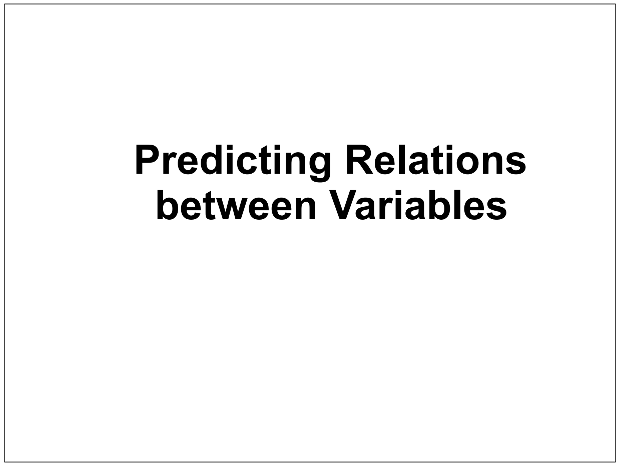
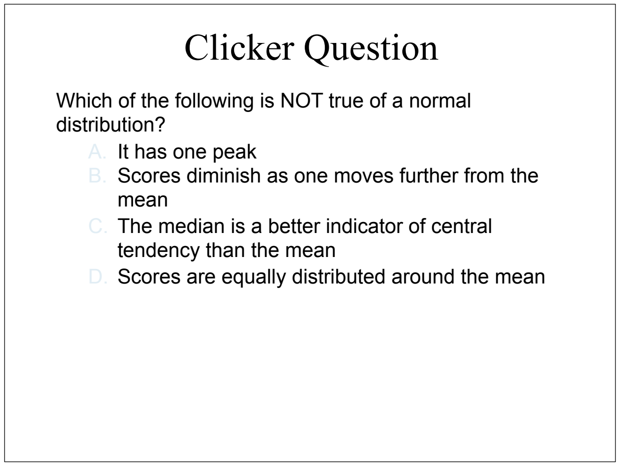


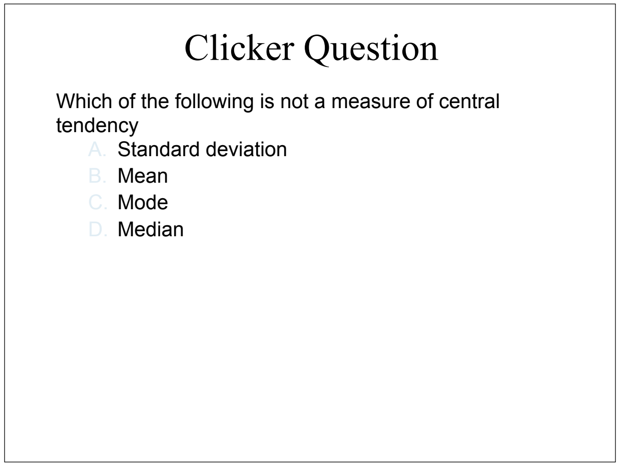
Predicting Relations between Variables



Clicker Question

Which of the following is NOT true of a normal distribution?

A. It has one peak
B. Scores diminish as one moves further from the mean
C. The median is a better indicator of central tendency than the mean
D. Scores are equally distributed around the mean



Clicker Question

Which of the following is not a measure of central tendency

A. Standard deviation
B. Mean
C. Mode
D. Median

Clicker Question

Why is it important to determine the standard deviation of a sample?

A. The standard deviation of the sample is the same as the standard deviation

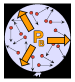
B. The stardard deviation specifies how reliable an estimate we can make about the mean in the population

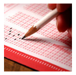
C. The standard deviation tells us whether the sample was fair or biased

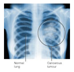
D. The standard deviation tells us whether the mean or the median is the best measure of the central tendency in the population

Hypotheses involving more than one variable

Many of the hypotheses of interest in science and in ordinary life involve relations between variables







- Amount of sleep and ability to recall information
- Pressure, volume, and temperature of a gas
- Experience and job performance
- SAT score and grades in college
- Vitamin intake and health condition
- Sexual activity and sexually transmitted diseases
- Smoking and lung cancer
- Miles per gallon and horsepower of cars

The Case Against Bread

- More than 98% of convicted felons are bread eaters.
- Fully half of all children who grow up in bread consuming households score below average on standardized tests.
- In the 18th century, when virtually all bread was baked in the home, the average life expectancy was less than 50 years.
- More than 90% of all violent crimes are committed within 24 hours of eating bread.
- Primitive tribal societies that have no bread exhibit a low incidence of cancer, Alzheimer's, and Parkinson's disease.
- Ask yourself: are the statistics meaningful!

Correlations and why they are interesting

- A correlational claim is a claim that the values on two variables vary systematically
  - Not necessarily in the same direction
- It is not a causal claim, although
  - correlations may be the result of causation
  - and correlations may be employed in establishing causal claims
- Why care about correlations if they are not (known to be) causal?
  - They can be used to make predictions about the unknown value of one variable from the known value of another variable

SAT and College Grades

- Should the SAT be used as a (or maybe the) basis for admission to the University of California?
- If so, then it must be justified
  - Does it predict success in college?
  - If it doesn't, then it may be an inappropriate measure to use in judging admissions
- Compare: basing admissions to UC on
  - Running speed for the mile
  - Length of one's index finger

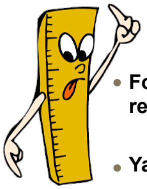
From the general to the testable

- Not all hypotheses relating variables are directly testable—hypotheses presented in general terms
  - Fitter people live longer
  - Better education correlates with greater happiness
  - Greater pollution correlates with greater global warming
- To test the correlation, one must link the general terms to specific, measurable variables

Operational "definitions"

- Relate the variables used in the hypothesis to measurable variables
- Variables such as force, memory ability, happiness, brain injury, etc., are not directly measurable (observable).
  - Must specify a measurement procedure and a variable we can measure
- The operational definitions of any non-observational terms are major *auxiliary assumptions* in any test of a hypothesis

Distance



- Inch: width of a grown man's thumb
  - King Edward II (14th C.): the length of an inch shall be equal to three grains of barley, dry and round, placed end to end lengthwise
- Foot: the name gives away its original reference
  - Standardized to 12 inches
- Yard: the length of a person's belt
  - King Henry I (13th C.): distance from his nose to the thumb of his outstretched arm, which was about 36 inches

Construct Validity

- Does the way you operationalize a variable really measure that variable?
  - Does a ruler (do grains of barley) really measure height?
  - Does an intelligence test measure intelligence?
  - Does a word-list test measure memory?
- The degree to which a measure measures what it is supposed to measure is referred to as its *construct validity*

Clicker Question

An operational definition

- Aims to provide necessary and sufficient conditions for the variable being being measured
- Employs operations to determine what something is
- Relates a variable used in a hypothesis to a way to detect and measure it
- Provides sufficient, but not necessary conditions for the variable being measured

Clicker Question

Construct validity is concerned with

- Whether the argument for the construct is valid
- Whether the operational definition really measures the variable used in the hypothesis
- Is only important if there is doubt about how to assign values to variables
- Replacing operational definitions with real definitions

Operational definitions are not definitions

An operational definition provides one way to measure a variable

- **There will typically be alternatives**
- **The alternatives may not always agree**

Even when construct validity is high, the operational definition does not provide necessary and sufficient conditions for the term

Two Types of Correlational Study

- When items have values on two score variables, correlate the scores on one with the scores on the other
  - Measure degree of correlation in terms of Pearson coefficient *r*
  - Predict value on one variable from that on the other using the regression line: y=ax+b
- When one nominal variable divides a population into two or more sub-populations, compare the two (or more) populations on another (score) variable in terms of their central tendencies
  - If the means are different, predict the value on the score variable depending on the value of the nominal variable

Relating Score Variables

- Same items measured on two score variables
- Is there any systematic relation between the score on one variable and the score on another?

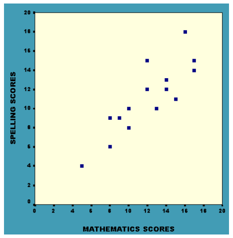
Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spelling	15	14	15	12	6	4	8	9	9	12	18	13	10	10	11
Math	12	17	17	12	8	5	10	9	8	14	16	14	10	13	15

Often it is difficult to determine if there is a regular pattern by just looking at scores (eyeballing the data)

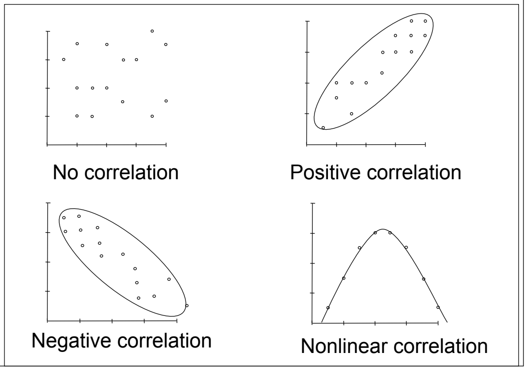
Important to graph or diagram the data

Scatterplots

Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spelling	15	14	15	12	6	4	8	9	9	12	18	13	10	10	11
Math	12	17	17	12	8	5	10	9	8	14	16	14	10	13	15



Scatterplots - 2



No correlation

Positive correlation

Negative correlation

Nonlinear correlation

Measuring correlation

Karl Pearson developed a measure of correlation, known as *Pearson's Product Moment Correlation (r)*

-1.0____________________0____________________1.0

Perfect negative   No Correlation   Perfect Positive

A Z score for an individual is how many standard deviations that individual is from the mean. From that there is an easy calculation of Pearson's r:

$$r = \sum(ZxZy) / N$$

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{N})\ (\Sigma Y^2 - \frac{(\Sigma Y)^2}{N})}}$$

Pearson Correlation Coefficient

Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Spelling	15	14	15	12	6	4	8	9	9	12	18	13	10	10	11
Math	12	17	17	12	8	5	10	9	8	14	16	14	10	13	15

- Pearson's Product Moment Correlation r = .857
  - Note: Positive Value—positively correlated
  - Value close to 1—strongly or highly correlated
- Strong positive correlation

Clicker Question

A Pearson correlation of 4.25 between height and salary

Represents a very strong positive correlation

Means that height is a very good predictor of salary

Means that height is a poor predictor of salary

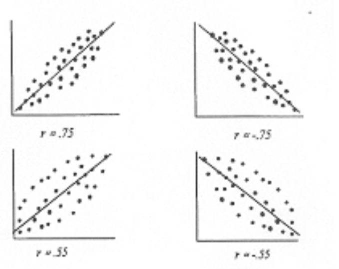
Makes no sense

How much does the correlation account for?

- Correlations are typically not perfect (r=1 or r=-1)
  - Evaluate the correlation in terms of how much of the variance in one variable is accounted for by the variance in another [variance=$\sum$ (X-mean)^2/N]
- Amount of variance accounted for (on the variable whose value is being predicted) equals:
  - Variance explained/total variance
- This turns out to be the square of the Pearson coefficient: r^2

Variance Accounted for

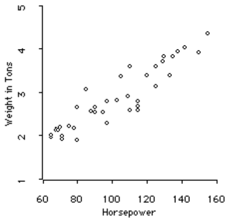
- r^2 = .56



- r^2 = .30

Variance accounted for - 2

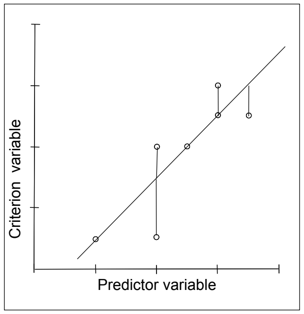
- Correlating automobile horsepower and weight
  - r = .92
  - r^2 = .81
- Horsepower accounts for 81% of the variance in car weight
  - Given only the horsepower of a car, you can make a quite reliable estimate of the car's weight



Prediction

- A major reason to be interested in correlation
  - If two variables are correlated, we can use the value of an item on one variable to predict the value on another
    - Employment prediction: future job performance based on years of experience
    - Actuarial prediction: how long one will live based on how often one skydives
    - Risk assessment: prediction of how much risk an activity poses in terms of its values on other variables
- Prediction employs the regression line

Regression line



- Prediction is based on the regression line
- Start with scatter plot of data points
- Find line which allows for the best prediction of the criterion variable (one to be predicted) from that of the predictor variable
- Line which minimizes the (square of the) distances of the blue lines

Regression line

- **y = a + bx**
- **y = predicted or criterion variable**
- **x = predictor variable**
- **a = y-intercept—regression constant**
- **b = slope—regression coefficient**
- **Note: the regression coefficient is not the same as the Pearson coefficient r**

Clicker Question

If the Pearson coefficient (r) between age and liking for chocolate is -.62, what can you infer about the slope of the regression line?

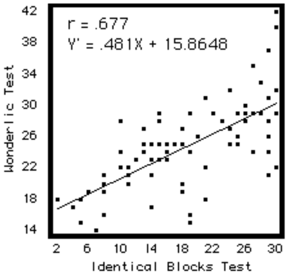
A. Nothing
B. The slope is also -.62
C. The slope will be .62
D. The slope will be negative

Understanding the Regression Line

- **Assume the regression line equation between the variables mpg (y) and weight (x) of several car models is**
  - **mpg = 62.85 - 0.011 weight**
  - **MPG is expected to *decrease* by 1.1 mpg for every additional 100 lb. in car weight**
  - **The regression constant, 62.85, represents the projected value of a car weighing 0 lbs.**

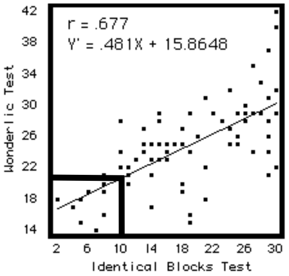
Interpolating from the regression line

- Correlation between
  - Identical Blocks Test (a measure of spatial ability)
  - Wonderlic Test (a measure of general intelligence)
- Calculate new value for x = 10:
- y = .48 x 10 + 15.86
  = 20.67



Interpolating from the regression line visually

- Draw line from the x-axis to the regression line
- Draw line from the intersection with the regression line to the y-axis



Clicker Question

You are told that the regression line relating a reasoning test score and a memory test score is

reasoning score = -3.25 + .7 memory score

You know that

A. There is a positive correlation between the scores
B. There is a negative correlation between the scores
C. Pearson’s r = .7
D. Pearson’s r = -3.25