03/17/2000

Relevant Data

Teaching Statistics with Relevant Data

1

Western Statistics Teachers Conference 2000

March 17, 2000

MILO SCHIELD

Augsburg College

www.augsburg.edu/ppages/schield

schield@augsburg.edu

03/17/2000

Relevant Data

Relation between Statistics and Data

2

David Moore:
Statistics is "the science of data" --
"the science of gaining information from data"
"Data are numbers with a context."

"We recommend that … modern statistics begin with data analysis, both because concrete experience with data motivates the more abstract parts of our subjects and because exploring even haphazardly produced data can provide insight." ***MAA Notes Number 21***

03/17/2000

Relevant Data

Goal-Related Criteria for Relevant Data

3

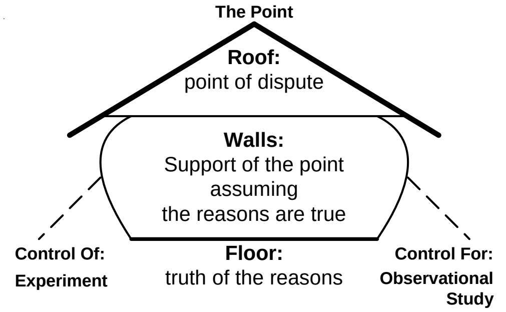
Statistical Inference	**Data Analysis and Modeling**
Experiment & Q/C: *Wardrup; MSMESB*	Relation of variables: *Macnaughton*
Causality: *Pearl,Robbins,Rubin*	Statistical Literacy: *Schield*

03/17/2000

Relevant Data

Statistical Literacy Supporting Arguments

4



The Point

Roof:
point of dispute

Walls:
Support of the point
assuming
the reasons are true

Control Of:
Experiment

Floor:
truth of the reasons

Control For:
Observational Study

03/17/2000

Relevant Data

Statistical Literacy Criteria for Relevant Data

5

External Criteria:

- "broad vistas" Eric Sowey
- "socially relevant" Donald Macnaughton

Internal Criteria :

- Multivariate (complex) data
- Mostly observational (non-experimental) data
- Allow modeled variable to be binary
- Includes spurious associations.

03/17/2000

Relevant Data

Relevant Data Sets Multiple Related Variables

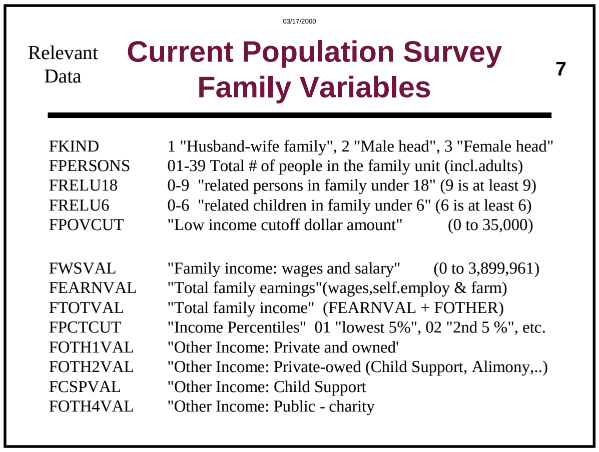
6

Current Population Survey Microdata
Subset prepared by Schield.

Framingham Study Data **from**
Entered by Schield from Epidemiology text.

The Bell Curve data
received by Schield from Charles Murray.

[Goal: put these datasets on the web for general access]



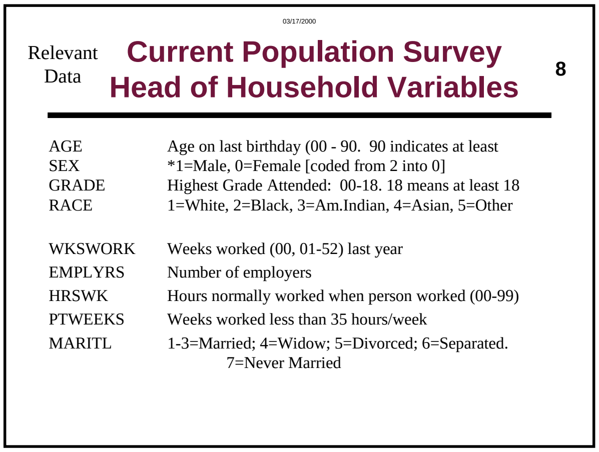
03/17/2000

Relevant Data

Current Population Survey Family Variables

7

FKIND	1 "Husband-wife family", 2 "Male head", 3 "Female head"
FPERSONS	01-39 Total # of people in the family unit (incl.adults)
FRELU18	0-9 "related persons in family under 18" (9 is at least 9)
FRELU6	0-6 "related children in family under 6" (6 is at least 6)
FPOVCUT	"Low income cutoff dollar amount" (0 to 35,000)
FWSVAL	"Family income: wages and salary" (0 to 3,899,961)
FEARNVAL	"Total family earnings"(wages,self.employ & farm)
FTOTVAL	"Total family income" (FEARNVAL + FOTHER)
FPCTCUT	"Income Percentiles" 01 "lowest 5%", 02 "2nd 5 %", etc.
FOTH1VAL	"Other Income: Private and owned'
FOTH2VAL	"Other Income: Private-owed (Child Support, Alimony,..)
FCSPVAL	"Other Income: Child Support
FOTH4VAL	"Other Income: Public - charity



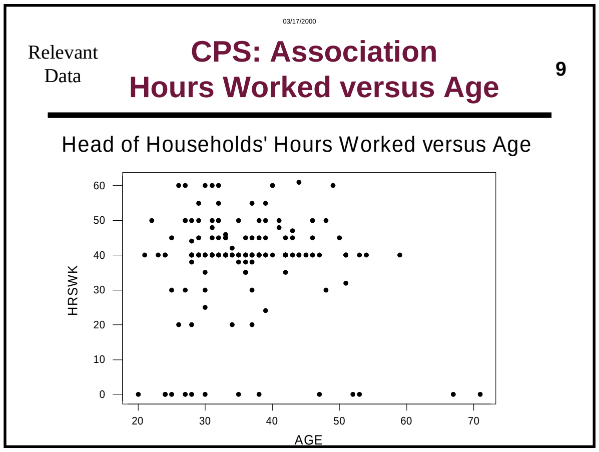
03/17/2000

Relevant Data

Current Population Survey Head of Household Variables

8

AGE	Age on last birthday (00 - 90. 90 indicates at least
SEX	*1=Male, 0=Female [coded from 2 into 0]
GRADE	Highest Grade Attended: 00-18. 18 means at least 18
RACE	1=White, 2=Black, 3=Am.Indian, 4=Asian, 5=Other
WKSWORK	Weeks worked (00, 01-52) last year
EMPLYRS	Number of employers
HRSWK	Hours normally worked when person worked (00-99)
PTWEEKS	Weeks worked less than 35 hours/week
MARITL	1-3=Married; 4=Widow; 5=Divorced; 6=Separated. 7=Never Married



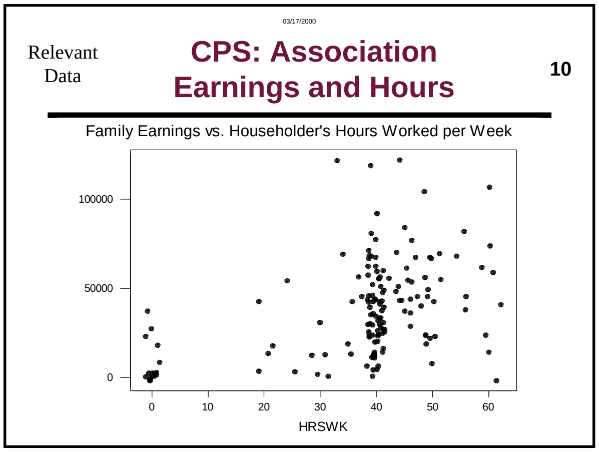
03/17/2000

Relevant Data

CPS: Association Hours Worked versus Age

9

Head of Households' Hours Worked versus Age



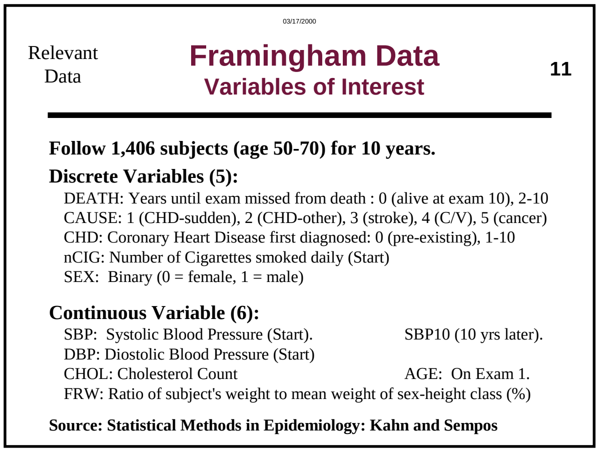
03/17/2000

Relevant Data

CPS: Association Earnings and Hours

10

Family Earnings vs. Householder's Hours Worked per Week



03/17/2000

Relevant Data

Framingham Data Variables of Interest

11

Follow 1,406 subjects (age 50-70) for 10 years.

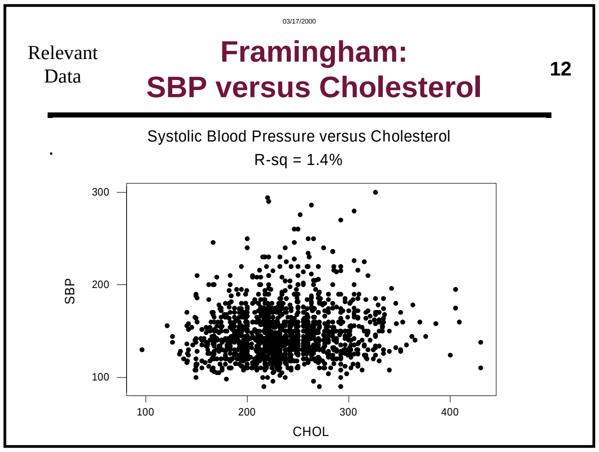
Discrete Variables (5):

DEATH: Years until exam missed from death : 0 (alive at exam 10), 2-10
CAUSE: 1 (CHD-sudden), 2 (CHD-other), 3 (stroke), 4 (C/V), 5 (cancer)
CHD: Coronary Heart Disease first diagnosed: 0 (pre-existing), 1-10
nCIG: Number of Cigarettes smoked daily (Start)
SEX: Binary (0 = female, 1 = male)

Continuous Variable (6):

SBP: Systolic Blood Pressure (Start). SBP10 (10 yrs later).
DBP: Diostolic Blood Pressure (Start)
CHOL: Cholesterol Count AGE: On Exam 1.
FRW: Ratio of subject's weight to mean weight of sex-height class (%)

Source: Statistical Methods in Epidemiology: Kahn and Sempos



03/17/2000

Relevant Data

Framingham: SBP versus Cholesterol

12

Systolic Blood Pressure versus Cholesterol
R-sq = 1.4%

03/17/2000

Relevant Data

The Bell Curve Binary Outcome Variables

13

- being in poverty (p. 136)
- being out of the labor force for a month or more (p. 159)
- being prevented from working by health problems (p. 161)
- being unemployed for a month in 1989 (p.164)
- having ever been married by age 30 (p. 172)
- first child conceived before marriage (p. 181)
- mothers having a first child born out of wedlock (p. 188)
- mothers on AFDC within a year of first birth (p. 193)
- being a chronic welfare recipient (p. 198)
- being interviewed in a penal institution (p.___)

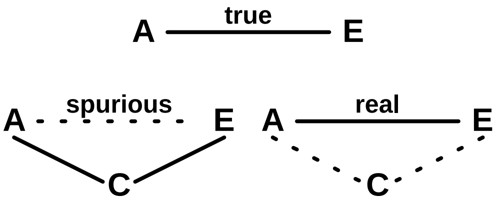
03/17/2000

Relevant Data

Association: Spurious?

14

Association Between A and E

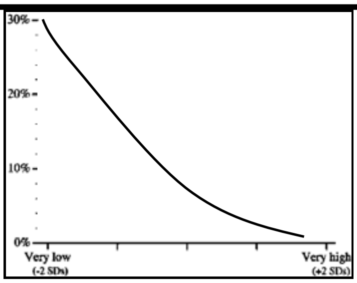


03/17/2000

Relevant Data

Risk of Poverty by Parents' SocioEconomic Status

15



30% of [these with low-SES parents] <u>are in poverty</u>.

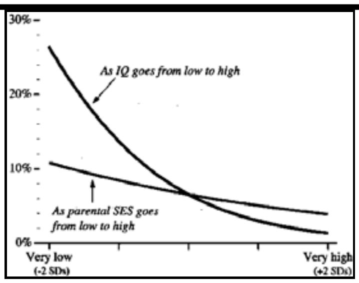
Your <u>risk of poverty</u> is 30% if your parents were low SES.

03/17/2000

Relevant Data

Risk of Poverty: Offset normalized to mean

16



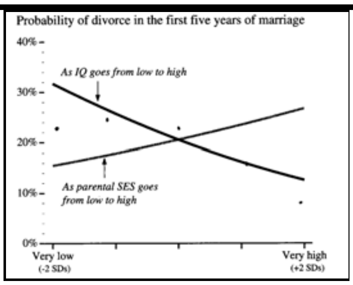
25% of [these very low-IQ adults who grew up in families with average socioeconomic status] <u>are in poverty</u>.

03/17/2000

Relevant Data

The Bell Curve: Risk of Divorce (1st 5 years)

17



32% of very low-IQ adults <u>were divorced in first 5 years</u>.
Your <u>divorce risk in first 5 years</u> is 32% if you have low IQ.

03/17/2000

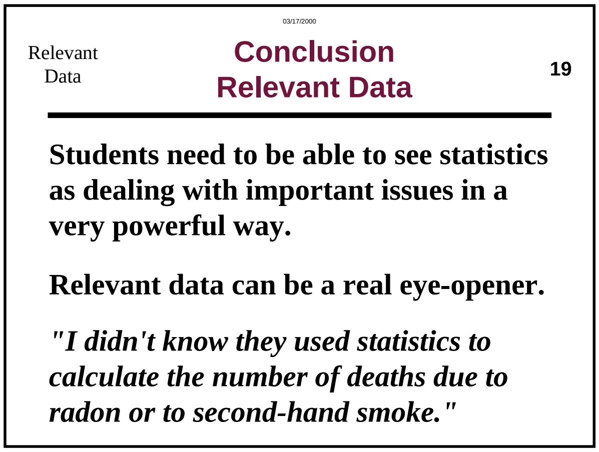
Relevant Data

Conclusion Spurious Associations

18

Students need to see that an association can be spurious even though the association is still true.

"I didn't know a race-based association could be overturned by taking into account confounding factors."



03/17/2000

Relevant Data

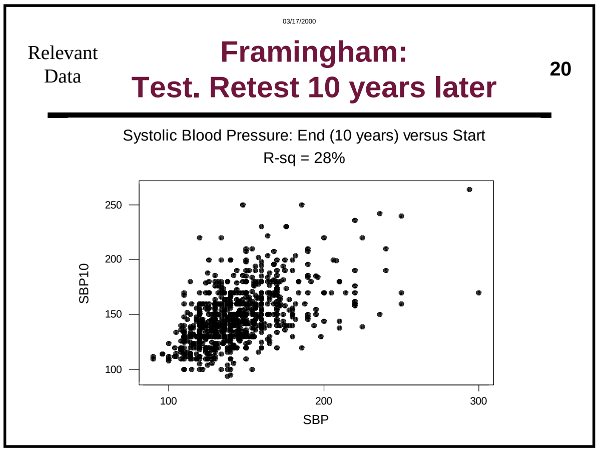
Conclusion Relevant Data

19

Students need to be able to see statistics as dealing with important issues in a very powerful way.

Relevant data can be a real eye-opener.

"I didn't know they used statistics to calculate the number of deaths due to radon or to second-hand smoke."



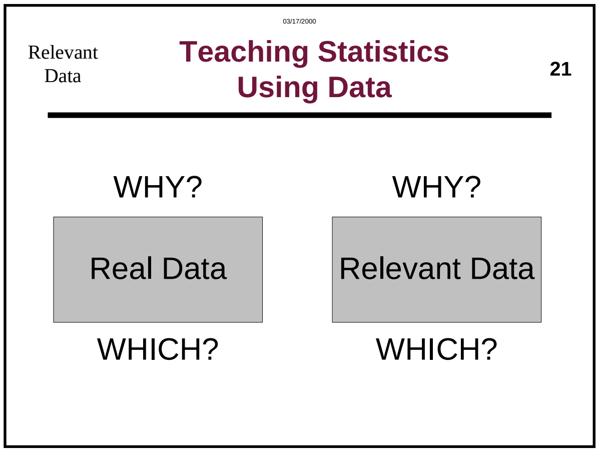
03/17/2000

Relevant Data

Framingham: Test. Retest 10 years later

20

Systolic Blood Pressure: End (10 years) versus Start
R-sq = 28%



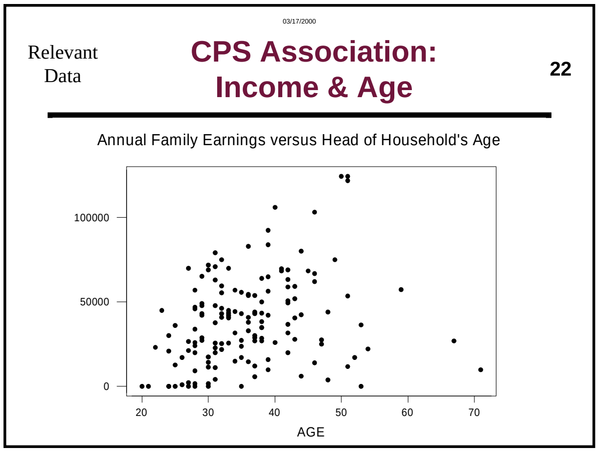
03/17/2000

Relevant Data

Teaching Statistics Using Data

21

WHY?	WHY?
Real Data	Relevant Data
WHICH?	WHICH?



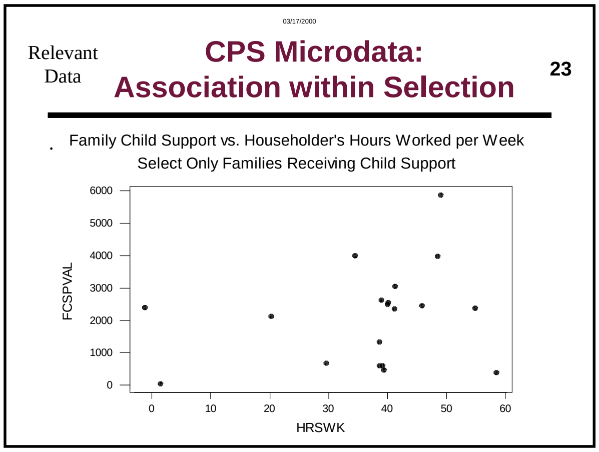
03/17/2000

Relevant Data

CPS Association: Income & Age

22

Annual Family Earnings versus Head of Household's Age



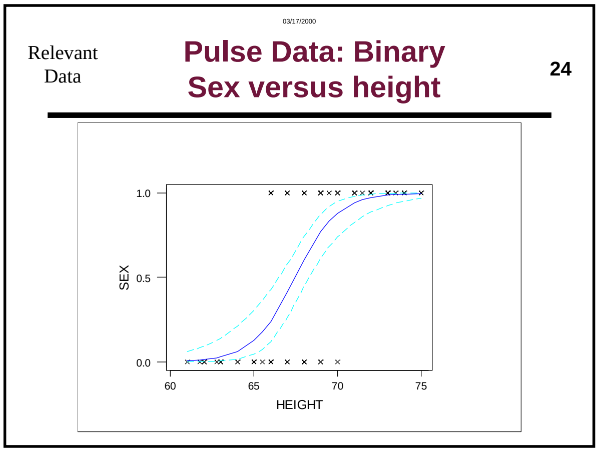
03/17/2000

Relevant Data

CPS Microdata: Association within Selection

23

. Family Child Support vs. Householder's Hours Worked per Week
Select Only Families Receiving Child Support



03/17/2000

Relevant Data

Pulse Data: Binary Sex versus height

24

03/17/2000

Relevant
Data

Statistics and Data
Importance of Real data

25

- '75: Exploratory Data Analysis, Tukey
- '78: Statistics: Freedman, Piasani, Purves
- '83: Statistics with Data, Mosteller et al.
- '85: Exploring Data, Mosteller and Tukey
- '90 Modern Data Analysis, Hamilton
- '98 Understanding Data, Griffiths et al.

03/17/2000

Relevant
Data

The Bell Curve
Explanatory Variables

26

Analysis of binary outcomes was done using three explanatory variables:

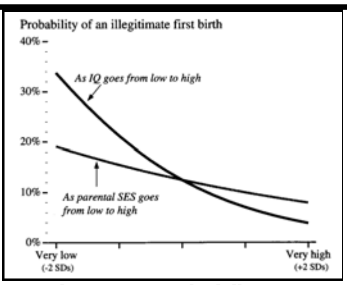
- intelligence (IQ),
- education, and
- family socioeconomic status (SES).

03/17/2000

Relevant
Data

The Bell Curve:
Risk of Illegitimate First Birth

27



34% of these very low-IQ women had illegitimate 1st births.
Woman's risk of illegitimate 1st birth is 34% if very low IQ.