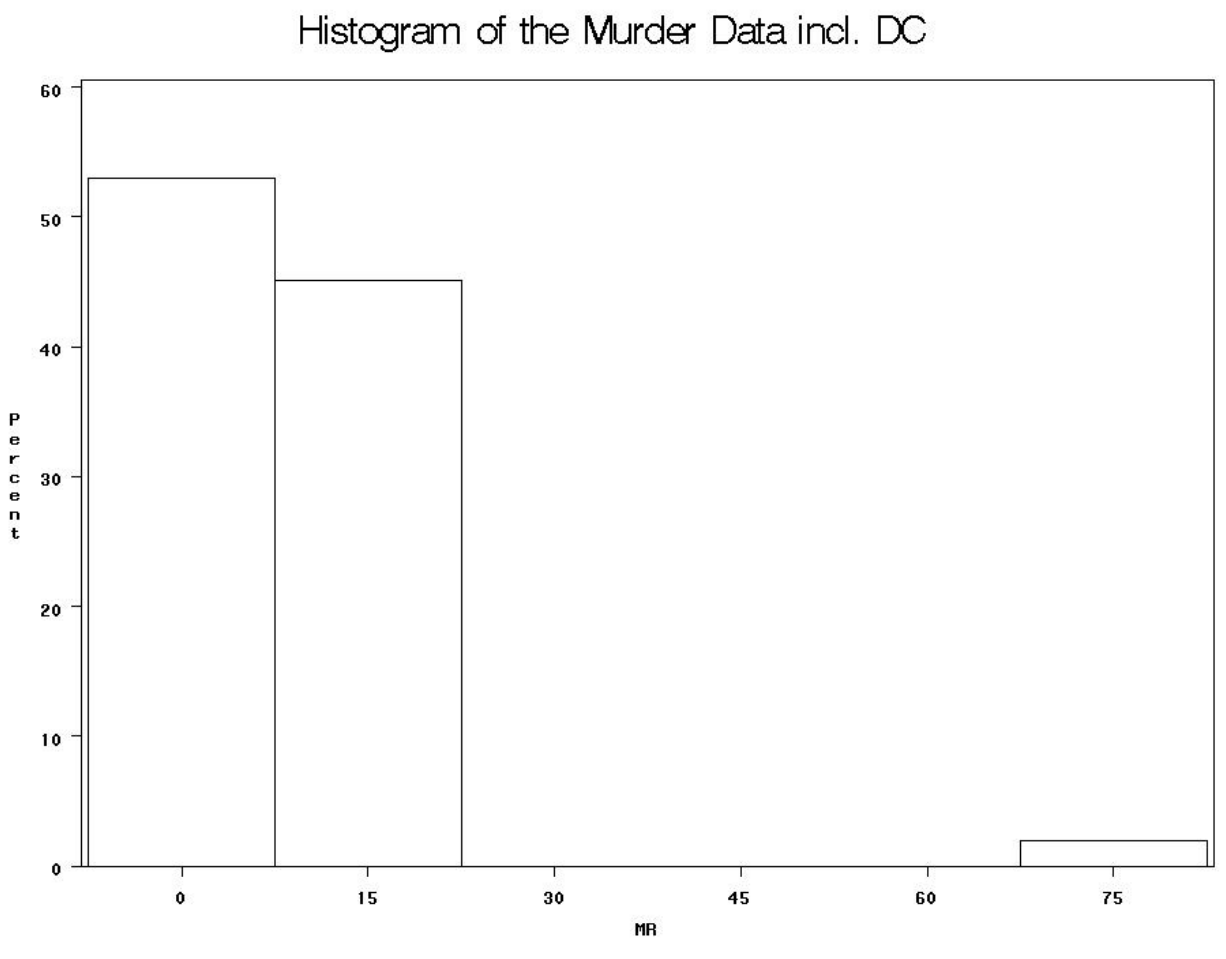
Cumulative Frequencies & Relative Frequencies

Murder Rate	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0 – 2.9	5	.10	5	.10
3 – 5.9	16	.31	21 (= 16 + 5)	.41 (=.31 +.10)
6 – 8.9	12	.24	33 (= 12 + 21)	.65(=.24+.41)
9 – 11.9	12	.24		
12 – 14.9	4	.08		
15 – 17.9	0	0		
18 – 20.9	1	.02		
> 21	1	.02		
Total	51	1		

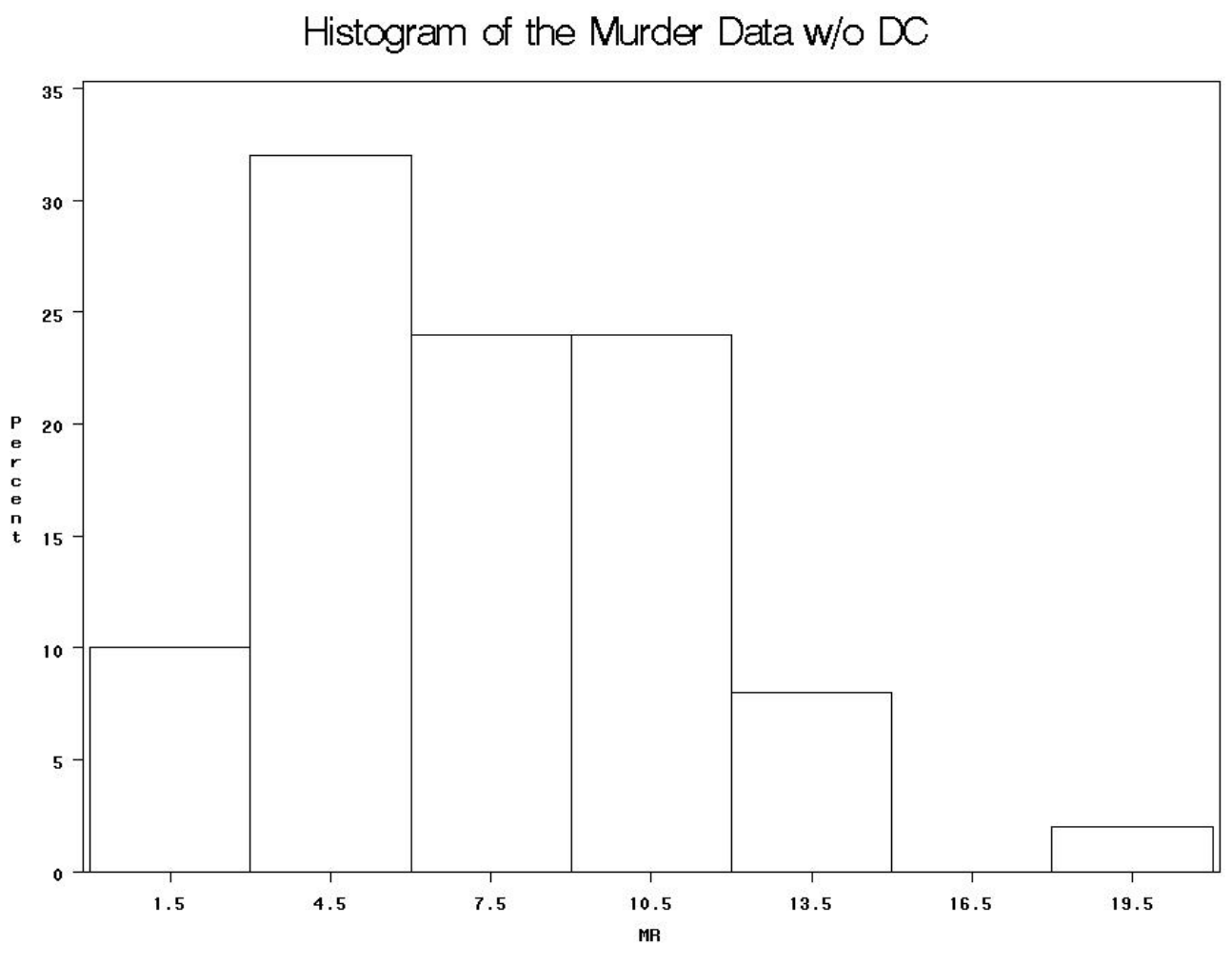
Histogram (Interval Data)

- Use the numbers from the frequency distribution to create a graph
- Draw a bar over each interval, the height of the bar represents the relative frequency for that interval
- Bars should be touching; i.e., equally extend the width of the bar at the upper and lower limits so that the bars are touching.

Histogram (version I)



Histogram (version II)



Bar Graph (Nominal/Ordinal Data)

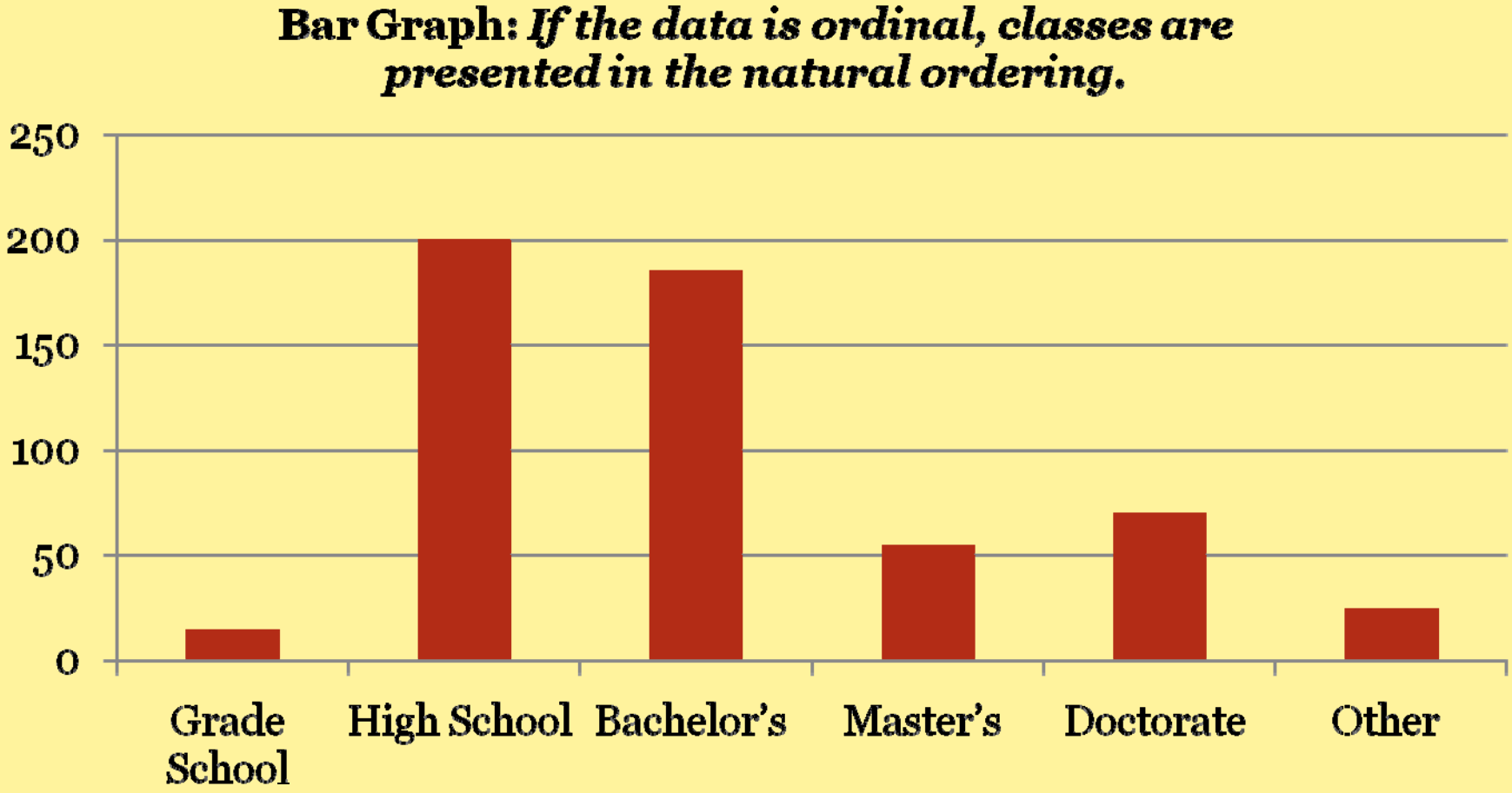
- Histogram: for *interval* (quantitative) data
- Bar graph is almost the same, but for *qualitative data*
- Difference:
  - The bars are ***usually separated*** to emphasize that the variable is categorical rather than quantitative
  - For nominal variables (no natural ordering), order the bars by frequency, except possibly for a category "other" that is always last

Pie Chart (Nominal/Ordinal Data)

- First Step: Create a Frequency Distribution

Highest Degree	Frequency (Number of Responses)	Relative Frequency
Grade School	15	
High School	200	
Bachelor's	185	
Master's	55	
Doctorate	70	
Other	25	
Total	550	

We could display this data in a bar chart…

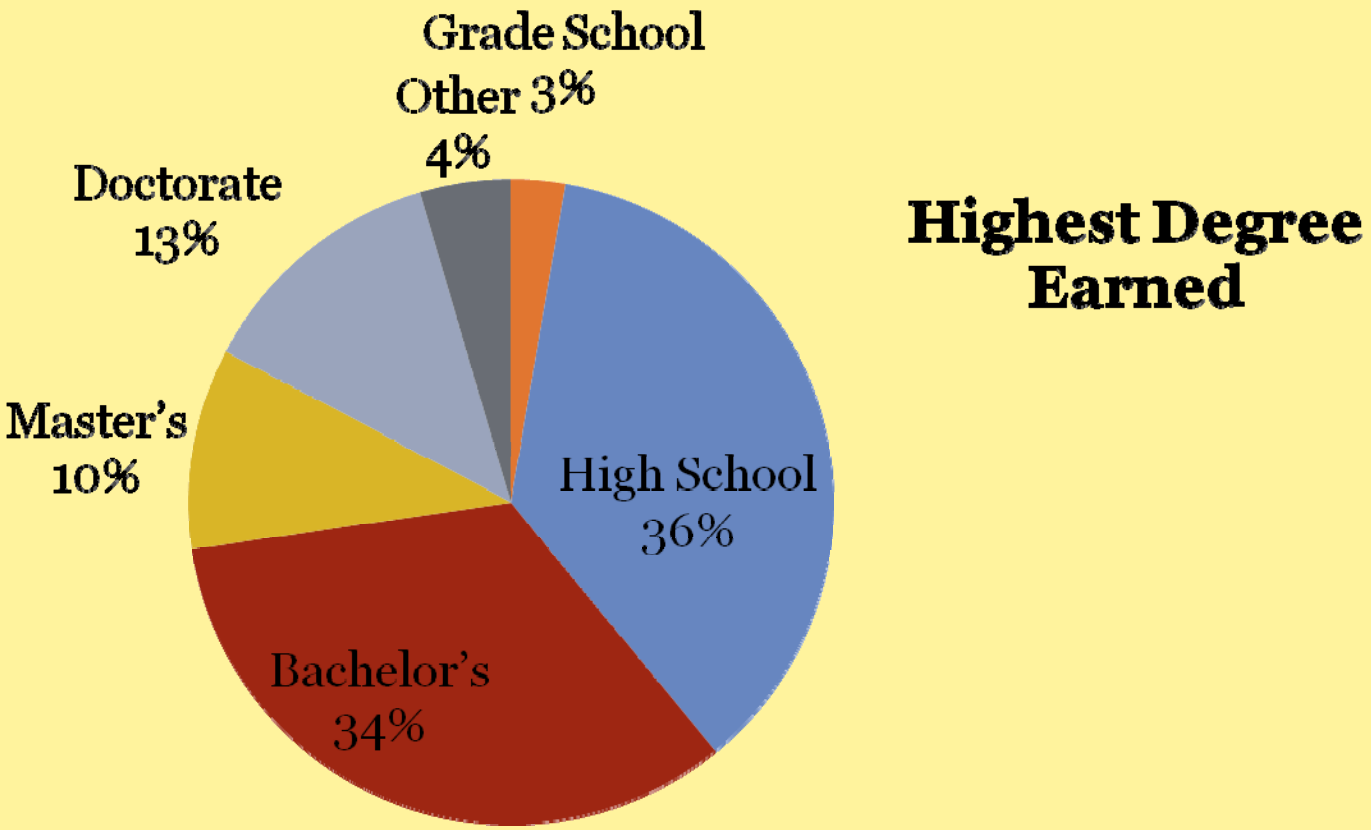


Pie Chart

- Pie Chart**: Pie is divided into slices; The area of each** slice is proportional to the frequency of each class.

Highest Degree	Relative Frequency	Angle (= Rel. Freq. * 360°)
Grade School	.027 (= 15/550)	9.72 (= .027 * 360°)
High School	.364	131.04
Bachelor's	.336	120.96
Master's	.100	36.0
Doctorate	.127	45.72
Other	.045	16.2

Pie Chart


Grade School
Other 3%
4%
Doctorate
13%
Master's
10%
High School
36%
Bachelor's
34%
Highest Degree Earned

Stem and Leaf Plot

- Write the observations ordered from smallest to largest
- Each observation is represented by a stem (leading digit(s)) and a leaf (final digit)
- Looks like a histogram sideways
- Contains more information than a histogram, because every single measurement can be recovered

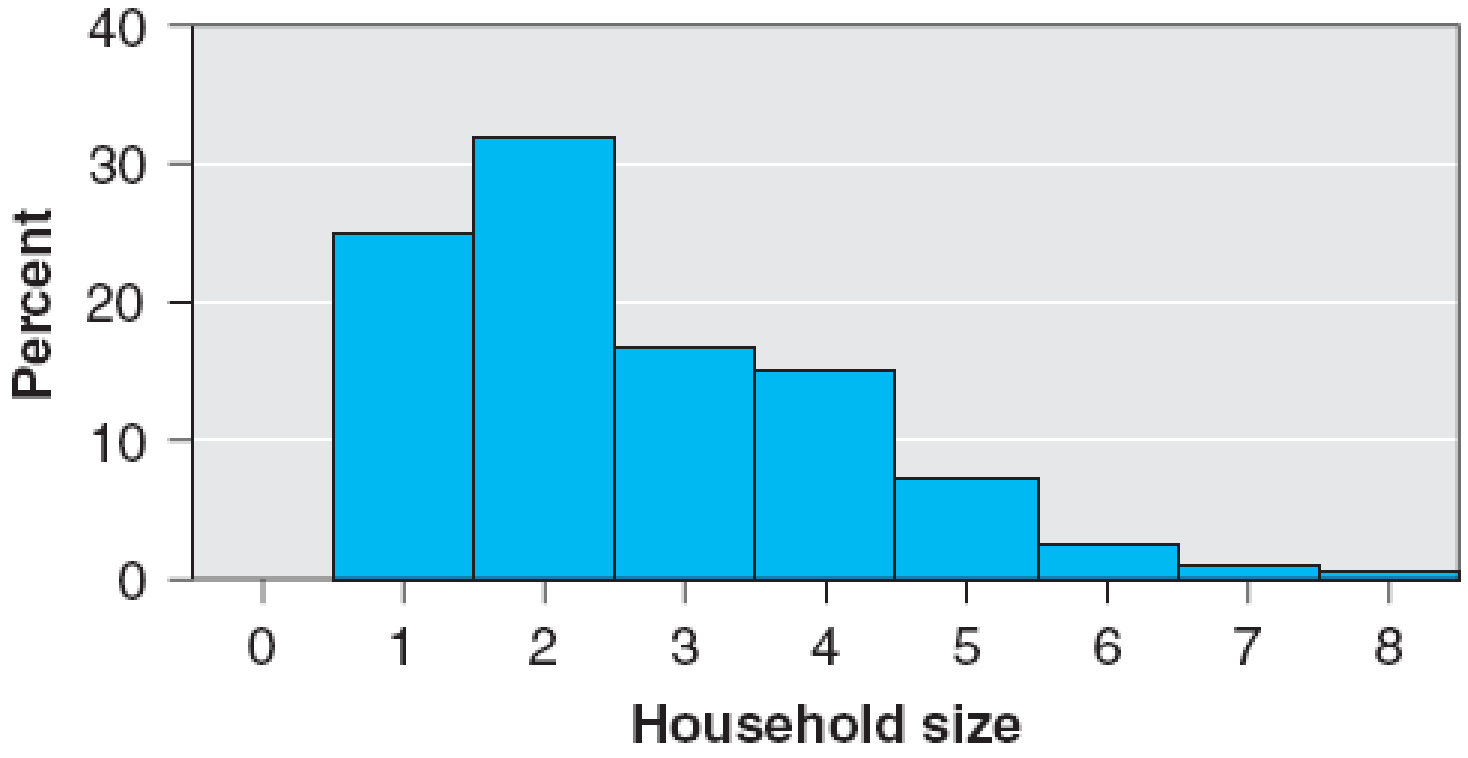
Stem and Leaf Plot

- Useful for small data sets (<100 observations)
  - Example of an *EDA*
- Practical problem:
  - What if the variable is measured on a continuous scale, with measurements like 1267.298, 1987.208, 2098.089, 1199.082, 1328.208, 1299.365, 1480.731, etc.
  - Use common sense when choosing "stem" and "leaf"

Stem-and-Leaf Example: Age at Death for Presidents

PRESIDENT	AGE	PRESIDENT	AGE	PRESIDENT	AGE
Washington	67	Fillmore	74	Roosevelt	60
Adams	90	Pierce	64	Taft	72
Jefferson	83	Buchanan	77	Wilson	67
Madison	85	Lincoln	56	Harding	57
Monroe	73	Johnson	66	Coolidge	60
Adams	80	Grant	63	Hoover	90
Jackson	78	Hayes	70	Roosevelt	63
Van Buren	79	Garfield	49	Truman	88
Harrison	68	Arthur	56	Eisenhower	78
Tyler	71	Cleveland	71	Kennedy	46
Polk	53	Harrison	67	Johnson	64
Taylor	65	McKinley	58	Nixon	81
				Reagan	93

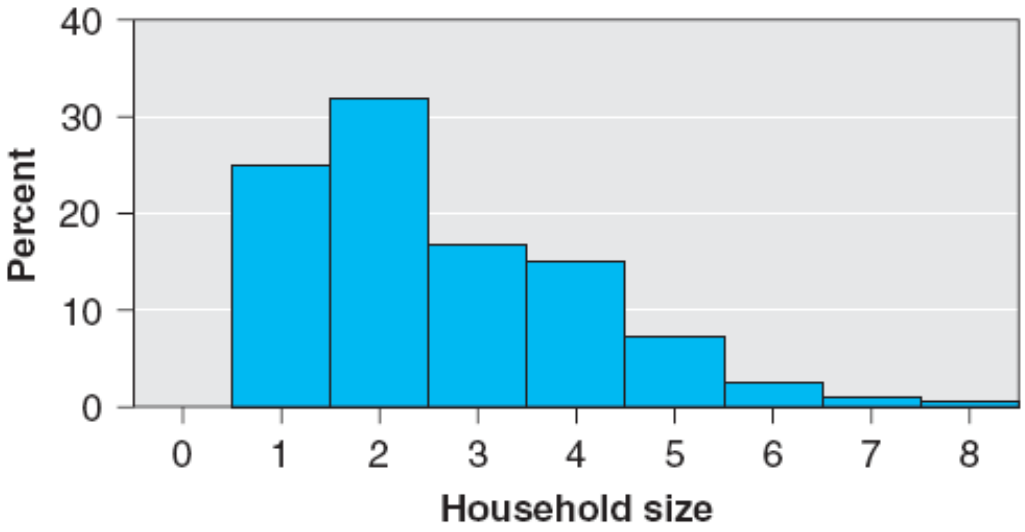
Example (Percentage) Histogram



Side by side?

Annual Earnings (Thousands of Dollars)

Stem	Leaf
1	6,7
2	9,9,9,2,7,7,8,3,5
3	4,5,1,2,6,6,3,9
4	6,9,7,7,7,5,9,8,6,3,0,3,4,8,3
5	3,5,3,7,7,9
6	3,2,2,4,4,0,0
7	7,4



Similarities/differences?

Sample/Population Distribution

- Frequency distributions and histograms exist for the population as well as for the sample
- Population distribution vs. sample distribution
- As the sample size increases, the sample distribution looks more and more like the population distribution

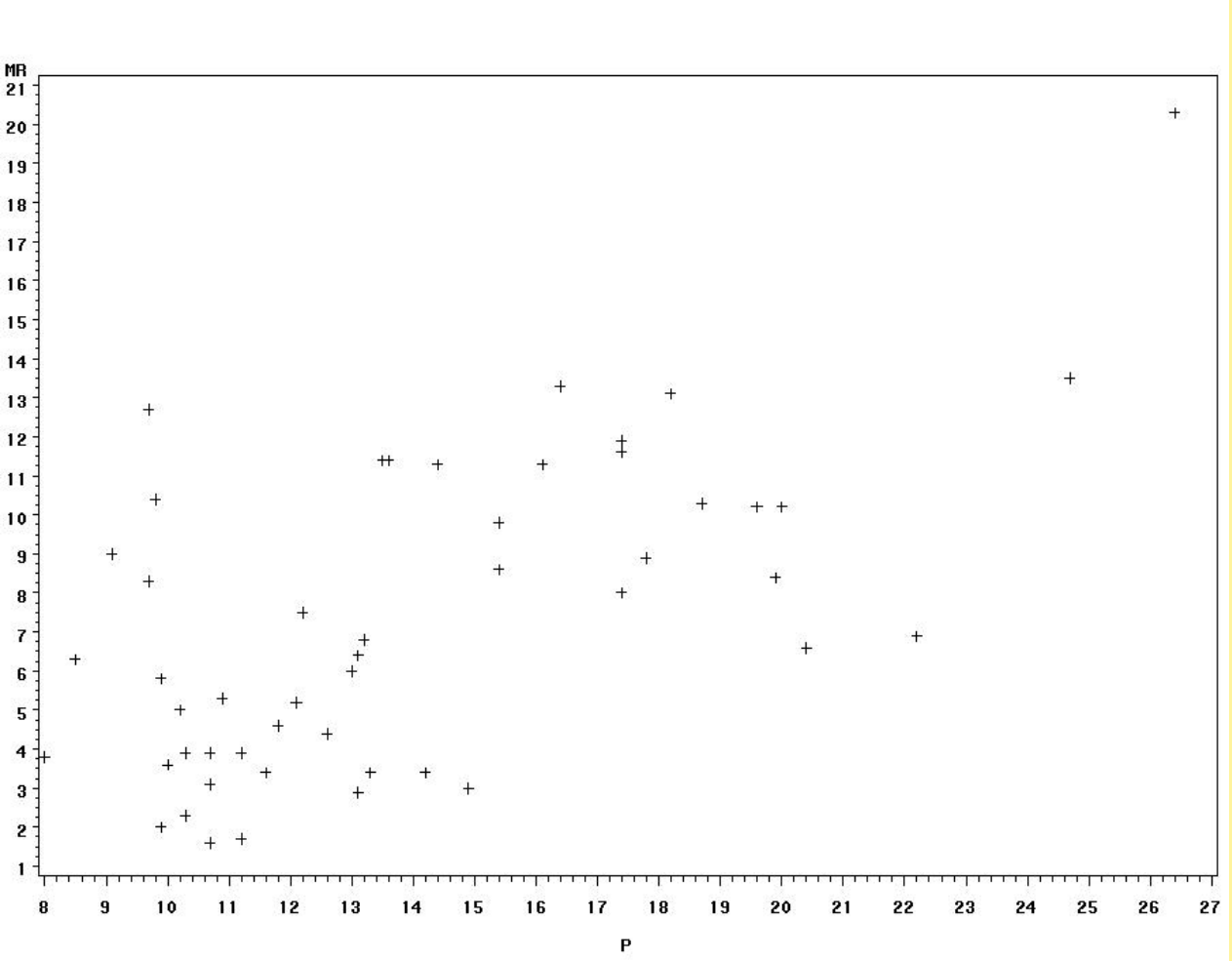
Describing Distributions

- Center, spread (numbers later)

- Symmetric distributions
  - Bell-shaped or U-shaped

- Not symmetric distributions:
  - Left-skewed or right-skewed

On to examining two variables for relationships . . .



Describing the Relationship Between Two Nominal (or Ordinal) Variables

Contingency Table

- Number of subjects observed at all the combinations of possible outcomes for the two variables
- Contingency tables are identified by their number of rows and columns
- A table with 2 rows and 3 columns is called a 2 x 3 table (“2 by 3”)

2 x 2 Contingency Table: Example

- 327 commercial motor vehicle drivers who had accidents in Kentucky from 1998 to 2002
- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

2 x 2 Contingency Table: Example, cont'd.

- How can we compare fatality rates for the two groups?
- Relative frequencies or percentages within each row
- Two sets of relative frequencies (for *seatbelt=yes* and for *seatbelt=no)*, called **row relative frequencies**
- If seat belt use and fatality of accident are related, then there will be differences in the row relative frequencies

Row relative frequencies

- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes			100
	No			100
				100

Describing the Relationship Between Two Interval Variables

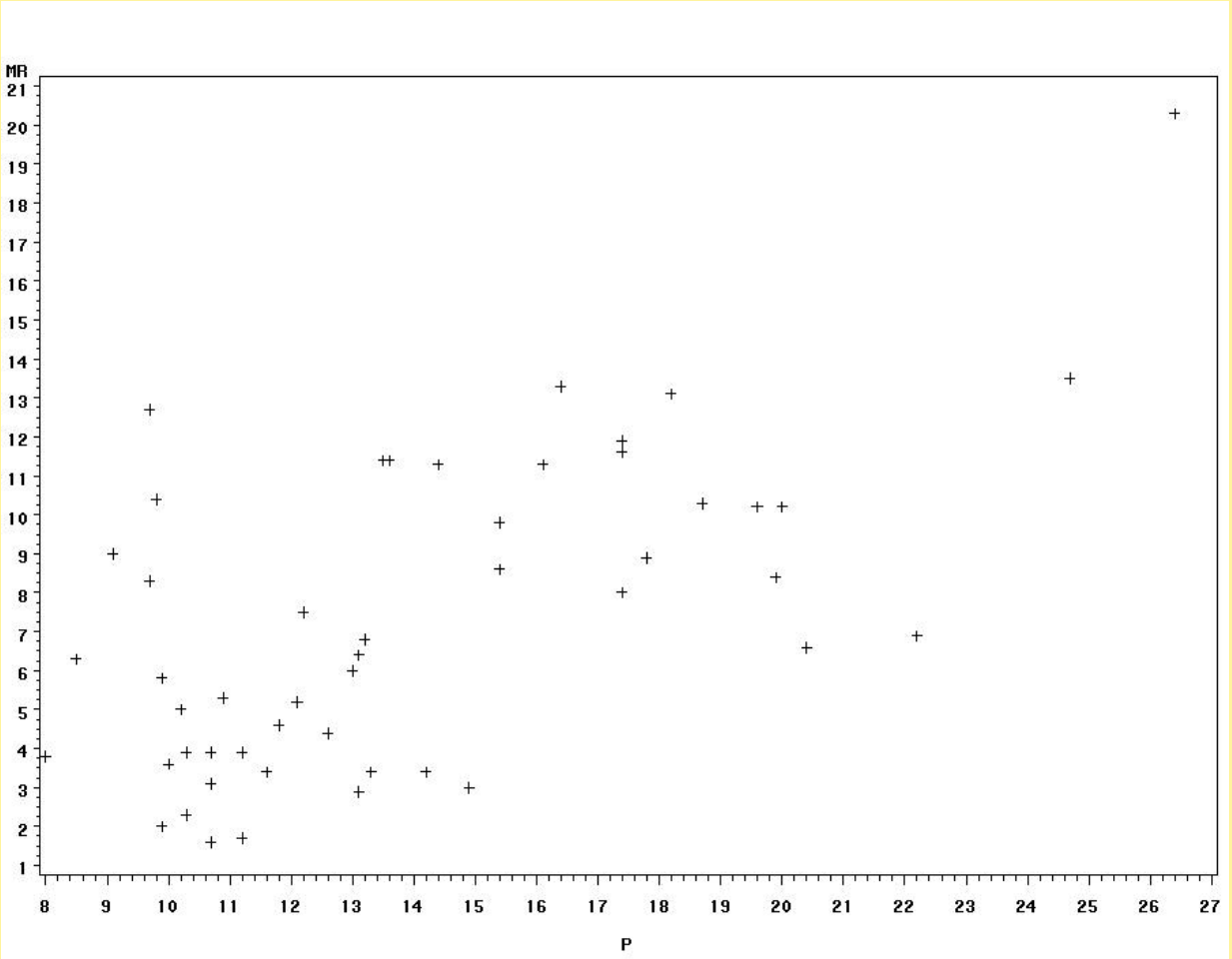
Scatter Diagram

- In applications where one variable depends to some degree on the other variables, we label the dependent variable Y and the independent variable X
- Example:

  Years of education = X

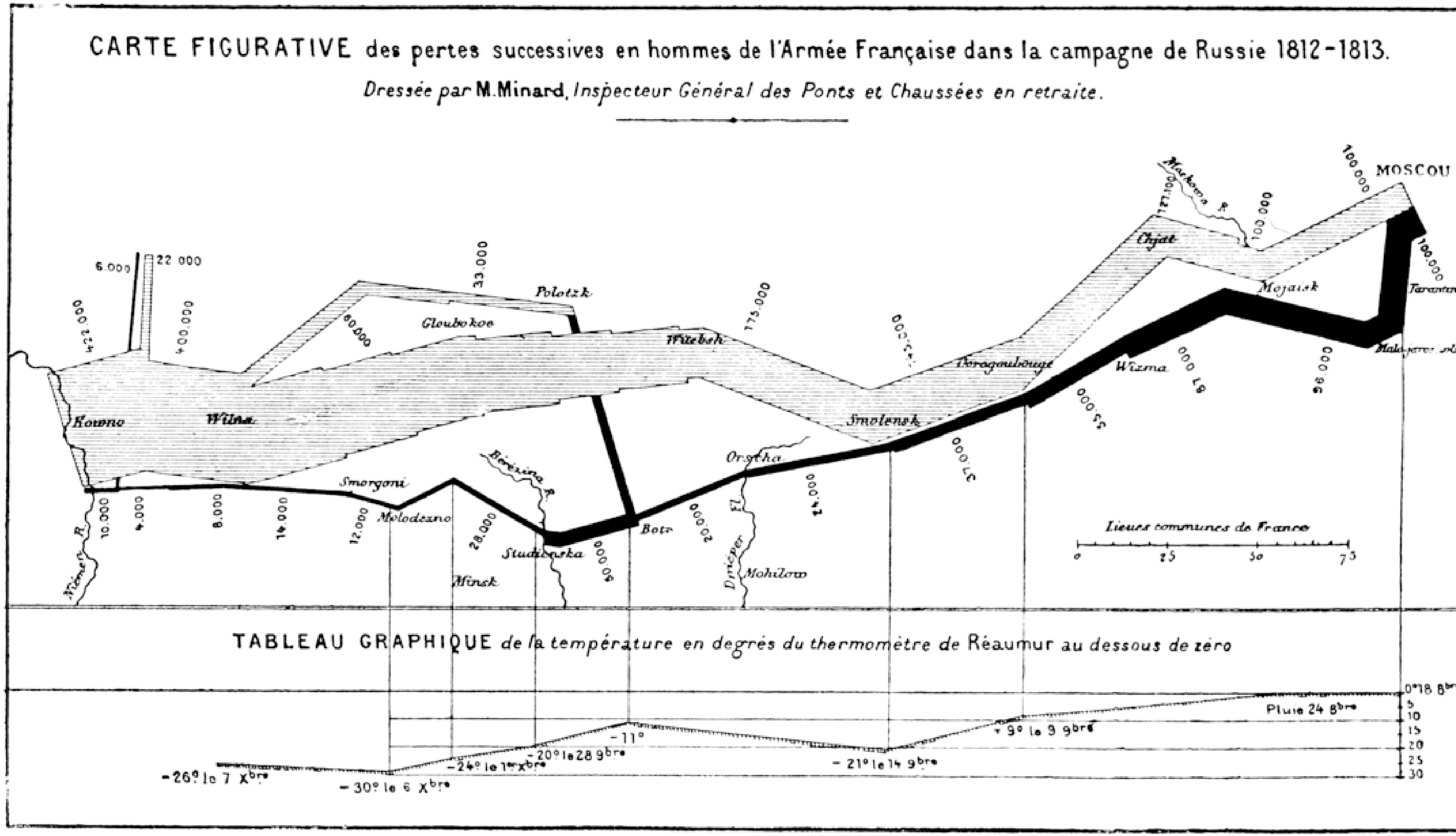
  Income = Y
- Each point in the scatter diagram corresponds to one observation

Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



3.1 Good Graphics …

- … present large data sets concisely and coherently
- … can replace a thousand words and still be clearly understood and comprehended
- … encourage the viewer to compare two or more variables
- … do not replace substance by form
- … do not distort what the data reveal
- … have a high “data-to-ink” ratio

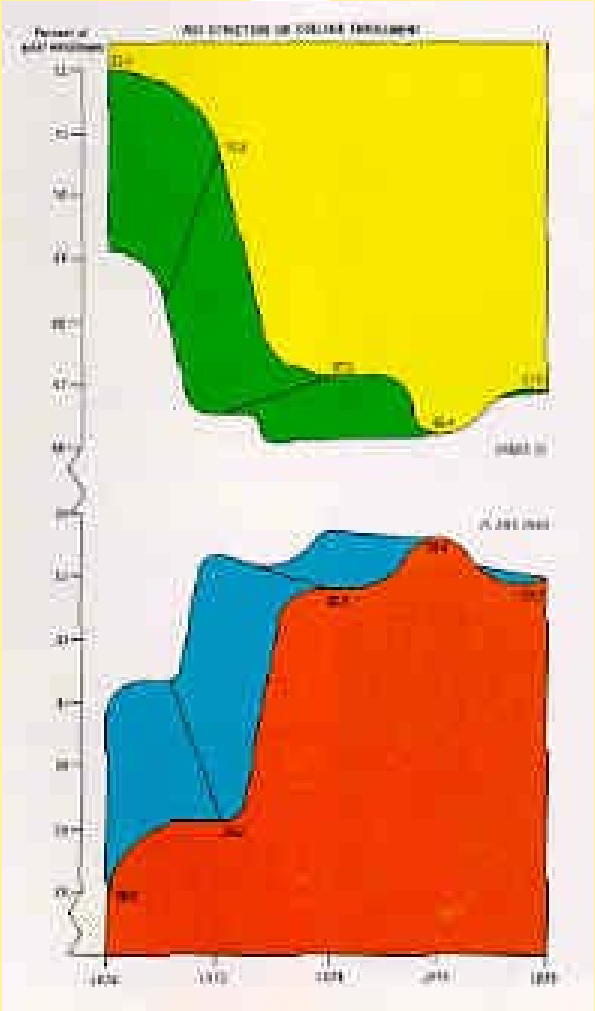


CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
MOSCOU
Kowno
Wilna
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wizma
Orscha
Bobr
Studienska
Minsk
Molodezno
Smorgoni
Mohilow
Lieues communes de France
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro
Pluie 24 8bre
0° 18 8bre
-9° le 9 9bre
-21° le 14 9bre
-11°
-20° le 28 9bre
-24° le 1er Xbre
-30° le 6 Xbre
-26° le 7 Xbre

3.2 Bad Graphics…

- …don't have a scale on the axis
- …have a misleading caption
- …distort by stretching/shrinking the vertical or horizontal axis
- …use histograms or bar charts with bars of unequal width
- …are more confusing than helpful

Bad Graphic, Example



Attendance Survey Question #5

- On an index card
  - Please write down your name and section number
  - Today's Question: