STATISTICS 201

Outline for today:

- Go over syllabus
- Provide requested information on survey (handed out in class)
- Brief introduction and hands-on activity

Survey: Provide this Info

- Name
- Major/Program
- Year in school or in graduate program
- Something interesting about yourself
- Why you are taking this class

How familiar are you with these?

(Not at all, somewhat, very, could teach it)

1. Summation notation
2. Hypothesis testing in general
3. P-values
4. Confidence intervals
5. Two-sample *t*-test
6. Sampling distributions
7. F-Distribution
8. Scatter plots
9. Simple linear regression
10. Matrices

Survey, continued

Provide the following data:

a. Your height, in *inches* (to nearest half inch) or *centimters* (note 1 cm = 0.3937 inches)
b. Your "handspan" in *centimeters*, defined as the distance covered on the ruler by your stretched hand from the tip of the thumb to the tip of the small finger.
c. Predicted handspan = −3 + 0.35 (ht in inches) or −3 + 0.1378 (ht in centimeters)
d. Your "residual" (to be explained!)

Regression and ANOVA

- Used to describe the relationship between a continuous "response" variable and one or more "predictor" variables (continuous = regression; categorical = ANOVA).
- Regression used to predict a future response using known, current values of the predictors, or estimate relationship.
- ANOVA used to figure out why means differ for different groups, treatments, etc.
- First need to discuss how data collection method affects potential conclusions – very important!
- Switch to power point slides modified from Brooks/Cole to accompany "Mind On Statistics" by Utts/Heckard

IMPORTANT NOTE

The remaining slides are modified from Power point presentations to accompany *Mind on Statistics,* by Utts and Heckard and are copyright Brooks/Cole. They are not to be copied or used for purposes other than this class.

Gathering Useful Data

See Section 1.4 of our 201 textbook

Principle Idea:

The knowledge of how the data were generated is one of the key ingredients for translating data intelligently.

Description or Decision? Using Data Wisely

- **Descriptive Statistics:** using numerical and graphical summaries to characterize a data set (and *only* that data set).
- **Inferential Statistics:** using sample information to make conclusions about a *broader range* of individuals than just those observed.

Two Important Issues Based on Data Collection Method

- **Extending results to a population:** This can be done if the *data are representative of a larger population* ***for the question of interest.*** Safest to use a ***random sample***.
- **Cause and effect conclusion:** Can *only* be made if data are from a ***randomized experiment,*** **not from an** ***observational study.***

10

Definitions of Types of Studies

Observational Study:

Researchers *observe* or *question* participants about opinions, behaviors, or outcomes. Participants not asked to do anything differently.

Two special cases:

sample surveys and *case-control studies*.

Experiment:

Researchers *manipulate* something and *measure the effect* of the manipulation on some outcome of interest.

Randomized experiments: participants are *randomly assigned* to participate in one condition (called *treatment*) or another.

Sometimes cannot conduct experiment due to practical/ethical issues.

NOT the same thing as random sampling.

Types of Variables (Measured or Not)

Explanatory variable (or independent variable) is one that may explain or may cause differences in a **response variable** (or outcome or dependent variable).

A **confounding variable** is a variable that *affects the response variable* and also is *related to the explanatory variable*. A potential confounding variable not measured in the study is called a **lurking variable**.

Obs. Study: *Lead Exposure and Bad Teeth*

"Children exposed to lead are more likely to suffer tooth decay …"
USA Today

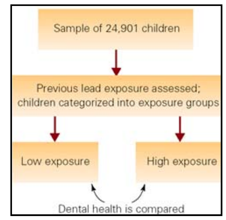
Observational study
involving 24,901 children.

Explanatory variable =
level of lead exposure.

Response variable = extent child
has missing/decayed teeth.

Possible confounding variables =
income level, diet,
time since last dental visit.

Lurking variables = amount of
fluoride in water, health care



CRUCIAL POINT

This study is an observational study.

We cannot conclude that lead exposure *causes* tooth decay.

It would be unethical to do a randomized experiment, so we need other (non-statistical) ways to establish cause and effect.

Randomized Experiment: *Quitting Smoking with Nicotine Patches*

"After the eight-week period of patch use, almost half (46%) of the nicotine group had quit smoking, while only one-fifth (20%) of the placebo group had." *Newsweek, March 9, 1993, p. 62*

Double-blind, Placebo-controlled Randomized Experiment
240 smokers recruited (volunteers)

Randomized to 22-mg nicotine patch or placebo (**controlled**) patch for 8 weeks.

Double-blind: neither the participants nor the nurses taking the measurements knew who had received the active nicotine patches.

CRUCIAL POINT

This study is a randomized experiment.

We *can* conclude that nicotine patches *cause* people to quit smoking.

Potential confounding variables should be similar in the placebo and nicotine patch groups because of random assignment.

Relationships Between Quantitative Variables

Three Tools we will use …

- **Scatterplot**, a two-dimensional graph of data values
- **Correlation**, a statistic that measures the *strength* and *direction* of a linear relationship
- **Regression equation**, an equation that describes the average relationship between a response and explanatory variable

Looking for Patterns with Scatterplots

Questions to Ask about a Scatterplot

- What is the *average* pattern? Does it look like a straight line or is it curved?
- What is the direction of the pattern?
- How much do individual points vary from the average pattern?
- Are there any unusual data points?

Positive/Negative Association

- Two variables have a **positive association** when the values of one variable tend to increase as the values of the other variable increase.
- Two variables have a **negative association** when the values of one variable tend to decrease as the values of the other variable increase.

Example: *Height and Handspan*

Data:

Height (in.)	Span (cm)
71	23.5
69	22.0
66	18.5
64	20.5
71	21.0
72	24.0
67	19.5
65	20.5
76	24.5
67	20.0
70	23.0
62	17.0

and so on,
for $n = 167$ observations.

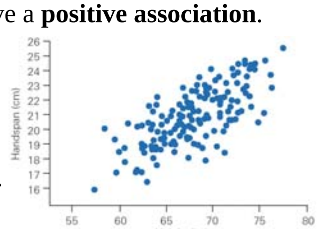
Data shown are the first 12 observations of a data set that includes the heights (in inches) and fully stretched handspans (in centimeters) of 167 college students.

Example, cont. *Height and Handspan*

Taller people tend to have greater handspan measurements than shorter people do.

When two variables tend to increase together, we say that they have a **positive association**.

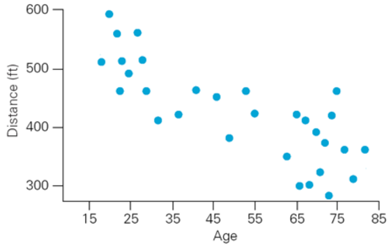
The handspan and height measurements may have a **linear relationship**.



Example: *Driver Age and Maximum Legibility Distance of Highway Signs*

- A research firm determined the **maximum distance** at which each of 30 drivers could read a newly designed sign.
- The 30 participants in the study ranged in **age** from 18 to 82 years old.
- We want to examine the **relationship** between age and the sign legibility distance.

Example *Driver Age and Maximum Legibility Distance of Highway Signs*

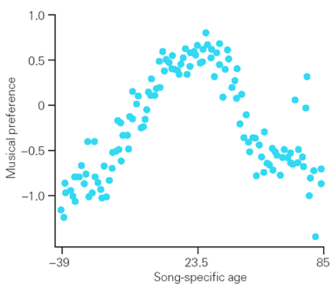


- We see a **negative** association with a **linear** pattern.
- We will use a **straight-line equation** to model this relationship.

Example: *The Development of Musical Preferences*

- The 108 participants in the study ranged in age from 16 to 86 years old.
- We want to examine the **relationship** between **song-specific age** (age in the year the song was popular) and **musical preference** (positive score => above average, negative score => below average).
- Note that a *negative* "song-specific age" means the person wasn't born yet when the song was popular.

Example: *The Development of Musical Preferences*



- Popular music preferences acquired in late adolescence and early adulthood.
- The association is **nonlinear**.

Describing Linear Patterns with a Regression Line

When the best equation for describing the relationship between *x* and *y* is a *straight line*, the equation is called the **regression line**.

Two purposes of the regression line:

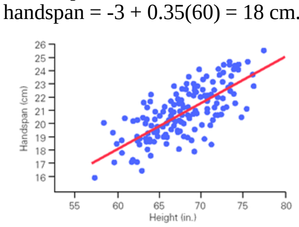
- to **estimate the average** value of *y* at any specified value of *x*
- to **predict the value** of *y* for an **individual**, given that individual's *x* value

Example: *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Estimate the average handspan for people 60 inches tall:
Average handspan = -3 + 0.35(60) = 18 cm.

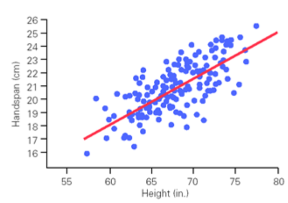
Predict the handspan for someone who is 60 inches tall:
Predicted handspan = -3 + 0.35(60) = 18 cm.



Example: *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Slope = 0.35 => Handspan increases by 0.35 cm, **on average**, for each increase of 1 inch in height.



In a **statistical relationship**, there is variation from the average pattern.

The Equation for the Regression Line (for a sample, not a population)

$$\hat{y} = b_0 + b_1 x$$

$\hat{y}$ is spoken as "**y-hat**," and it is also referred to either as predicted y or estimated y.

b_0 is the **intercept** of the straight line. The intercept is the value of y when $x = 0$.

b_1 is the **slope** of the straight line. The slope tells us how much of an increase (or decrease) there is for the y variable when the x variable increases by one unit. The sign of the slope tells us whether y increases or decreases when x increases.

Prediction Errors and Residuals

- **Prediction Error** = difference between the observed value of y and the predicted value $\hat{y}$.
- **Residual** $= (y - \hat{y})$

Let's predict your handspan Record these on your survey

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

Handspan (cm) = -3 + 0.35 Height (inches)

Or -3 + 0.1378 Height (cm)

Calculate your predicted handspan:

Examples: -3 + (0.35)(70 inches) = 21.5 cm

-3 + (0.35)(65 inches) = 19.75 cm

-3 + (0.1378) (165 cm) = 19.73 cm

Find your residual:

(actual handspan – predicted handspan)

33

Measuring Strength and Direction with Correlation

Correlation *r* *indicates the **strength** and the **direction** of a straight-line relationship.*

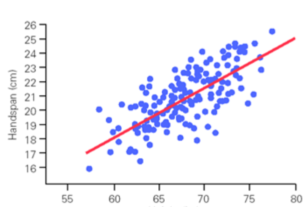
- The ***strength*** of the relationship is determined by the *closeness of the points to a straight line.*
- The ***direction*** is determined by whether one variable generally increases or generally decreases when the other variable increases.

Example: *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Correlation *r* = +0.74 =>

a somewhat **strong positive linear** relationship.

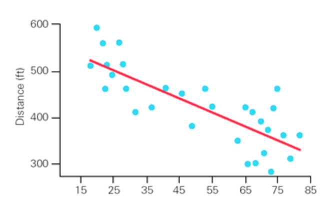


Example: *Driver Age and Maximum Legibility Distance of Highway Signs (cont)*

Regression equation: Distance = 577 - 3 Age

Correlation *r* = -0.8 =>

a **somewhat strong negative linear** association.

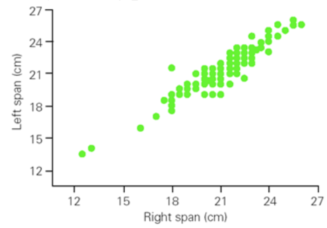


Example: *Left and Right Handspans*

If you know the span of a person's right hand, can you accurately predict his/her left handspan?

Correlation *r* = +0.95 =>

a **very strong positive linear** relationship.

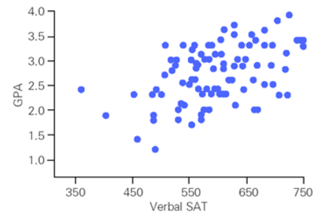


Example: *Verbal SAT and GPA*

Grade point averages (GPAs) and verbal SAT scores for a sample of 100 university students.

Correlation *r* = 0.485 =>

a **moderately strong positive linear** relationship.

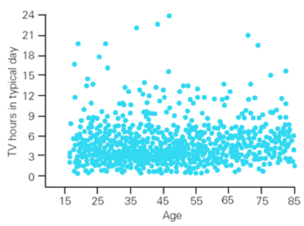


Example: *Age and Hours of TV Viewing*

Relationship between age and hours of daily television viewing for 1913 survey respondents.

Correlation *r* = 0.12 => a **weak connection**.

Note: a few claimed to watch more than 20 hours/day!

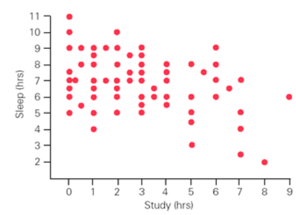


Example: *Hours of Sleep and Hours of Study*

Relationship between reported hours of sleep the previous 24 hours and the reported hours of study during the same period for a sample of 116 college students.

Correlation *r* = –0.36 => a **not too strong negative association**.

(More study, less sleep)



Summary

Regression is used to do two things:

- **Predict** future values using information available now. (Predict response from explanatory variable.)
- **Estimate** the average relationship between a response and one or more explanatory variables.
- Regression only works for *linear* relationships.

41

Homework

Problems 1.12, 1.13, 1.29 and 1.30

Due next Wed (October 9) by 6pm