STOR 155 Introductory Statistics

Lecture 7:
Scatterplots, Correlation
Relationships between Quantitative Variables

• Chapter 1 talks about distribution of one variable.
• Often we are interested in the relationships between two quantitative variables.
• Examples:
  – Heights of parents and children
  – High school GPA and college GPA
  – Stock returns of two different corporations
Chapter 2: Looking at Data – Relationships

- Association (statistical dependence)
- Response and explanatory variables
- Scatterplots
- Correlation
• Two variables are associated (dependent) if some values of one variable tend to occur more often along with some values of the second variable.
POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

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

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

Definition, pg 106
*Introduction to the Practice of Statistics, Fifth Edition*
© 2005 W. H. Freeman and Company
Examples

- weight (in kilogram) and height (in centimeter)
- An insurance company reports that heavier cars have less fatal accidents per 10,000 vehicles than lighter cars do.
- A medical study finds that short women are more likely to have heart attacks than women of average height, while tall women have even fewer heart attacks.

Note: no explanation yet, just findings …
Response and Explanatory Variables

• A *response variable* (or *dependent variable*) measures an outcome of a study.

• An *explanatory variable* (or *independent variable*) explains or causes changes in the *response variable*.

• In most cases, we set values of one variable to see how it affects another variable.
  – biological, chemical experiments

• Not always causal relation!
  – SAT scores vs college grades
Scatterplots

• Two-dimensional plot, with one variable’s values plotted along the vertical axis and the other along the horizontal axis.
  – X-axis: explanatory
  – Y-axis: response

• Display the general relationship between two quantitative variables graphically.

• Two variables measured on the same “individuals”.
Example

• A statistician wanted to purchase a house in a neighborhood. He decided to develop a model to predict the selling price of a house.

• He took a random sample of 100 houses that recently sold and recorded the selling price, the number of bedrooms, and the size (in square foot) for each.
Bivariate Fit of Price By H Size

The graph shows a scatter plot of house price versus house size. The data points are plotted on a grid with price on the y-axis and house size on the x-axis. The scatter plot suggests a trend where larger houses tend to be more expensive.
Examining a Scatterplot

• **Overall Pattern**
  – Direction: positive or negative association
  – Form: linear or nonlinear (eg. curved or clustered)
  – Strength: strong if there is very little deviation from the trend

• **Deviation**
  – Deviation in *form* or *direction*
  – Outliers
Typical Patterns of Scatterplots

- **Positive linear relationship**
- **No relationship**
- **Negative linear relationship**
- **Negative nonlinear relationship**
  - This is a weak linear relationship. A non-linear relationship seems to fit the data better.
- **Nonlinear (concave) relationship**
Figure 2-1
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company
Categorical variable in a scatterplot

Figure 2-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company
Relationship between Categorical and Quantitative Variables

- Back-to-back stemplots: two categories
- Side-by-side boxplots: any number of categories
Review: Scatterplot

• The plot shows relationship between two quantitative variables.
• It plots observations of different individuals in a two-dimensional graph.
• Each point in a scatterplot corresponds to two variables of the same individual.
• Just **graphical**, not **numerical**.
Which one shows a stronger relationship?
Correlation

- A quantity used to measure the direction and strength of the linear relationship between two quantitative variables.
- Often written as $r$

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

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data are summarized as follows.

\[ n = 100, \bar{x} = 36,009.5, \bar{y} = 5411.4 \]

\[ s_x = 6597.6, s_y = 254.9, \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = -1,356,256. \]

- Find the correlation
- \( r = -0.806 \)
Properties of $r$

$-1 \leq r \leq 1$

- Positive $r$ indicates positive linear association.
- Negative $r$ indicates negative linear association.
- The closer that $r$ moves toward 1 or $-1$, the stronger the linear association is.
- $r = -1$ or 1 occurs only when the points in the scatterplot lie along a straight line.
Properties of $r$

- It measures only the **linear** relationship between two quantitative variables.
- Invariant to the order of the variables
- Invariant to rescaling (why?)
- Unit-free
- Sensitive to outliers
- Question: True or False?
  
  $r = 0$ means there is no association between two variables.
Different correlations

Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$
Blunders about Correlation (why ?)

• There is a high correlation between the gender of American workers and their income.

• We found a high (r = 1.09) between students’ ratings of faculty teaching and ratings made by other faculty members.

• The correlation between planting rate and yield of corn was found to be r = 0.23 bushel.
Take Home Message

- Association (dependence)
- Scatterplot
- Correlation
- Properties of Correlation