

STA 291

Fall 2009

1

LECTURE 4
8 September 2009

Administrative

2

- (Review) Chapter 3: Sampling Plans
- Chapter 4: Graphical and Tabular Techniques
- Suggested problems from the textbook (not graded, but good as exam preparation): 3.1, 3.2, 3.3, 3.4, 3.5 , 3.7

Where we left off: Types of Bias

3

- **Selection Bias**

- Selection of the sample systematically excludes some part of the population of interest

- **Measurement/Response Bias**

- Method of observation tends to produce values that systematically differ from the true value

- **Nonresponse Bias**

- Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

Where we left off: Sampling Error

4

- Assume you take a random sample of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another random sample of 100 UK students
- Will you get the same percentages?

Sampling Error (cont'd)

5

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you
 1. take a “sample” consisting of the whole population of 30,000 students (this would be called a ?)
- or -
 2. you get very lucky

Sampling Error

6

- **Sampling Error** is the error that occurs when a statistic based on a sample *estimates* or *predicts* the value of a population parameter.
- In random samples, the sampling error can usually be quantified.
- In nonrandom samples, there is also sampling variability, but its extent is not predictable.

Nonsampling Error

7

- Any error that *could* also happen in a census, that is, when you ask the whole population
- Examples: Bias due to question wording, question order, nonresponse (people refuse to answer), wrong answers (especially to delicate questions)

Descriptive Statistics

8

- Summarize data
- Use graphs, tables (and numbers, see Chapter 4)
- Condense the information from the dataset
- Interval data: Histogram
- Nominal/Ordinal data: Bar chart, Pie chart

Data Table: Murder Rates

9

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

10

- 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

11

- 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

12

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

13

- 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

14

- **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

15

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		

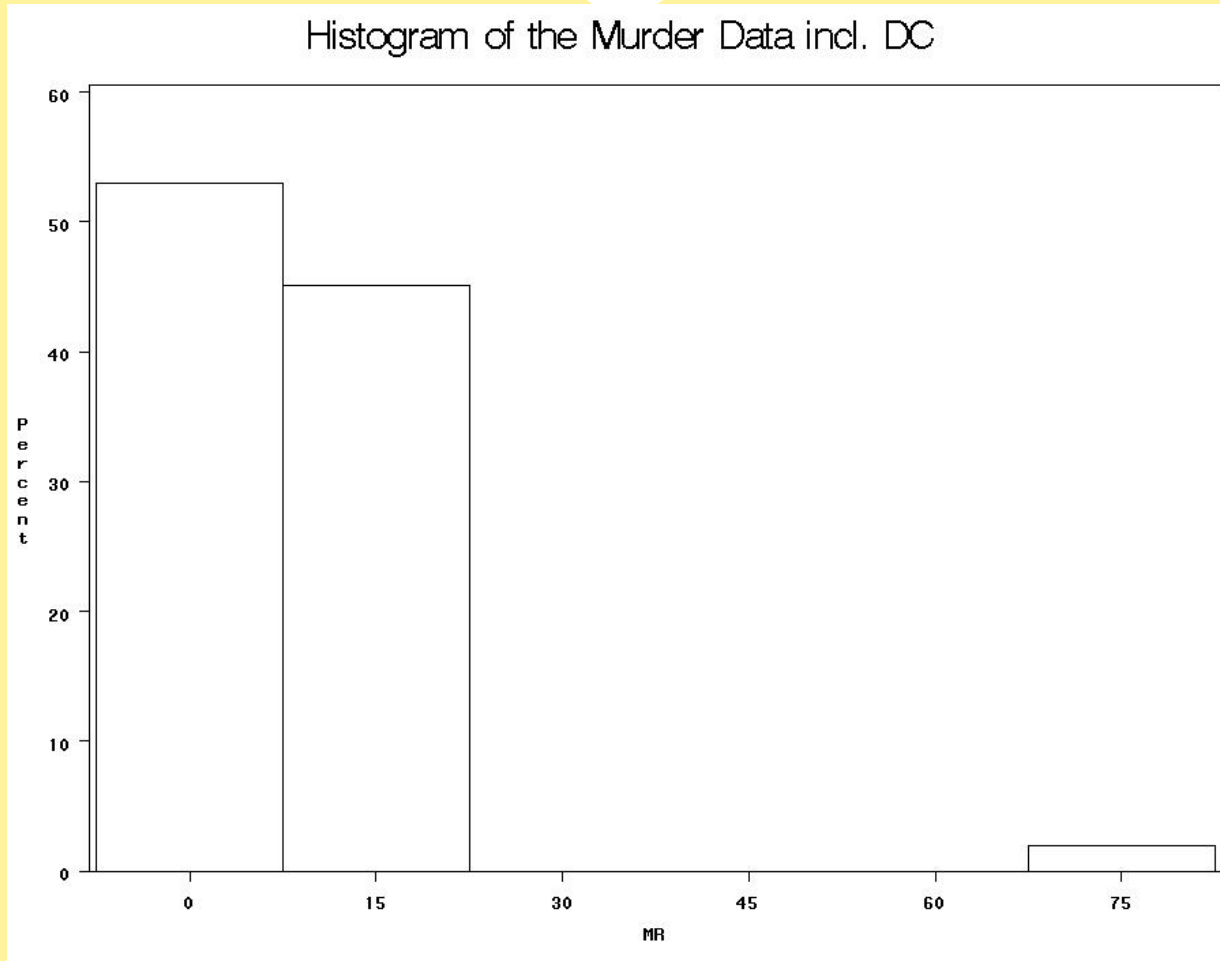
Histogram (Interval Data)

16

- 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.

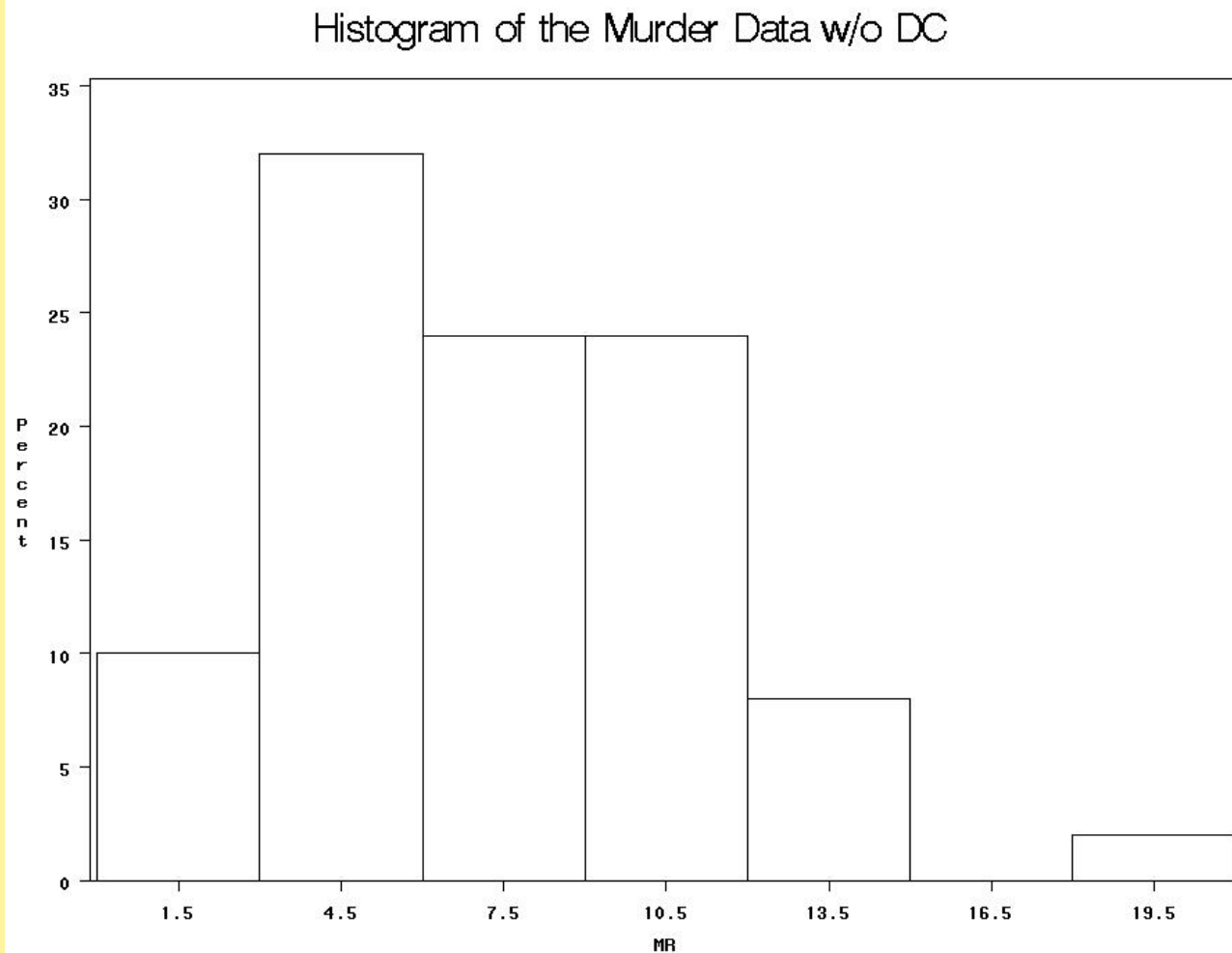
Histogram (version I)

17



Histogram (version II)

18



Bar Graph (Nominal/Ordinal Data)

19

- 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)

20

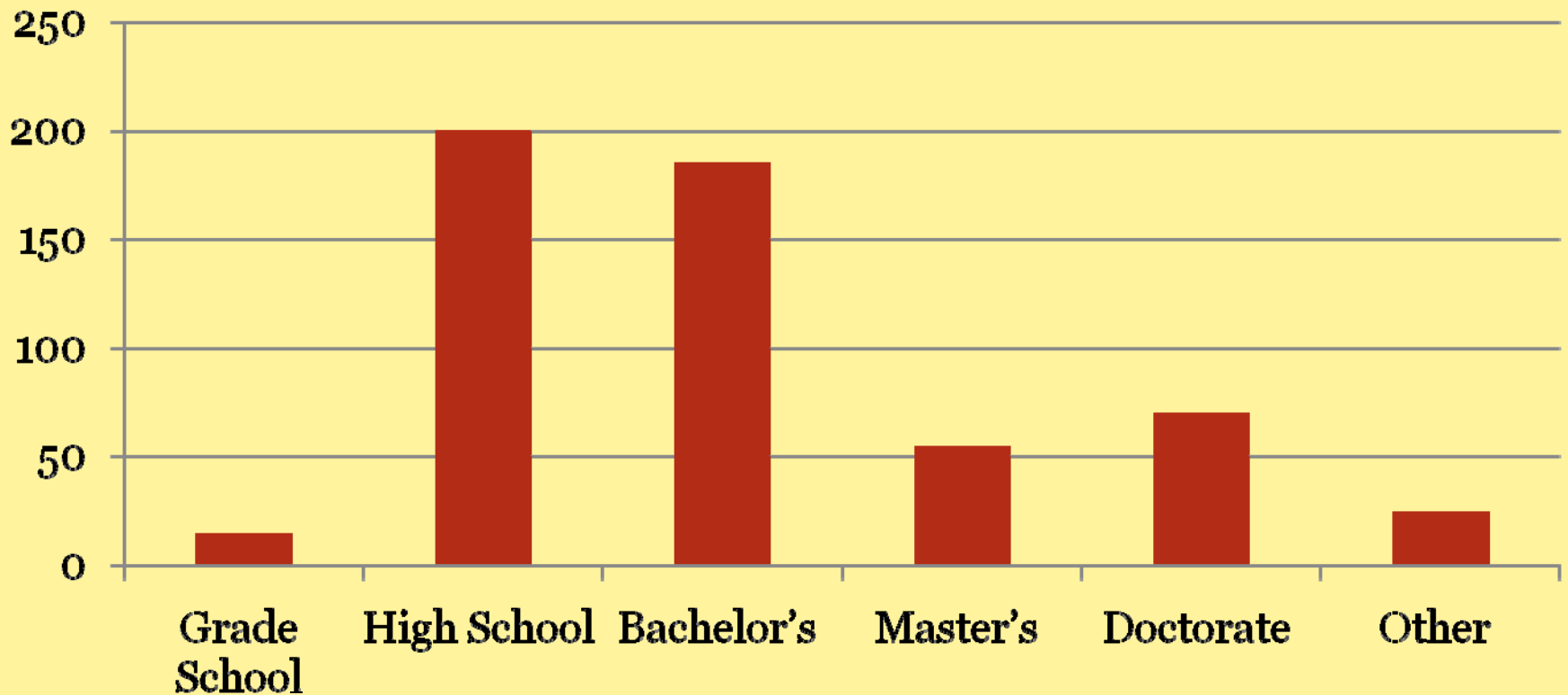
- First Step: Create a Frequency Distribution

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...

21

Bar Graph: *If the data is ordinal, classes are presented in the natural ordering.*



Pie Chart

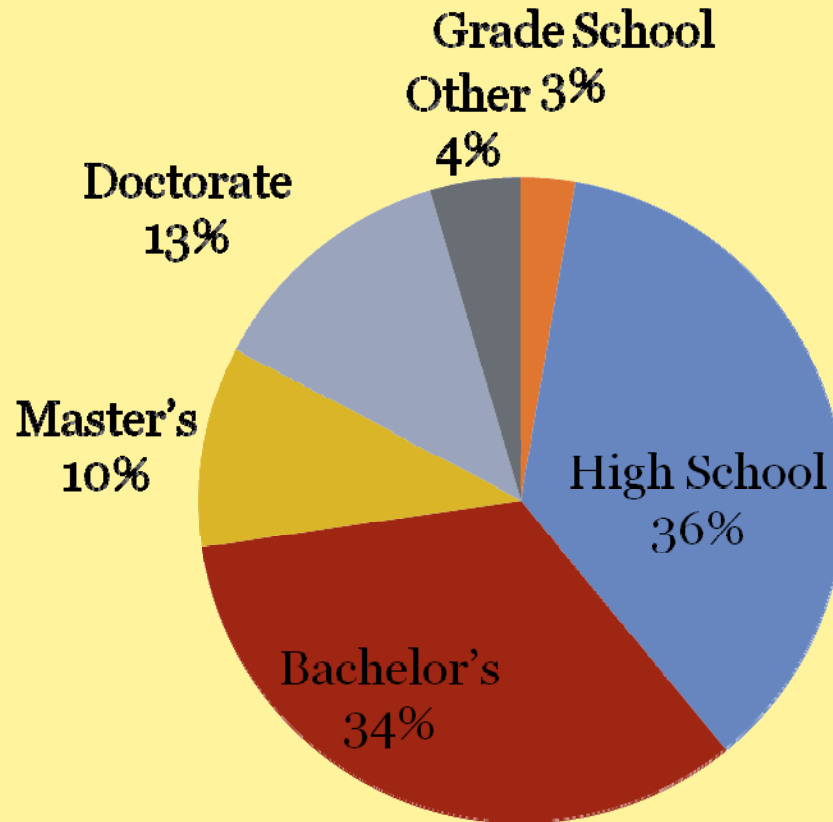
22

- **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

23



**Highest Degree
Earned**

Stem and Leaf Plot

24

- 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

25

- 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

26

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

Attendance Survey Question #4



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