MSIT 431 Probability and Statistical Methods

Lecture 2 Looking at Data - Relationship

Dongning Guo

Fall 2018

Outline

2.1 Relationships

2.2 Scatterplots

2.3 Correlation

2.4 Least-Squares Regression

2.5 Cautions about Correlation and Regression

2.6 Data Analysis for Two-Way Tables

2.7 The Question of Causation

2.1 Relationships

- What is an association between variables?
- Explanatory and response variables
- Key characteristics of a data set

3

Associations Between Variables

Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

4

Key Characteristics of a Data Set

Certain characteristics of a data set are key to exploring the relationship between two variables. These should include the following:

- ✓ Cases: Identify the cases and how many there are in the data set.
- ✓ Label: Identify what is used as a label variable if one is present.
- ✓ Categorical or quantitative: Classify each variable as categorical or quantitative.
- ✓ Values: Identify the possible values for each variable.
- ✓ Explanatory or response: If appropriate, classify each variable as explanatory or response.

5

2.2 Scatterplots

- Scatterplots
- Interpreting scatterplots
- Categorical variables in scatterplots

6

Scatterplot

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual/case corresponds to one point on the graph.

How to Make a Scatterplot

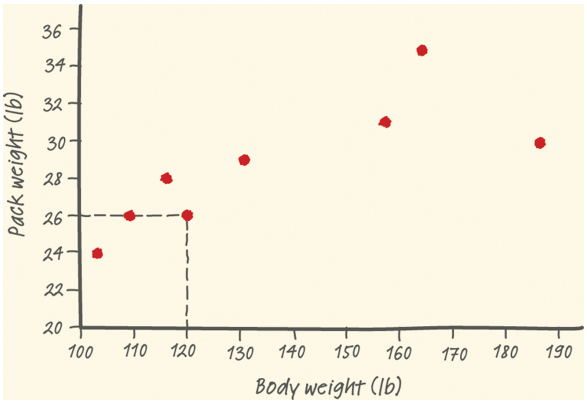
1. Decide which variable should go on each axis. If a distinction exists, plot the explanatory variable on the *x* axis and the response variable on the *y* axis.
2. Label and scale your axes.
3. Plot individual data values.

7

Scatterplot

Example: Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28



9

Interpreting Scatterplots

To interpret a scatterplot, look for patterns and important departures from those patterns.

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *deviations* from that pattern.

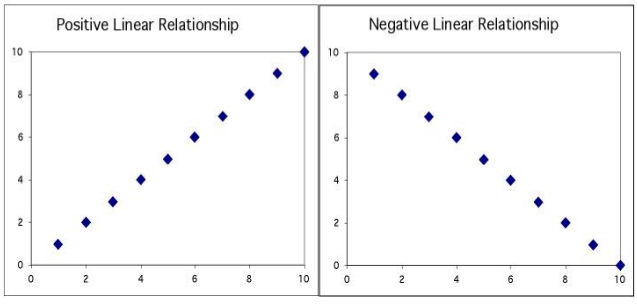
- You can describe the overall pattern of a scatterplot by the **direction, form,** and **strength** of the relationship.
- An important kind of departure is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

10

Interpreting Scatterplots

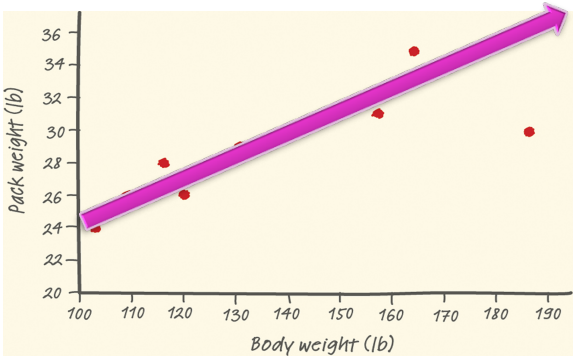
Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice-versa.



11

Interpreting Scatterplots



Outlier

✓ There is one possible outlier—the hiker with the body weight of 187 pounds seems to be carrying relatively less weight than are the other group members.

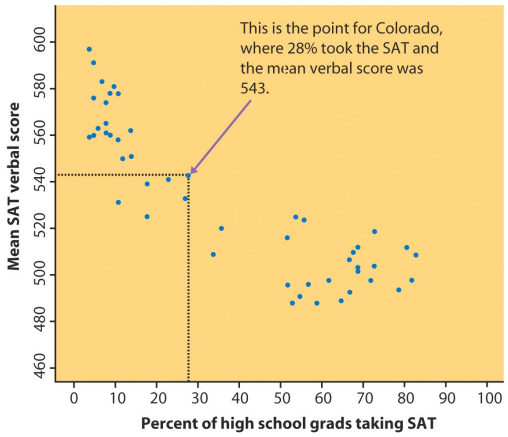
Strength **Direction** **Form**

✓ There is a moderately strong, positive, linear relationship between body weight and backpack weight.
✓ It appears that lighter hikers are carrying lighter backpacks.

12

Adding Categorical Variables

- Consider the relationship between mean SAT verbal score and percent of high school grads taking the SAT for each state.

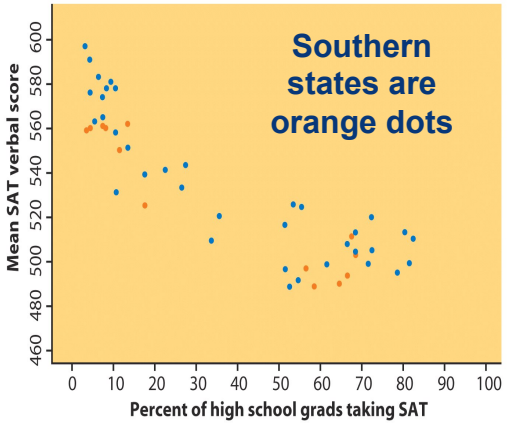


10

Adding Categorical Variables

- Consider the relationship between mean SAT verbal score and percent of high school grads taking the SAT for each state.

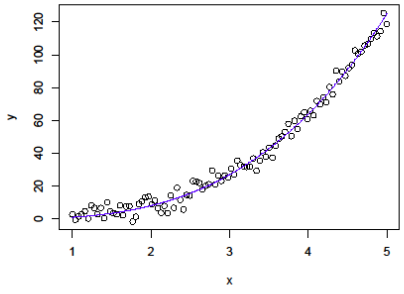
To add a *categorical variable*, use a different plot color or symbol for each category.



10

Nonlinear Relationships

There are other forms of relationships besides linear. The scatterplot below is an example of a nonlinear form. (a curve rather than a straight line)

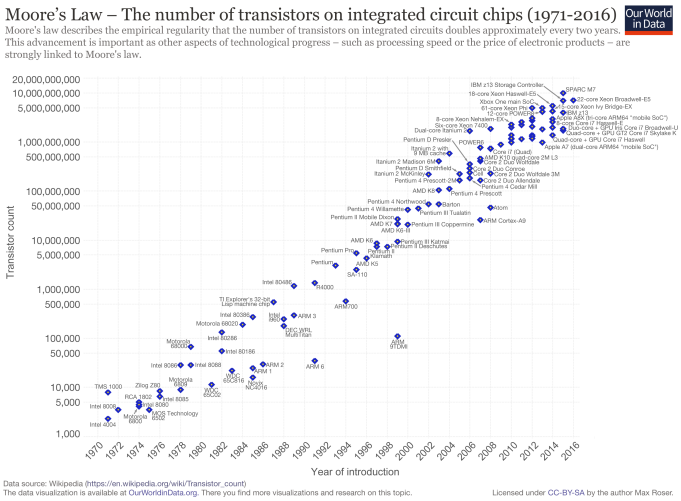


15

Notable example: Moore's law

The number of transistors in a dense integrated circuit doubles approximately every two years. --Gordon E. Moore, co-founder of Intel, 1965.

logarithmic scale



16

2.3 Correlation

- The correlation coefficient *r*
- Properties of *r*
- Influential points

17

Measuring Linear Association

Linear relationships are important because a straight line is a simple pattern that is quite common.

Our eyes are not always good judges of how strong a relationship is. Therefore, we use a numerical measure to supplement our scatterplot and help us interpret the strength of the linear relationship.

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

18

Measuring Linear Association

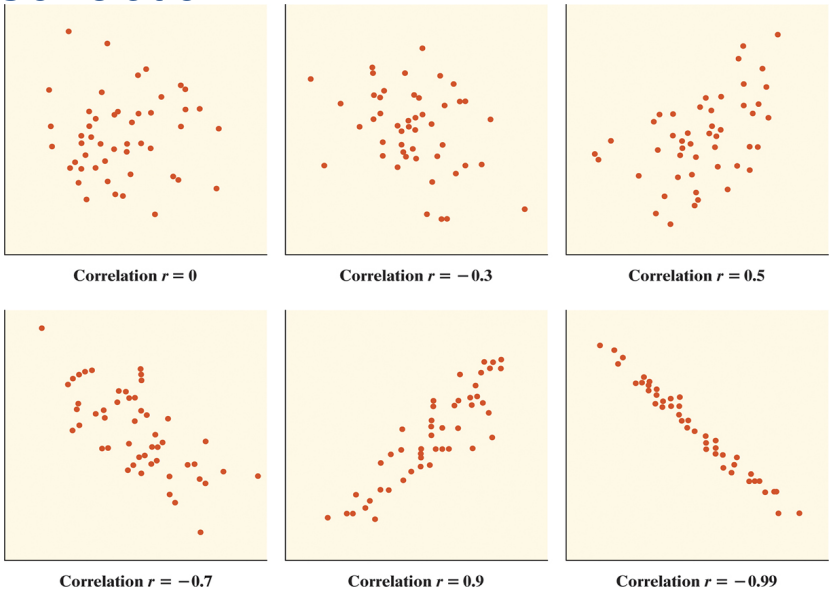
We say a linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line.

Properties of Correlation

- *r* is always a number between –1 and 1.
- *r* > 0 indicates a positive association.
- *r* < 0 indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward –1 or 1.
- The extreme values *r* = –1 and *r* = 1 occur only in the case of a perfect linear relationship.
- No distinction between explanatory and response variables.
- *r* has no units and does not change when we change the units of measurement of *x*, *y*, or both.

19

Correlation



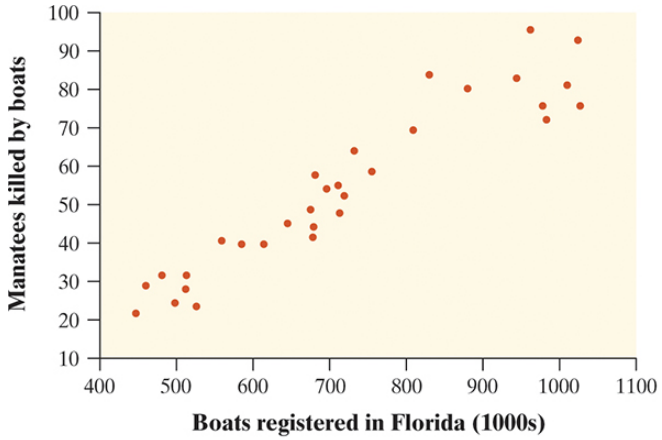
20

Cautions

- Correlation requires that both variables be quantitative.
- Correlation *does not describe curved relationships* between variables, no matter how strong the relationship is.
- The correlation *r* is not resistant; it can be strongly affected by a few outlying observations.
- Correlation is not a complete summary of two-variable data.

21

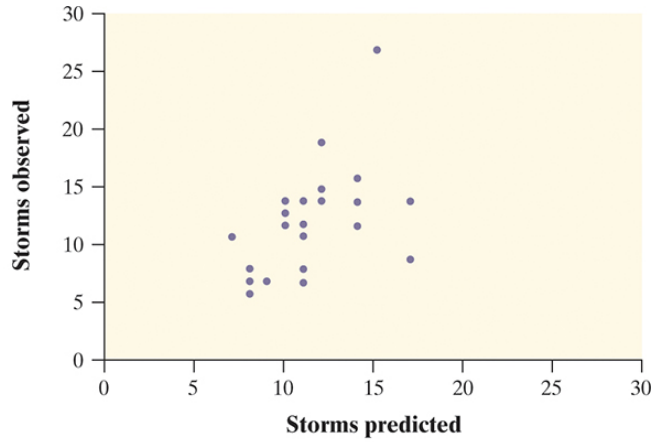
Correlation Examples



(a)

22

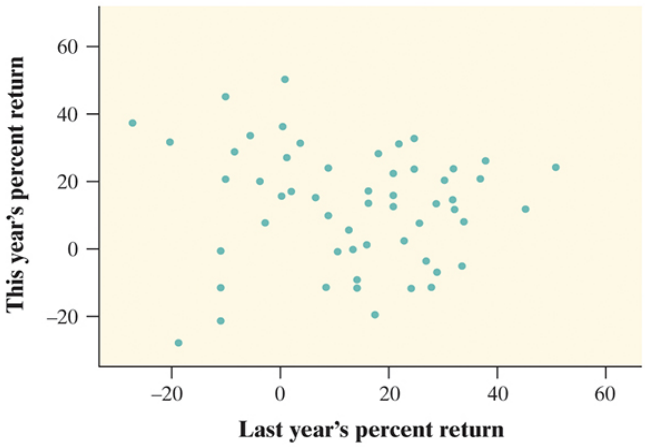
Correlation Examples



(b)

23

Correlation Examples



(d)

24

2.4 Least-Squares Regression

- Regression lines
- Least-squares regression line
- Facts about least-squares regression
- Correlation and regression

25

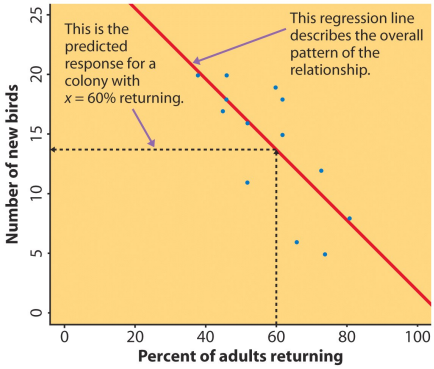
Regression Line

A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

We can use a regression line to predict the value of *y* for a given value of *x*.

Example: Predict the number of new adult birds that join the colony based on the percent of adult birds that return to the colony from the previous year.

- **If 60% of adults return, how many new birds are predicted?**



26

Regression Line

When a scatterplot displays a linear pattern, we can describe the overall pattern by drawing a straight line through the points. **Fitting a line** to data means drawing a line that comes as close as possible to the points.

Regression equation: $\hat{y} = b_0 + b_1 x$

- ***x*** is the value of the explanatory variable.
- **"*y*-hat"** is the predicted value of the response variable for a given value of *x*.
- b_1 is the **slope,** the amount by which *y* changes for each one-unit increase in *x*.
- b_0 is the **intercept,** the value of *y* when $x = 0$.

27

Least-Squares Regression Line

- The **least-squares regression line of *y* on *x*** is the line that minimizes the sum of the squares of the vertical distances of the data points from the line.
- If we have data on an explanatory variable *x* and a response variable *y*, the equation of the least-squares regression line is:

$$y = b_0 + b_1 x$$

where

$$b_1 = r\, s_y / s_x$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

28

Facts About Least-Squares Regression

Regression is one of the most common statistical settings, and least-squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

- **Fact 1:** A change of one standard deviation in *x* corresponds to a change of *r* standard deviations in *y*.
- **Fact 2:** The least-squares regression line always passes through (*x*-bar, *y*-bar)
- **Fact 3:** The distinction between explanatory and response variables is essential.

29

Correlation and Regression

Least-squares regression looks at the distances of the data points from the line only in the *y* direction. As a result, the variables *x* and *y* play different roles in regression. Even though correlation *r* ignores the distinction between *x* and *y,* there is a close connection between correlation and regression.

The **square of the correlation, r^2,** is the fraction of the variation in values of *y* that is explained by the least-squares regression of *y* on *x*.

- r^2 is called the **coefficient of determination.**

30

Using R

[cars data set: stopping distance vs. speed]

```
cars.lm <- lm(dist ~ speed, data = cars)

coef(cars.lm)

plot(dist ~ speed, data = cars, pch = 16)

abline(coef(cars.lm))

predict(cars.lm, newdata = data.frame(speed = c(6, 8, 21)))

residuals(cars.lm)

summary(cars.lm)
```

31

2.5 Cautions About Correlation and Regression

- Predictions
- Residuals and residual plots
- Outliers and influential observations
- Lurking variables
- Correlation and causation

32

Predictions via Regression Line

For the returning birds example, the least-squares regression line is:

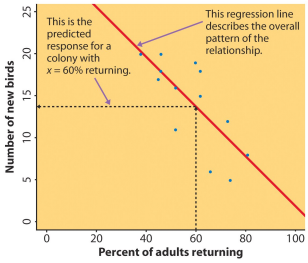
y*-hat = 31.9343 - 0.3040*x

y*-hat** is the predicted number of new birds for colonies with ***x percent of adults returning.

Suppose we know that an individual colony has 60% returning. What would we ***predict*** the number of new birds to be for just that colony?

For colonies with 60% returning, we ***predict*** the average number of new birds to be:

31.9343 - (0.3040)(60) = **13.69** birds



33

Residuals

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

residual = observed *y* – predicted *y*

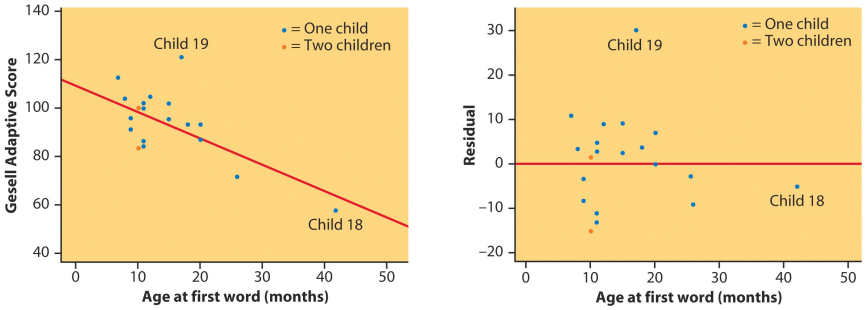
$$= y - \hat{y}$$

34

Residual Plots

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

- Ideally there should be a "random" scatter around zero.
- Residual *patterns* suggest deviations from a linear relationship.



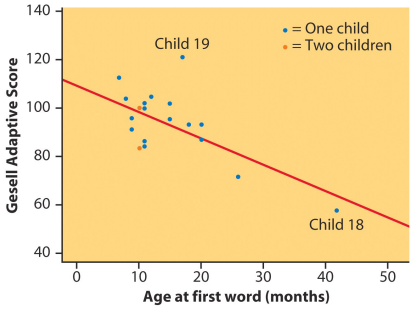
35

Outliers and Influential Points

An ***outlier*** is an observation that lies outside the overall pattern of the other observations.

Outliers in the *y* direction have large residuals.

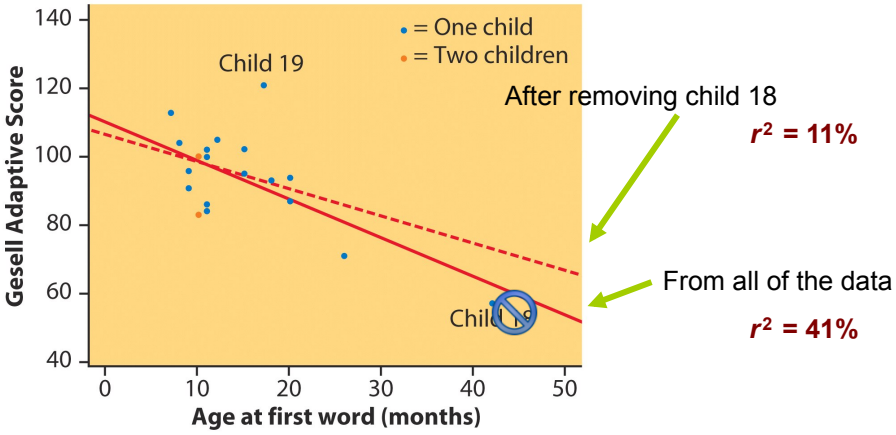
Outliers in the *x* direction are often ***influential*** for the least-squares regression line, meaning that the removal of such points would markedly change the equation of the line.



36

Outliers and Influential Points

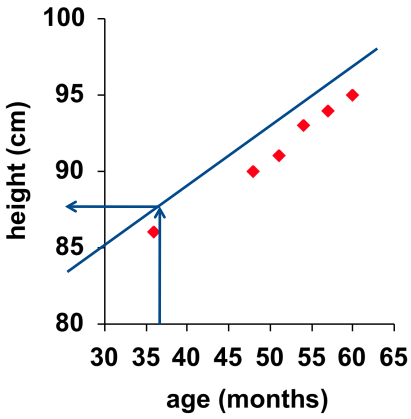
Gesell Adaptive Score and Age at First Word



37

Extrapolation

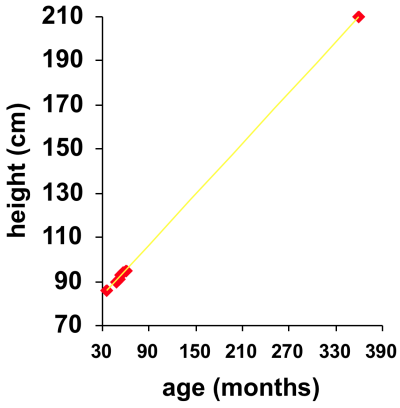
- Sarah's height was plotted against her age.
- Can you guess (predict) her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?



38

Extrapolation

- Regression line:
  y*-hat = 71.95 + .383 *x
- Height at age 42 months?
  ***y*-hat = 88**
- Height at age 30 years?
  ***y*-hat = 209.8**
- She is predicted to be 6' 10.5" at age 30! *What's wrong?*



39

Cautions About Correlation and Regression

- Both describe linear relationships.
- Both are affected by outliers.
- Always plot the data before interpreting.
- Beware of ***extrapolation.***
  - Use caution in predicting *y* when *x* is outside the range of observed *x*'s.
- Beware of ***lurking variables.***
  - These have an important effect on the relationship among the variables in a study, but are not included in the study.
- **Correlation does not imply causation!**

40

2.6 Data Analysis for Two-Way Tables

- The two-way table
- Joint distribution
- Conditional distributions
- Simpson's paradox

41

Categorical Variables

- Recall that categorical variables place individuals into one of several groups or categories.
- The values of a categorical variable are labels for the different categories.
- The distribution of a categorical variable lists the count or percent of individuals who fall into each category.
- When a dataset involves two categorical variables, we begin by examining the counts or percentage in various categories for *one* of the variables.

A **Two-way table** describes two categorical variables, organizing counts according to a **row variable** and a **column variable.** Each combination of values for these two variables is called a **cell.**

42

The Two-Way Table

Young adults by gender and chance of getting rich by age 30			
	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

What are the variables described by this two-way table?

How many young adults were surveyed?

43

Marginal Distribution

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among *all* individuals described by the table.

Note: Percentage is often more informative than counts, especially when comparing groups of different sizes.

To examine a marginal distribution:

1. Use the data in the table to calculate the marginal distribution (in percentage) of the row or column totals.

2. Make a graph to display the marginal distribution.

44

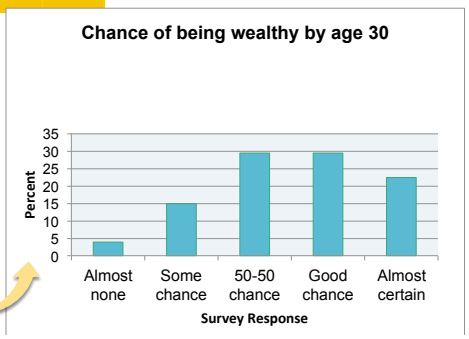
Marginal Distribution

Young adults by gender and chance of get

	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Examine the **marginal distribution** of chance of getting rich.

Response	Percent
Almost no chance	194/4826 = 4.0%
Some chance	712/4826 = 14.8%
A 50-50 chance	1416/4826 = 29.3%
A good chance	1421/4826 = 29.4%
Almost certain	1083/4826 22.4%



45

Conditional Distribution

Marginal distributions tell us nothing about the relationship between two variables. For that, we need to explore the conditional distributions of the variables.

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.

To examine or compare conditional distributions:

1. Select the row(s) or column(s) of interest.
2. Use the data in the table to calculate the conditional distribution (in percents) of the row(s) or column(s).
3. Make a graph to display the conditional distribution.
   - Use a **side-by-side bar graph** or **segmented bar graph** to compare distributions.

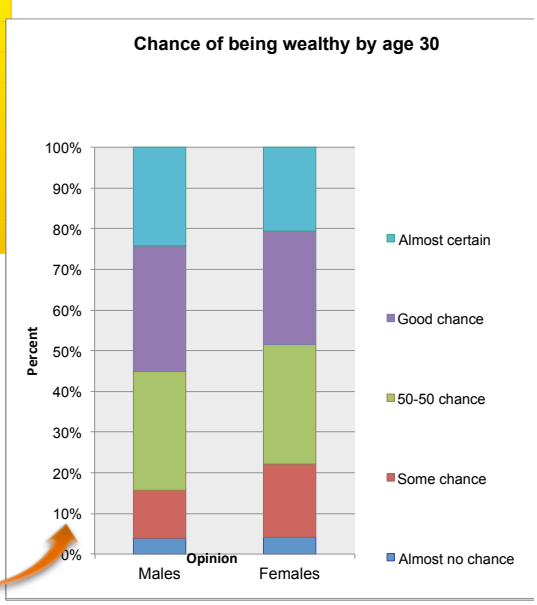
46

Conditional Distribution

Young adults by gender and chance of getting rich

	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Response	Male	Female
Almost no chance	98/2459 = 4.0%	96/2367 = 4.1%
Some chance	286/2459 = 11.6%	426/2367 = 18.0%
A 50-50 chance	720/2459 = 29.3%	696/2367 = 29.4%
A good chance	758/2459 = 30.8%	663/2367 = 28.0%
Almost certain	597/2459 = 24.3%	486/2367 = 20.5%



Simpson's Paradox

When studying the relationship between two variables, there may exist a **lurking variable** that creates a reversal in the direction of the relationship when the lurking variable is ignored as opposed to the direction of the relationship when the lurking variable is considered.

The lurking variable creates subgroups, and failure to take these subgroups into consideration can lead to misleading conclusions regarding the association between the two variables.

> An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox.**

48

Simpson's Paradox

Consider the acceptance rates for the following groups of men and women who applied to college.

Counts	Accepted	Not accepted	Total
Men	198	162	360
Women	88	112	200
Total	286	274	560

Percents	Accepted	Not accepted
Men	**55%**	45%
Women	44%	56%

A higher percentage of **men** were accepted: Is there evidence of discrimination?

49

Simpson's Paradox

Consider the acceptance rates when broken down by type of school.

BUSINESS SCHOOL

Counts	Accepted	Not accepted	Total
Men	18	102	120
Women	24	96	120
Total	42	198	240

Percents	Accepted	Not accepted
Men	15%	85%
Women	**20%**	80%

ART SCHOOL

Counts	Accepted	Not accepted	Total
Men	180	60	240
Women	64	16	80
Total	244	76	320

Percents	Accepted	Not accepted
Men	75%	25%
Women	80%	20%

50

Simpson's Paradox

- **Lurking variable:** Applications were split between the Business School (240) and the Art School (320).

  Within each school a higher percentage of women were accepted than men.

  No discrimination against women in this case!

  This is an example of **Simpsons Paradox.**

  When the lurking variable (Type of School: Business or Art) is ignored the data seem to suggest discrimination against women.

  However, when the type of school is considered, the association is reversed and suggests discrimination against men.

51

2.7 The Question of Causation

- Explaining association
- Causation
- Common response
- Confounding
- Establishing causation

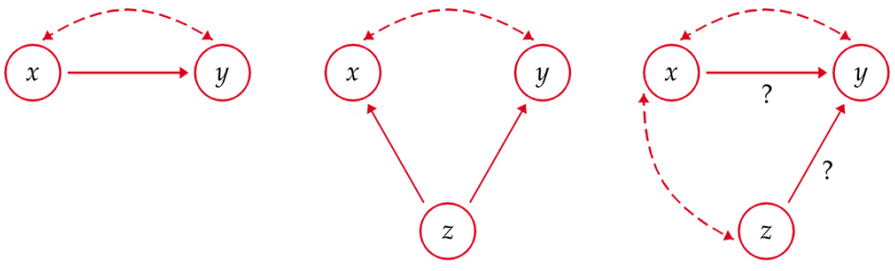
52

Explaining Association: Causation

Association, however strong, does NOT imply causation.

Some possible explanations for an observed association

The dashed lines show an association. The solid arrows show a cause-and-effect link. *x* is explanatory, *y* is response, and *z* is a lurking variable.



Causation Common response Confounding

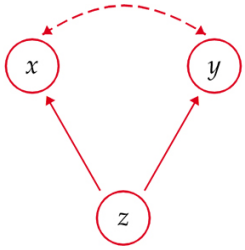
53

Explaining Association: Common Response

"Beware the lurking variable" is good advice when thinking about an association between two variables. The observed relationship between the variables can be explained by a lurking variable. Both *x* and *y* may change in response to changes in *z*.

Most students who have high SAT scores (*x*) in high school have high GPAs (*y*) in their first year of college.

- This positive correlation can be explained as a *common response* to students' ability and knowledge.
- The observed association between *x* and *y* could be explained by a third lurking variable *z*. In this example, "ability and knowledge" is the lurking variable.
- Both *x* and *y* change in response to changes in *z*. This creates an association even though there is no direct causal link.



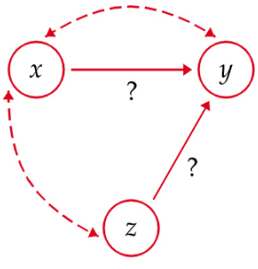
Common response

54

Explaining Association: Confounding

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explaterory variables or lurking variables.

- Example: Studies have found that religious people live longer than nonreligious people.
- Religious people also take better care of themselves and are less likely to smoke or be overweight.



Confounding

55

Establishing Causation

It appears that lung cancer is associated with smoking.

How do we know that both of these variables are not being affected by an unobserved third (lurking) variable?

For instance, what if there is a genetic predisposition that causes people to both get lung cancer *and* become addicted to smoking, but the smoking itself doesn't CAUSE lung cancer?

We can evaluate the association using the following criteria:

1. The association is strong.
2. The association is consistent.
3. Higher doses are associated with stronger responses.
4. Alleged cause precedes the effect.
5. The alleged cause is plausible.

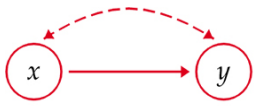
56

Evidence of Causation

A properly conducted ***experiment*** may establish causation.

Other considerations when we cannot do an experiment:

- The association is *strong.*
- The association is *consistent.*
  - The connection happens in *repeated trials.*
  - The connection happens under *varying conditions.*



- Higher doses are associated with stronger responses.
- Alleged cause *precedes* the effect in time.
- Alleged cause is *plausible* (reasonable explanation).

57