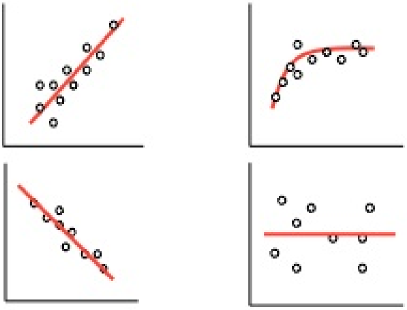
PSY 501: Review of statistics (part 1 – descriptives)

Week 11

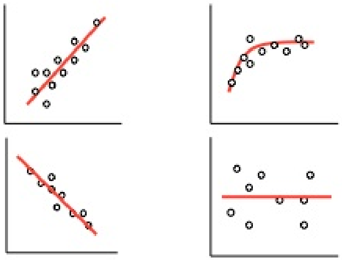
Statistics: Why do we use them?

- Descriptive statistics (this week)
  - Used to describe, simplify, and organize data sets
  - Describing *distributions* of scores
- Inferential statistics (next week)
  - Used to test claims about the population, based on data gathered from samples
  - Takes sampling error into account
  - "Are the results above and beyond what you would expect from chance?'

Correlation

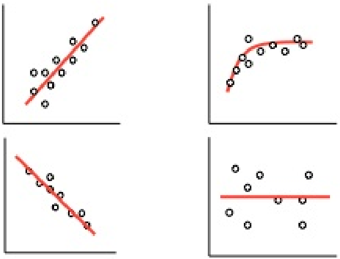


Correlation



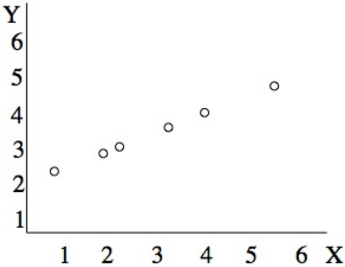
- Properties of a correlation
  - Form (linear vs. nonlinear)
  - Direction (positive vs. negative)
  - Strength (none, weak, strong, perfect)
- To examine this relationship, you should:
  - Make a scatterplot
  - Compute the correlation coefficient

Correlation

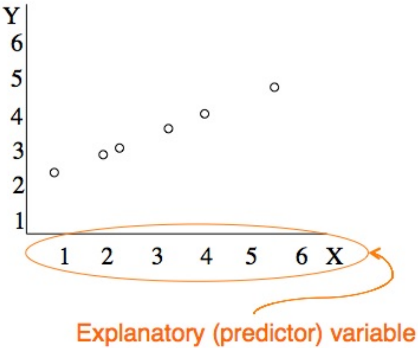


- Correlation coefficient
  - a numerical description of the relationship between two variables, ranges between -1 and 1, with 0 = no relationship
  - Pearson's *r*: describes relationship between two continuous variables
  - "As X goes up, what happens to Y?"

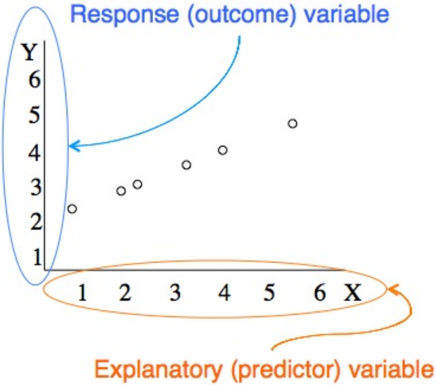
Regression: Making predictions



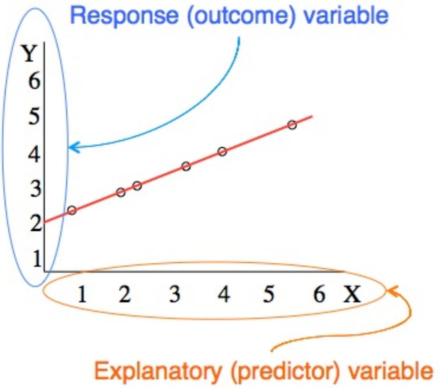
Regression: Making predictions



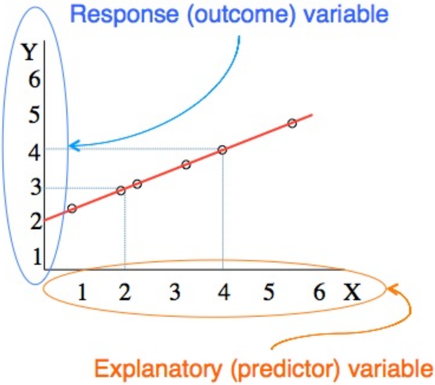
Regression: Making predictions



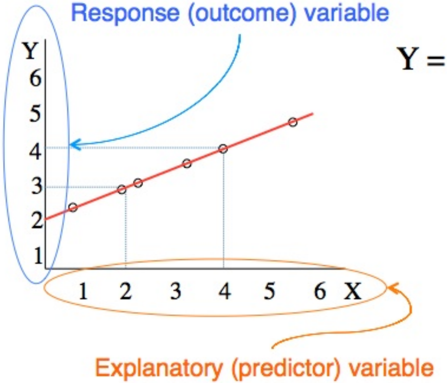
Regression: Making predictions



Regression: Making predictions



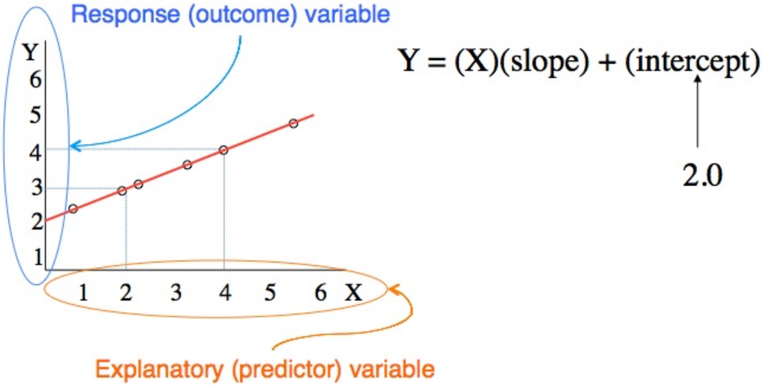
Regression: Making predictions


Response (outcome) variable

Explanatory (predictor) variable

$$Y = (X)(\text{slope}) + (\text{intercept})$$

Regression: Making predictions



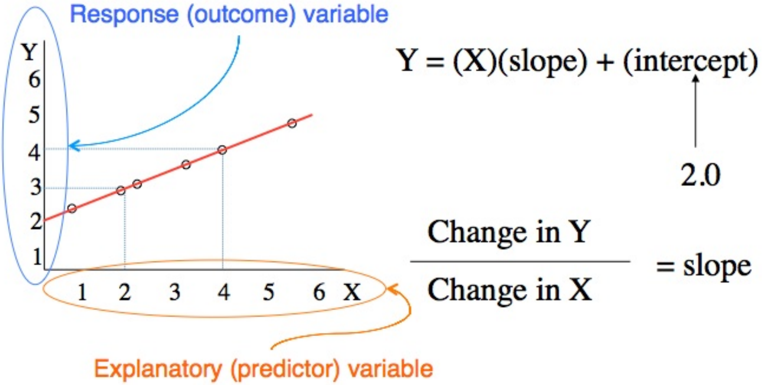
Response (outcome) variable

$$Y = (X)(\text{slope}) + (\text{intercept})$$

2.0

Explanatory (predictor) variable

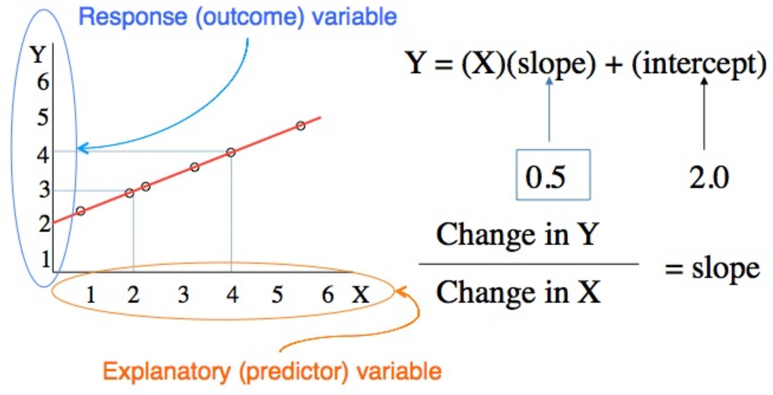
Regression: Making predictions



$$Y = (X)(\text{slope}) + \underset{\overset{\uparrow}{2.0}}{(\text{intercept})}$$

$$\frac{\text{Change in } Y}{\text{Change in } X} = \text{slope}$$

Regression: Making predictions

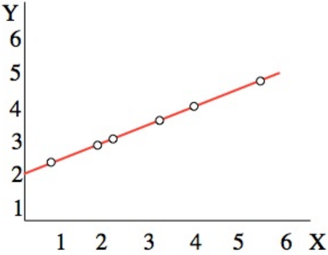


$$Y = (X)(\underset{\boxed{0.5}}{\uparrow\text{slope}}) + (\underset{2.0}{\uparrow\text{intercept}})$$

$$\frac{\text{Change in Y}}{\text{Change in X}} = \text{slope}$$

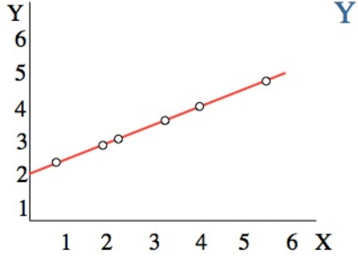
Regression: Making predictions

Can make specific predictions about Y based on X



Regression: Making predictions

Can make specific predictions about Y based on X

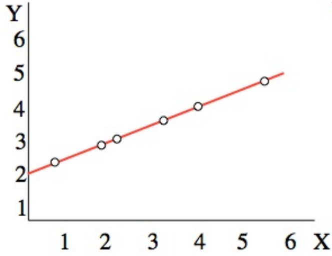


X = 5

Y = ?

Regression: Making predictions

Can make specific predictions about Y based on X



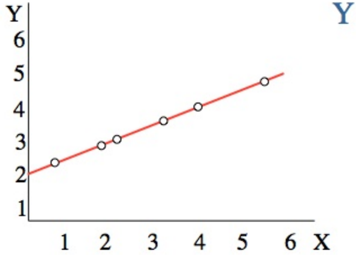
X = 5

Y = ?

$$Y = (X)(.5) + (2.0)$$

Regression: Making predictions

Can make specific predictions about Y based on X



$X = 5$

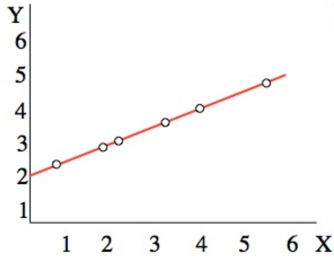
$Y = ?$

$Y = (X)(.5) + (2.0)$

$Y = (5)(.5) + (2.0)$

Regression: Making predictions

Can make specific predictions about Y based on X



X = 5

Y = ?

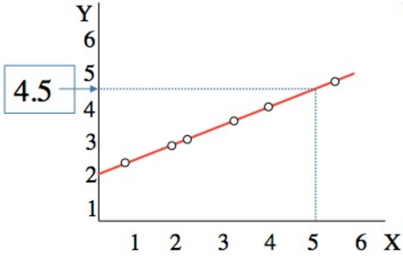
$$Y = (X)(.5) + (2.0)$$

$$Y = (5)(.5) + (2.0)$$

$$Y = 2.5 + 2 = 4.5$$

Regression: Making predictions

Can make specific predictions about Y based on X



$X = 5$

$Y = ?$

$Y = (X)(.5) + (2.0)$

$Y = (5)(.5) + (2.0)$

$Y = 2.5 + 2 = 4.5$

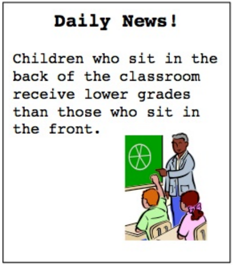
Cautions with correlation and regression

- Don't extrapolate
- Extreme scores (outliers) can strongly influence the calculated relationship
- Don't make causal claims
  - Be careful of misinterpretation

Example: Misunderstood correlational design

Suppose you notice that kids who sit in the front of the class typically get higher grades

- This suggests there is a relationship between where you sit in class and grades.



Daily News!

Children who sit in the back of the classroom receive lower grades than those who sit in the front.

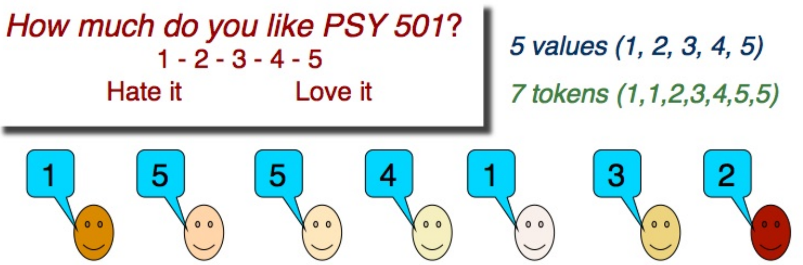
- Possibly implied: "[All] Children who sit in the back of the classroom [always] receive worse grades than [each and every child] who sits in the front."
- Better: "Researchers found that children who sat in the back of the classroom were more likely to receive lower grades than those who sat in the front."

Statistics: Why do we use them?

- Descriptive statistics
  - Used to describe, simplify, and organize data sets
  - Describing *distributions* of scores
    - Graphic and tabular descriptions
    - Numeric descriptions

Distributions

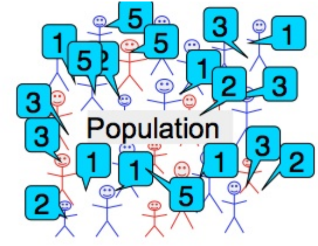
- Recall that a variable is a characteristic that can take different values
- The distribution of a variable is a summary of all the different (observed) values of a variable
  - Both *type* (each value) and *token* (counts of each instance)



Distributions

Many important distributions

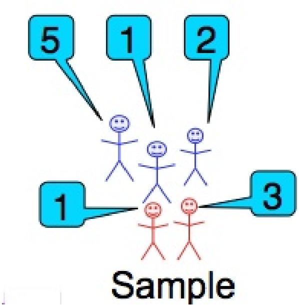
- Population
  - All the scores of interest



Distributions

Many important distributions

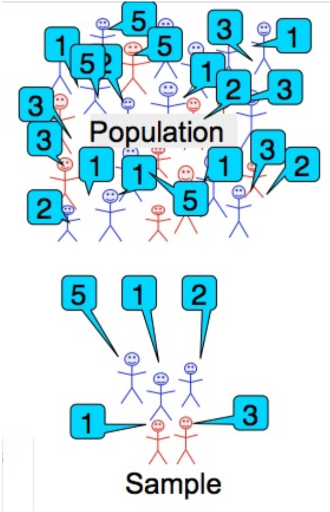
- Population
  - All the scores of interest
- Sample
  - All of the scores observed (your data)
  - Used to estimate population characteristics



Distributions

Many important distributions

- Population
  - All the scores of interest
- Sample
  - All of the scores observed (your data)
  - Used to estimate population characteristics
- Distribution of sample distributions
  - Used to estimate sampling error



How do we describe these distributions?

Describing distributions

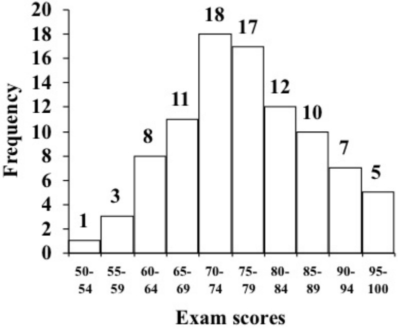
Focus on 3 properties of distributions

- Shape
  - Symmetric vs. asymmetric (skew)
  - Unimodal vs. multimodal
- Center
  - Where most of the data in the distribution are located
    - Mean, median, mode
- Spread (variability)
  - How similar/dissimilar are the scores in the distribution?
    - Standard deviation (variance), range

Graphs for continuous variables

Frequency histogram

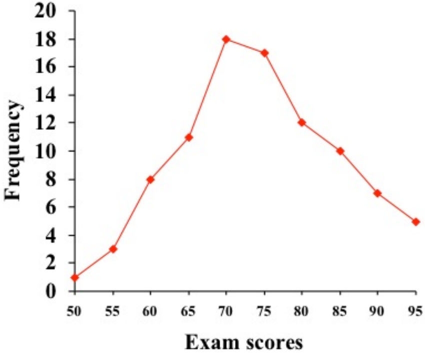
- Example: distribution of scores on an exam



Graphs for continuous variables

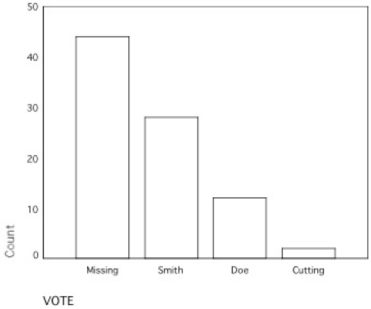
Line graph

- Example: distribution of scores on an exam



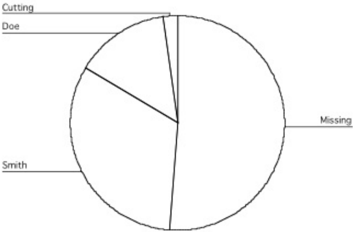
Graphs for categorical variables

Bar chart



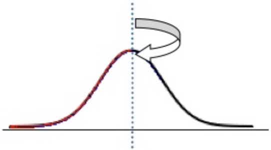
Graphs for categorical variables

Pie chart



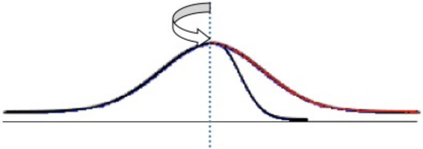
Properties of distributions: Shape

Symmetric



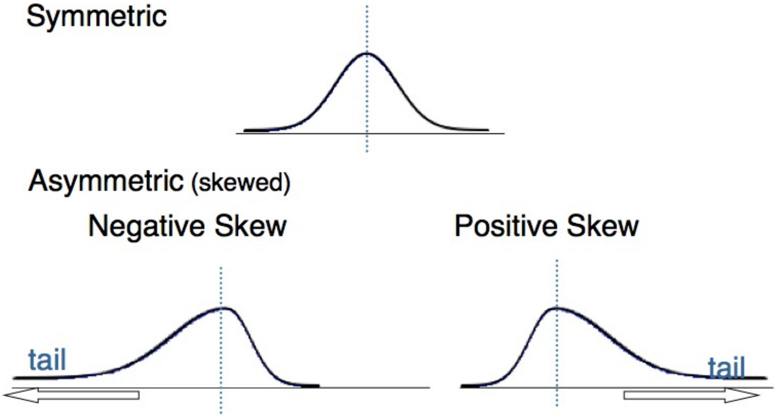
- The two sides line up

Asymmetric (skewed)

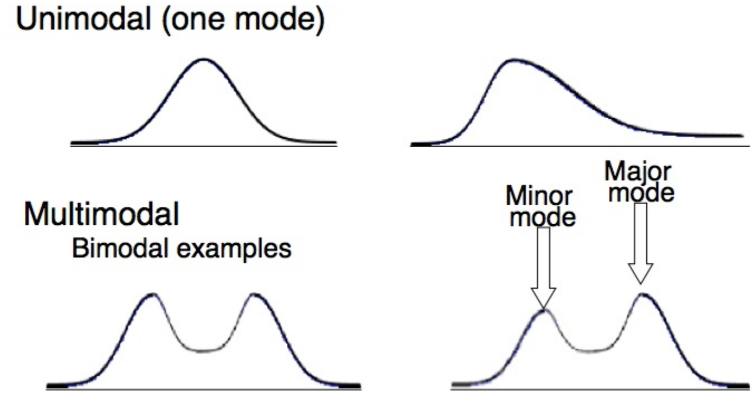


- The two sides do not line up

Properties of distributions: Shape



Properties of distributions: Shape



Properties of distributions: Center

There are three main measures of center

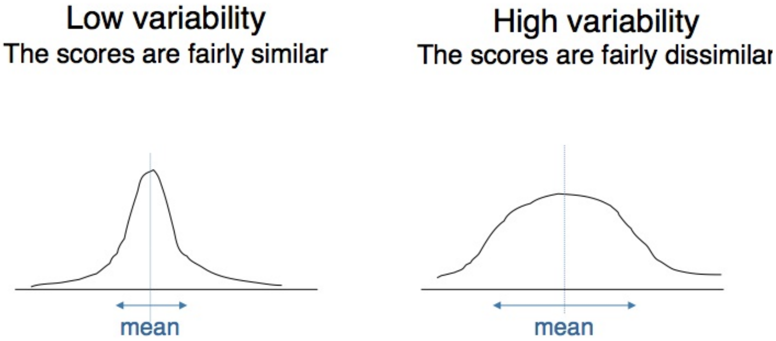
- Mean: the average
  - add up all the scores and divide by the total number
  - Most used measure of center
- Median: the middle score
  - the score that separates the top 50% from the bottom 50%
  - good for skewed distributions (e.g., home prices, reaction times)
- Mode: the most frequent score
  - Good for nominal scales (e.g., eye color)
  - A must for multi-modal distributions

Properties of distributions: Spread

How similar are the scores?

- Range: max - min
  - Only takes two scores from distribution into account
  - Influenced by outliers
- Standard deviation: the average amount that the scores in the distribution deviate from the mean
  - Takes all of the scores into account
  - Also influenced by outliers, but not as much as range
- Variance: standard deviation squared

Visualizing variability



Low variability
The scores are fairly similar

High variability
The scores are fairly dissimilar