

Lecture 2

Displaying and Describing Categorical Data

Relative Frequency Table

ISU Student Enrollment by Class, 2005

Class	Frequency	Relative Frequency (%)	
Freshmen	4,189	23.50	= 4,189 / 17,827 * 100
Sohomore	3,271	18.35	= 3,271 / 17,827 * 100
Junior	4,648	26.07	= 4,648 / 17,827 * 100
Senior	5,658	31.74	= 5,658 / 17,827 * 100
Unclassified	61	0.34	= 61 / 17,827 * 100
Total	17,827	100	

4

Today's Objectives

- Identify and discuss appropriate methods for displaying categorical data
 - Frequency tables, bar charts and pie charts
- Explore whether an association exist between two categorical variables
 - Contingency tables, and clustered and stacked bar charts

2

Comparisons Across Groups

Which table does a better job of comparing class distributions across universities?

Table 1

Class	Public Frequency	Private Frequency
Freshmen	4,189	642
Sohomore	3,271	513
Junior	4,648	487
Senior	5,658	498
Unclassified	61	0
Total	17,827	2,140

Table 2

Class	Public (%)	Private (%)
Freshmen	23.50	30.00
Sohomore	18.35	23.97
Junior	26.07	22.76
Senior	31.74	23.27
Unclassified	0.34	0.00
Total	100	100

5

Frequency Table

ISU Student Enrollment by Class, 2005

Class	Frequency
Freshmen	4,189
Sohomore	3,271
Junior	4,648
Senior	5,658
Unclassified	61
Total	17,827

- A frequency table records each value (i.e. category) that a variable might have and gives the counts of observations in each category.

Source: *Illinois State University Facts*, Planning and Institutional Research, October 2005, www.pirilstu.edu

3

More ISU Facts!

Racial/Ethnic Designation of ISU Students in 2005

	Frequency	(%)
American Indian / Alaskan Native	53	0.26
Black/Non-Hispanic	1,212	5.98
Asian/Pacific Islander	340	1.68
Hispanic	624	3.08
White/Non-Hispanic	16,936	83.57
Not Reported	1100	5.43
Total	20,265	100

Note: Total includes graduate students

6

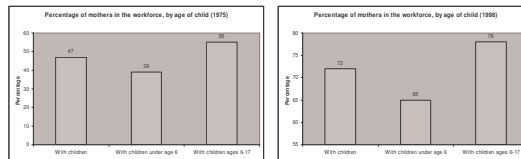
Even More ISU Facts!

Enrollment By College in 2005

	Frequency	(%)
Applied Science and Technology	3,560	17.57
Arts and Sciences	5,939	29.31
Business	3,081	15.20
Education	3,010	14.85
Fine Arts	1,185	5.85
Nursing	568	2.80
Other	2,922	14.42
Total	20,265	100

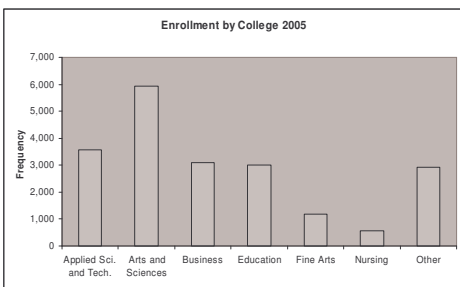
7

Comparisons Using Bar Charts



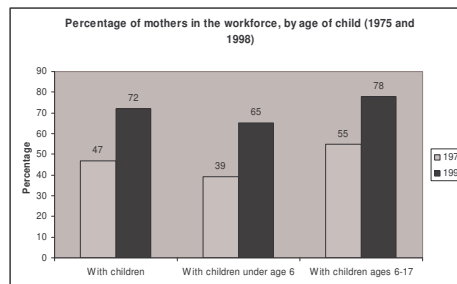
10

Bar Charts



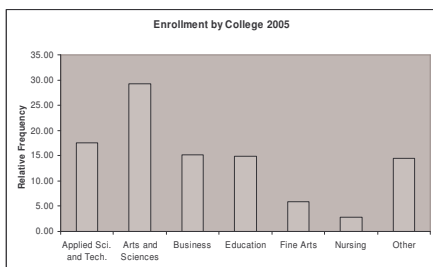
8

Clustered Bar Charts



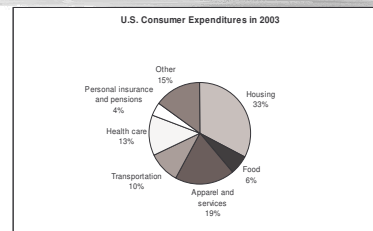
11

Bar Charts (Continued)



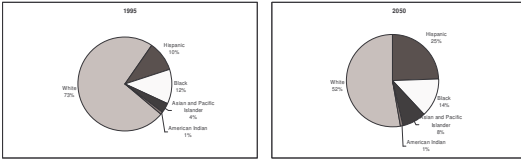
9

Pie Charts



A pie chart represents data in the form of slices or sections of a circle. Each slice represents a category, and the size of the slice is proportional to the relative frequency of the category. Data source: Bureau of Labor Statistics, Consumer Expenditure Survey, 2003 12

Comparisons Using Pie Charts – Predicting Population Trends



Source: U.S. Census Bureau

13

Contingency Table (Two-Way Table)

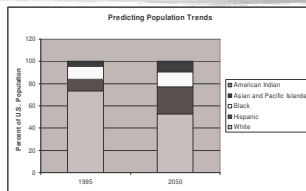
	Smoker	Nonsmoker	Total
High School	32	61	93
2-yr College	5	17	22
4+yr College	13	72	85
Total	50	150	200

•A Contingency table is a table with rows that represent the possible values of one variable and columns that represent the possible values for a second variable.

•Excellent tool for exploring whether there is an association between two categorical variables.

16

Segmented Bar Charts



- Alternative to comparing multiple pie charts
- The data for the selected variable (race/ethnicity) are represented as a percentage of the total for each category (year) of the second variable.
- Bars are stacked to total 100%

14

Conditional Distribution

What percent of highly educated people smoke?

		Smoker	Nonsmoker	Total
4+yr College	Count	13	72	85
	% of row	15.3	84.7	100

Conditional distribution - shows the distribution of one variable (smoking status) for just the observations that satisfy some condition on another variable (education - 4+yr college).

17

Exploring Association between two Variables

Education		Smoking Status	
Class	Frequency	Class	Frequency
High School	93	Smoker	50
2-yr College	22	Nonsmoker	150
4+yr College	85	Total	200
Total	200		

Is there an association between education level and smoking?

15

Comparing Conditional Distributions

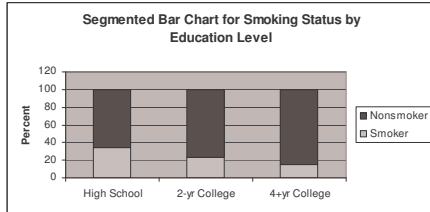
Are highly educated individuals more or less likely to be smokers than those with less education?

		Smoker	Nonsmoker	Total
High School	Count	32	61	93
	% of row	34.4	65.6	100.0
2-yr College	Count	5	17	22
	% of row	22.7	77.3	100.0
4+yr College	Count	13	72	85
	% of row	15.3	84.7	100

Are the conditional distributions similar or different? If they are similar, we say that the variables are independent. If they are different, we say that there appears to be an association between the two variables.

18

Comparing Conditional Distributions



19

Expanded Contingency Table

		Smoker	Nonsmoker	Total
High School	Count	32	61	93
	% of row	34.4	65.6	100.0
	% of column	64.0	40.7	46.5
	% of total	16.0	30.5	42.5
2-yr College	Count	5	17	22
	% of row	22.7	77.3	100.0
	% of column	10.0	11.3	11.0
	% of total	2.5	8.5	11.0
4+yr College	Count	13	72	85
	% of row	15.3	84.7	100.0
	% of column	26.0	48.0	42.5
	% of total	6.5	36.0	42.5
Total	Count	50	150	200
	% of row	25.0	75.0	100.0
	% of column	100.0	100.0	100.0
	% of total	25.0	75.0	100.0

22

Comparing Conditional Distributions

Are nonsmokers more likely to be highly educated?

		Smoker	Nonsmoker
High School	Count	32	61
	% of column	64.0	40.7
2-yr College	Count	5	17
	% of column	10.0	11.3
4+yr College	Count	13	72
	% of column	26.0	48.0
Total	Count	50	150
	% of column	100	100

Table shows the conditional distribution of education, conditional on smoking status

Does it appear that education and smoking status are independent? Explain

20

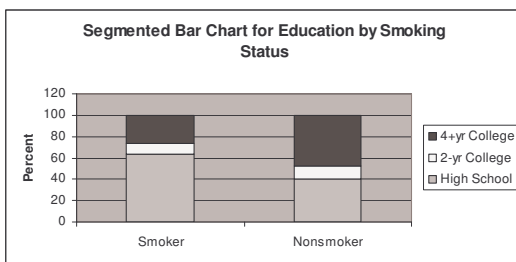
Example

		Smoker	Nonsmoker	Total
Education	High School	32	61	93
	2-yr College	5	17	22
	4+yr College	13	72	85
	Total	50	150	200

- What percent of individuals are high school grads?
- What percent of high school grads are smokers?
- What percent of smokers are high school grads?
- What percent of individuals are high school grads who smoke?

23

Comparing Conditional Distributions



21

Assignment

- ☐ Read Chapter 4: Displaying Quantitative Data
- ☐ Try the following exercises from Chapter 3:
 - #7,11,14,19,21,23,29

24