

STA 291 Summer 2008

Lecture 2

- 5.3 Sampling Plans
- 2.2 Graphical and Tabular Techniques for Nominal Data
- 2.3 Graphical Techniques for Interval Data

Review: Basic Terminology I

- **Population**
 - total set of all subjects of interest
 - the entire group of people, animals or things about which we want information
- **Sample**
 - subset of the population from which the study actually collects information
 - used to draw conclusions about the whole population

Review: Basic Terminology II

- **Parameter**
 - numerical characteristic of the population
 - calculated using the whole population
- **Statistic**
 - numerical characteristic of the sample
 - calculated using the sample

Data Collection and Sampling Theory

Why not measure all of the the units in the population? Why not take a census?

Problems:

- *Accuracy*: May not be able to list them all— may not be able to come up with a **frame**.
- *Time*: Speed of Response
- *Expense*: Cost
- *Infinite Population*
- *Destructive Sampling or Testing*

Flavors of Statistics

- **Descriptive Statistics**
 - Summarizing the information in a collection of data
- **Inferential Statistics**
 - Using information from a sample to make conclusions/predictions about the population

Example 1

- University Health Services at UK conducts a survey about alcohol abuse among students.
 - 200 of the 30,000 students are sampled and asked to complete a questionnaire.
 - One question is "have you regretted something you did while drinking"?
 - What is the population? Sample?
 - For the 30,000 students, of interest is the percentage who would respond "yes".
- This value is computed for the students sampled.
Is this a parameter or a statistic?

STA 291 - Fall 2007 - Lecture 2

7

Review: Qualitative Variables (=Categorical Variables) **Nominal or Ordinal**

- Nominal variables have a **scale of unordered categories**
- Ordinal variables have a **scale of ordered categories**

STA 291 - Fall 2007 - Lecture 3

8

Review: Quantitative Variables

- Quantitative variables are measured numerically, that is, for each subject, a **number is observed**
- The scale for quantitative variables is called **interval scale**

STA 291 - Fall 2007 - Lecture 3

9

Review: Nominal, Ordinal, or Interval Scale

- Which scale of measurement is most appropriate for the following variables?
- Attitude towards legalization of marijuana (favor, neutral, oppose)
- Gender (male, female)
- Number of siblings (0,1,2,...)
- Political party affiliation (Democrat, Republican, Independent)

STA 291 - Fall 2007 - Lecture 3

10

Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups ("**strata**") according to some criterion.
- Select a simple random sample independently from each group.

STA 291 - Lecture 2

11

Why could stratification be useful?

- We may want to draw inference about population parameters for each subgroup
- Sometimes, ("proportional stratified sample") estimators from stratified random samples are more precise than those from simple random samples

STA 291 - Lecture 2

12

Proportional Stratification

- The proportions of the different strata are the same in the sample as in the population

- Mathematically:

Population size N , subpopulation sizes N_i

Sample size n , subsample sizes n_i

$$\frac{n_i}{n} = \frac{N_i}{N}$$

Proportional Stratification

- Example:

- Total population of the US: 303 Million
- Population of Kentucky: 4 Million (1.3%)
- Suppose you take a sample of size $n=303$ of people living in the US.
- If stratification is proportional, then 4 people in the sample need to be from Kentucky
- Suppose you take a sample of size $n=1000$. If you want it to be proportional, then 13 people (1.3%) need to be from Kentucky.

Cluster Sampling

- The population can be divided into a set of non-overlapping subgroups (the clusters)
- The clusters are then selected at random, and all individuals in the selected clusters are included in the sample
- Example

Systematic Sampling

- An initial name is selected at random
- every K th name is thereafter selected
- K is computed by dividing membership list length by the desired sample size
- Not a simple random sample (why?), but often almost as good as one
- Example

Summary of Important Sampling Plans

- **Simple Random Sampling (SRS)**
 - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
 - Non-overlapping subgroups (strata)
 - SRSs are drawn from each strata
- **Cluster Sampling**
 - Non-overlapping subgroups (clusters)
 - Clusters selected at random
 - All individuals in the selected clusters are included in the sample
- **Systematic Sampling**
 - Useful when the population consists as a list
 - A value K is specified. Then one of the first K individuals is selected at random, after which every K th observation is included in the sample

Review: Sampling Error

- **Sampling Error** is the error that occurs because a statistic is calculated based on a sample (instead of being based on a population).
- In random samples, the sampling error can usually be quantified.
- In nonrandom samples, there is also sampling variability, but its extent is not predictable.

Chapter 2 Descriptive Statistics

- 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

Alabama	11.6	Alaska	9.0
Arizona	8.6	Arkansas	10.2
California	13.1	Colorado	5.8
Connecticut	6.3	Delaware	5.0
D C	78.5	Florida	8.9
Georgia	11.4	Hawaii	3.8
...		...	

- Difficult to see the “big picture” from these numbers and recognize any patterns
- Try to condense the data...

Frequency Distribution

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

Frequency Distribution

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
- Intervals must be collectively exhaustive (there must be some interval for each observation)
- Rule of thumb:
If you have n observation, the number of intervals should be about \sqrt{n}

Relative Frequencies

- Relative frequency for an interval: The proportion of sample observations that fall in that interval
- Sometimes, percentages are preferred to relative frequencies

Frequency and Relative Frequency and Percentage Distribution

Murder Rate	Frequency	Relative Frequency	Percentage
0-2.9	5	.10	10
3-5.9	16	.31	31
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

STA 291 - Lecture 2

25

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)

STA 291 - Lecture 2

26

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

STA 291 - Lecture 2

27

Frequency and Cumulative Frequency

Murder Rate	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
0-2.9	5	.10		
3-5.9	16	.31		
6-8.9	12	.24		
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		

STA 291 - Lecture 2

28

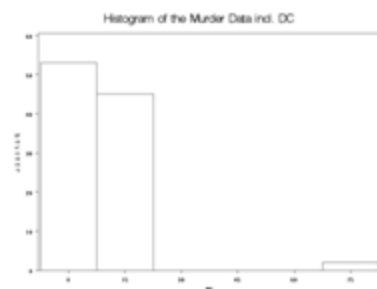
Histogram (Interval Data)

- 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.

STA 291 - Lecture 2

29

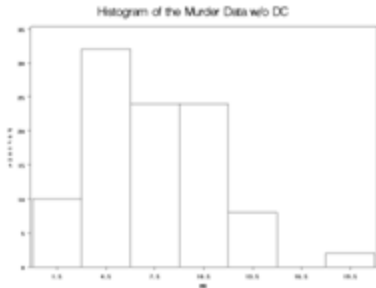
Histogram



STA 291 - Lecture 2

30

Histogram w/o DC



STA 291 - Lecture 2

31

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

STA 291 - Lecture 2

32

Pie Chart (Nominal/Ordinal Data)

First Step: Create a Frequency Distribution

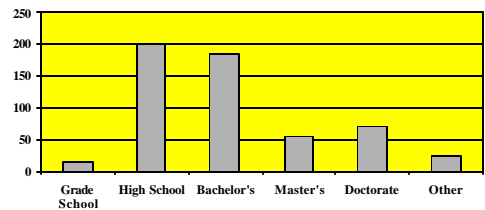
Highest Degree	Frequency (Number of Employees)	Relative Frequency
Grade School	15	
High School	200	
Bachelor's	185	
Master's	55	
Doctorate	70	
Other	25	
Total	550	

STA 291 - Lecture 2

33

We could display this data
in a bar chart...

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



STA 291 - Lecture 2

34

Pie Chart

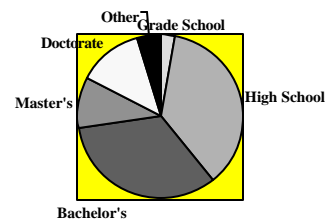
- Pie Chart: Pie is divided into slices; The area of each slice is proportional to the frequency of each class.

Highest Degree	Relative Frequency	Angle (= Rel. Freq. x 360°)
Grade School	$15/550 = .027$	9.72
High School	$200/550 = .364$	131.04
Bachelor's	$185/550 = .336$	120.96
Master's	$55/550 = .1$	36.0
Doctorate	$70/550 = .127$	45.72
Other	$25/550 = .045$	16.2

STA 291 - Lecture 2

35

Pie Chart for Highest Degree Achieved



STA 291 - Lecture 2

36

Stem and Leaf Plot (Interval Data)

Stem	Leaf	#
20	3	1
19		
18		
17		
16		
15		
14		
13	135	3
12	7	1
11	334469	6
10	2234	4
9	08	2
8	03469	5
7	5	1
6	034689	6
5	0238	4
4	46	2
3	0144468999	10
2	039	3
1	67	2

STA 291 - Lecture 2

37

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

STA 291 - Lecture 2

38

Stem and Leaf Plot

- Useful for small data sets (<100 observations)
- 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”

STA 291 - Lecture 2

39

Stem and Leaf Plot

- Can also be used to compare groups: Back-to-Back Stem and Leaf Plots, using the same stems for both groups.
- Murder Rate Data from U.S. and Canada
- By the way, it doesn't really matter whether the smallest stem is at top or bottom of the table

STA 291 - Lecture 2

40

AGE AT DEATH OF U.S. PRESIDENTS

PRESIDENT	AGE	PRESIDENT	AGE	PRESIDENT	AGE
Washington	87	Fillmore	74	Goosevelt	80
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

Making Stem and Leaf plots

Stems Leaves

4
5
6
7
8
9

STA 291 - Lecture 2

41

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

STA 291 - Lecture 2

42

Describing Distributions

- Symmetric distributions
 - Bell-shaped or U-shaped
- Not symmetric distributions:
 - Left-skewed or right-skewed

STA 291 - Lecture 2

43

Summary of Univariate Graphical and Tabular Techniques

- Discrete data: Frequency distribution
- Continuous data: Grouped frequency distribution
- Small data sets: Stem and leaf plot
- Interval data: Histogram
- Categorical data: Bar chart, Pie chart
- Grouping intervals should be of same length, but may be dictated more by subject-matter considerations
- Always use common sense

STA 291 - Lecture 2

44

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 2x3 table ("2 by 3")

STA 291 - Lecture 2

45

2x2 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

STA 291 - Lecture 2

46

Contingency Table: Example, contd.

- 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

STA 291 - Lecture 2

47

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

STA 291 - Lecture 2

48

Describing the Relationship Between Two Interval Variables

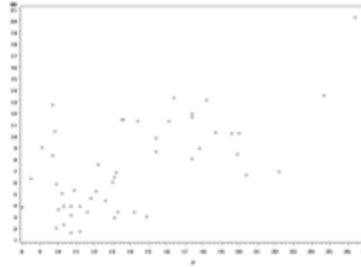
Scatter Diagram

- 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

STA 291 - Lecture 2

49

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



[Correlation and Scatterplot Applet](#)

[Correlation by Eye Applet](#)

[Simple Regression Analysis Tool](#)

STA 291 - Lecture 2

50

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

STA 291 - Lecture 2

51

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

STA 291 - Lecture 2

52

Good vs. Bad Graphics

- Please read Chapter 3 about the Art & Science of graphical presentations

STA 291 - Lecture 2

53