

Statistics One

Lecture 4 Correlations

Three segments

- Examples
- Computation of the correlation coefficient
- Interpretation of correlations

Lecture 4

Segment 1

Correlation examples

Correlations 1

- Important concepts/topics
 - What is a correlation?
 - What are they used for?
 - Scatterplots

Correlations 1

- Correlation
 - A statistical procedure used to measure and describe the relationship between two variables
 - Correlations can range between +1 and -1
 - +1 is a perfect positive correlation
 - -1 is a perfect negative correlation

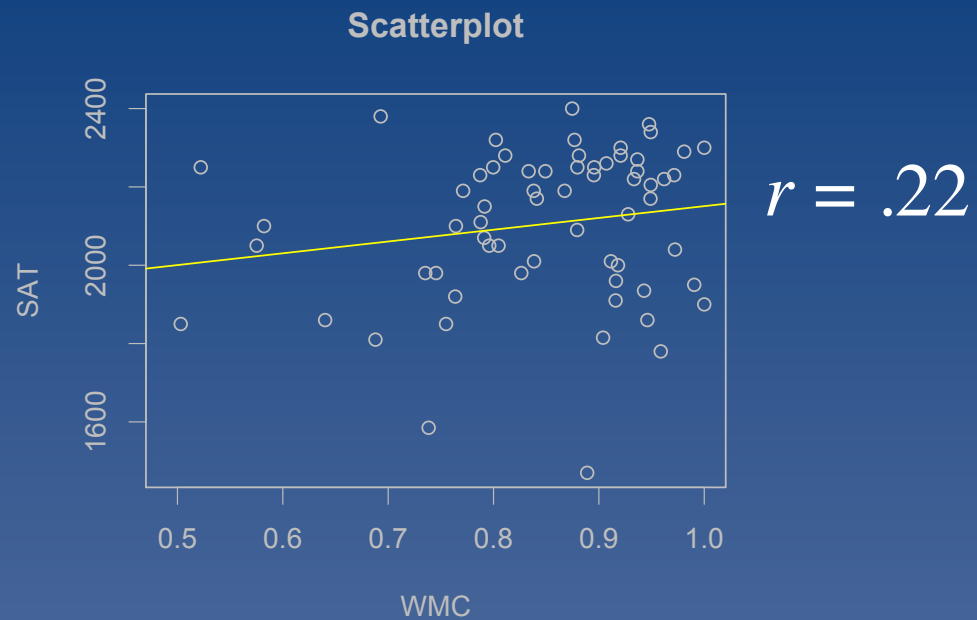
Correlations 1

- Assume two variables, let's call them X and Y, are correlated, then one variable can be used to predict the other variable
 - More precisely, a person's score on X can be used to predict his or her score on Y

Correlations 1

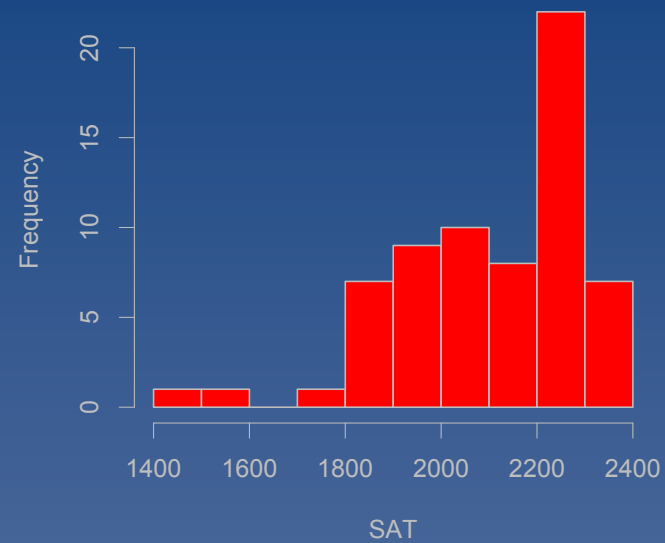
- For example, working memory capacity (X) is strongly correlated with SAT score (Y)
 - So if I know a person's score on a working memory task then I can predict that person's score on the SAT

R scatterplot: `plot(SAT~WMC)`

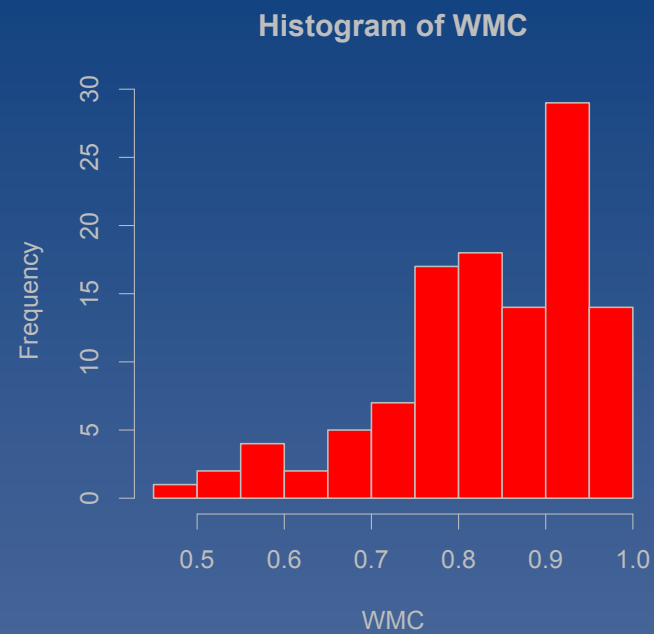


SAT scores

Histogram of SAT



WMC scores



Correlations 1

- CAUTION!
- Accuracy of prediction will depend upon:
 - The magnitude of the correlation

Correlations 1

- CAUTION!
- The magnitude of the correlation depends upon
 - Reliability of X
 - Reliability of Y
 - Sampling (random and representative?)

Correlations 1

- CAUTION!
- The validity of the prediction depends upon
 - Validity of X
 - Validity of Y
 - Several other assumptions (see segment 3)

Correlations 1

- CAUTION!
- Most important:
 - The correlation coefficient is a sample statistic
 - It does not apply to individual cases in the sample!

A more serious example

- Intelligence testing & World War 1 (WW1)
 - The United States became officially involved in WW1 in April, 1917
 - President Woodrow Wilson successfully kept the U.S. neutral for 2.5 years prior to April, 1917
 - His diplomacy was admirable but as a consequence the U.S. and its citizens did not expect to go to war so quickly in 1917, nor were they prepared for the magnitude of the war

A more serious example

- A challenge for the U.S. military
 - How to quickly recruit a large number of men and then assign them to positions in the U.S. military
 - Specifically, which men should be designated as officers and/or assigned to officer training?

A more serious example

- Historical coincident
 - The turn of the century witnessed the birth of intelligence testing and a scientific approach to individual differences in intelligence

The Army Alpha Battery

- Prominent Psychology researchers were recruited by the U.S. military to develop an aptitude test that could be administered to large groups of men and quickly scored
 - Multiple choice and short-answer questions
 - The contemporary equivalent is the American Services Vocational Aptitude Battery (ASVAB)

The Army Alpha Battery

- Robert Yerkes, one of the Psychologists, argued that the test measured “native intellectual ability” and claimed that the test was unaffected by culture
- However, later research showed that it was clearly culturally biased

Questions

- What type of statistical information could be presented to support Yerkes' claim that the test was not affected by culture?
- What type of information could be presented to refute Yerkes' claim?

A fun example

- Baseball statistics
 - Sabermetrics

Quick baseball lesson



Quick baseball lesson

- PA = Plate Appearances
- AB = At Bats
- H = Hits
- BB = Walks (Base on Balls)
- HBP = Hit by Pitch
- SF = Sacrifice Fly

Quick baseball lesson

- Batting average (AVG)
 - $AVG = (H) / (AB)$
- On-base percentage (OBP)
 - $OBP = (H + BB + HBP) / (AB + BB + HBP + SF)$

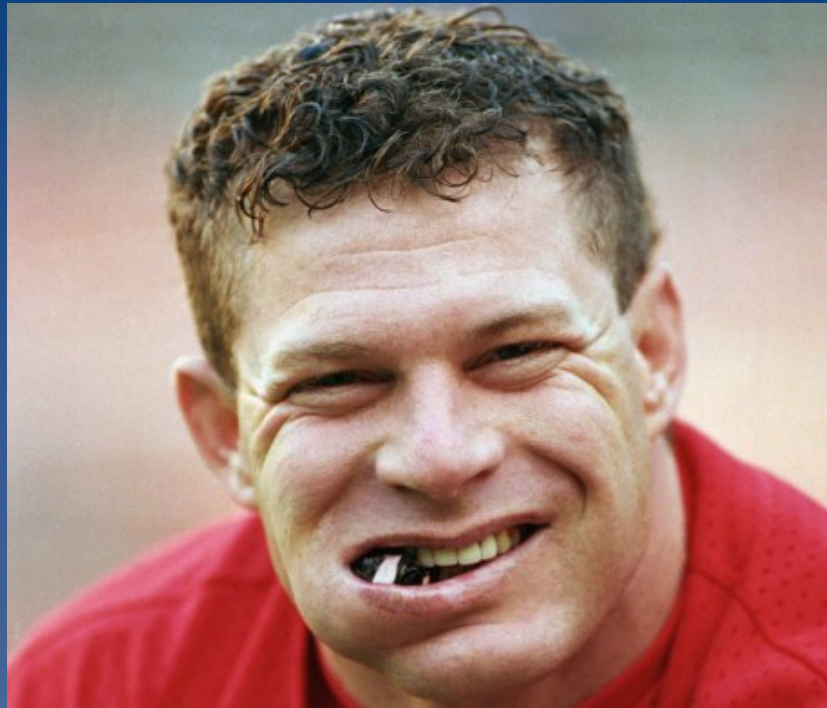
Fun example

- Which statistic is a better predictor of a player's contribution to the team's offense?
 - AVG or OBP?

Baseball history tangent

- Many people credit Billy Beane for recognizing the importance of OBP
- However, OBP was adopted by Major League Baseball as an officially recorded statistic in 1984
- In 1984 Billy Beane was playing for the New York Mets minor league team, sharing a hotel room with this guy...

Lenny Dykstra



Baseball history tangent

- Dykstra was 5 feet 10 inches, 180 pounds
- Beane was 6 feet 4 inches, 210 pounds
- Yet, Dykstra was promoted, instead of Beane
- Two years later, in 1986, Dykstra and the New York Mets won the World Series
- Beane's description of Dykstra:
 - He is "perfectly designed, emotionally" for baseball. He has "no concept of failure."

Baseball history tangents

- So who invented on-base percentage and demonstrated its importance?
 - Bill James (father of Sabermetrics)

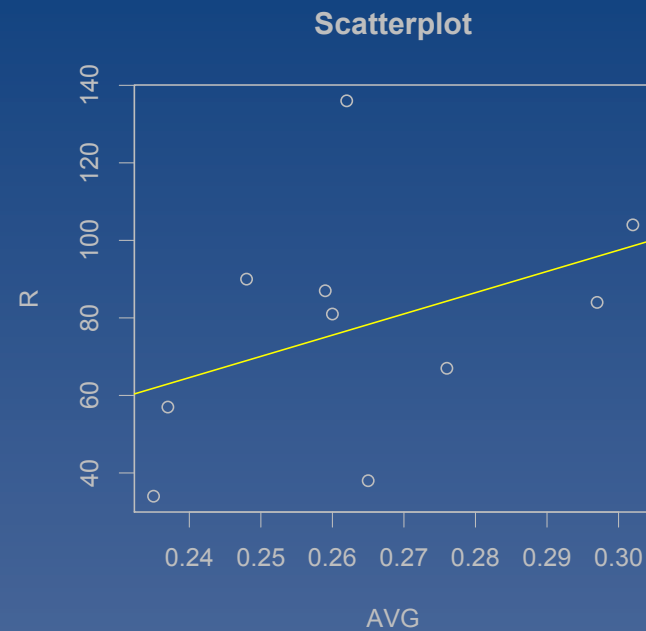
Back to the fun example

- What is the correlation between Runs Scored and AVG?
- What is the correlation between Runs Scored and OBP?

Back to the fun example

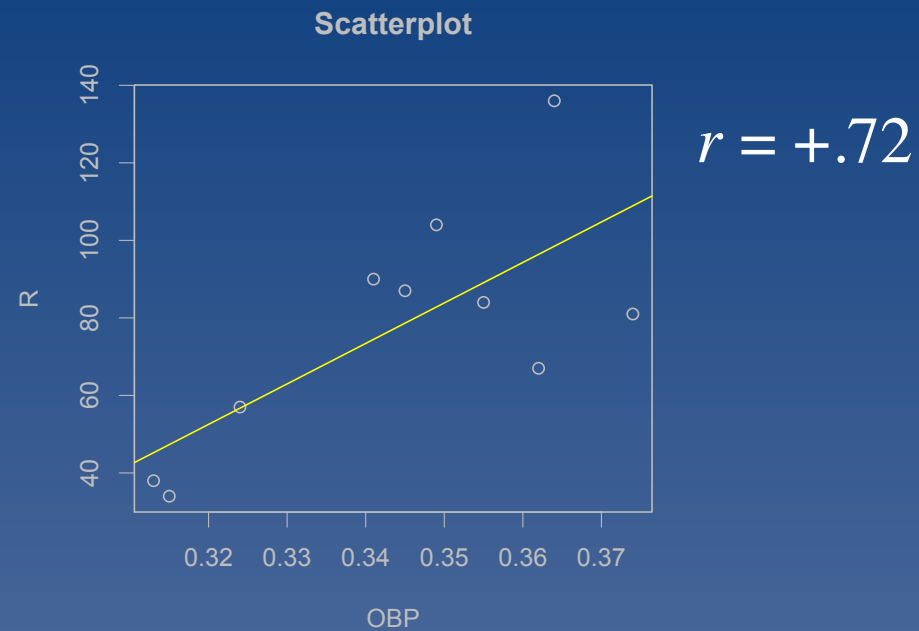
- Data from the New York Yankees (2011)
 - 10 offensive players appeared in more than half the games so I obtained the data from just those 10 players (the data are available online)
 - So, $N = 10$

R scatterplot: `plot(R~AVG)`



$r = +.40$

R scatterplot: `plot(R~OBP)`



Correlations 1: Review

- Important concepts/topics
 - What is a correlation?
 - What are they used for?
 - Scatterplots

Image in slide 22 was retrieved from
<http://static.ddmcdn.com/gif/lou-gehrig-hof.jpg>

Image in slide 27 was retrieved from
<http://dailybail.com/storage/lenny-dykstra.jpg>

© 2012 Andrew Conway