PSY 201: Statistics in Psychology

Lecture 11

Correlation

Is there a relationship between IQ and problem solving ability?

Greg Francis

Purdue University

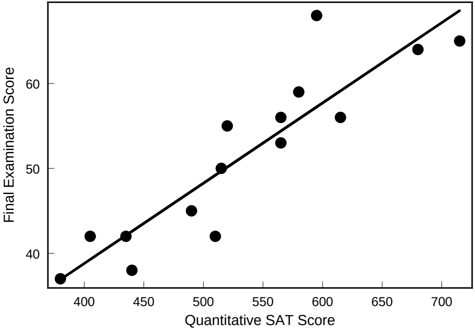
Fall 2023

CORRELATION

- suppose you get $r \approx 0$.
- Does that mean there is no correlation between the data sets?
- many aspects of the data may affect the value of r
  - Linearity of data.
  - Homogeneity of group.
  - Size of group.
  - Restricted range.

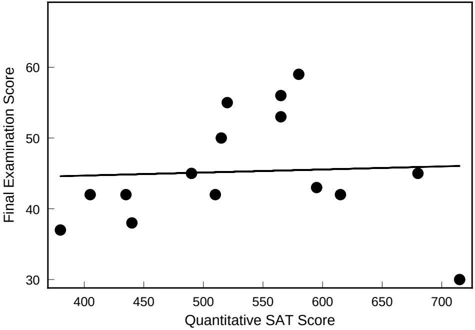
LINEARITY

- r is partly an index of how well a straight line fits the data set
- Here, $r = 0.903$



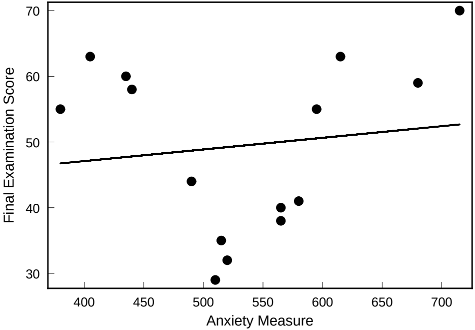
NONLINEARITY

- when data points don't fall along a single line (nonlinear data)
- Here, $r = 0.05$



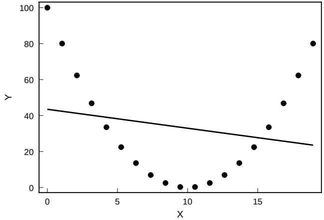
NONLINEARITY

- there are lots of types of nonlinearities
- curvilinear relationship
- Here, $r = 0.131$

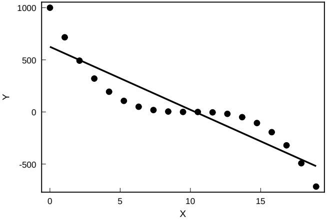


NONLINEARITY

- It can get complicated
- $r = -0.20$



- $r = -0.91$



BOTTOM LINE

- Pearson r is an index of a **linear** relationship between variables
- if another (nonlinear) relationship exists, r might not notice it
- Pearson r measures only simple relationships between variables
- if r is small, you might want to plot a scattergram to look at the data to notice if other relationships exist

HOMOGENEITY

- suppose you get $r \approx 0$, and you cannot detect any type of nonlinear relationship
- Does this mean there is no correlation between the variables?
- Not necessarily, it may be that the data does not have enough variation in it
- Correlation measures how variable X changes with variable Y
- if one doesn't change much, there won't be a strong correlation

HOMOGENEITY

- consider the covariance formula

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- where, covariance is

$$s_{xy} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n - 1}$$

- if there is little change in Y from $\overline{Y}$, s_{xy} is going to be small because $+/-$ variations in $X - \overline{X}$ will be weighted by small values of $Y - \overline{Y}$
- similarly, s_y is going to be small, so we divide a small number by a small number

HOMOGENEITY

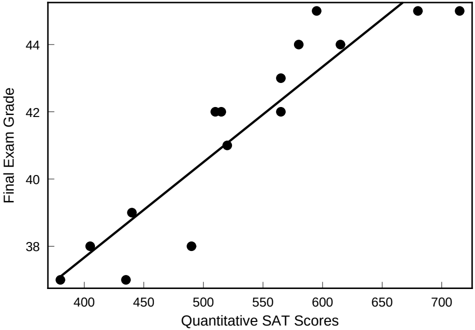
- intutively

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

- if one of those variables (or both) is not varying much at all, r will be small
- you need enough variability across both sets of scores to adequately measure correlation

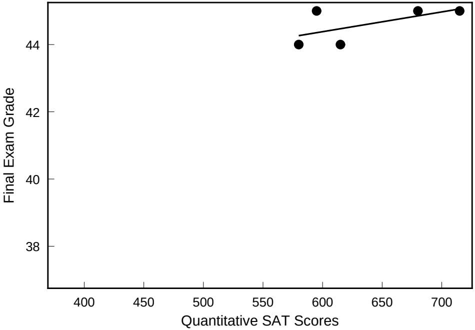
HOMOGENEITY

- the effects of homogeneity can be subtle
- relationship between SAT scores and Final exam grade
- $r = 0.92$



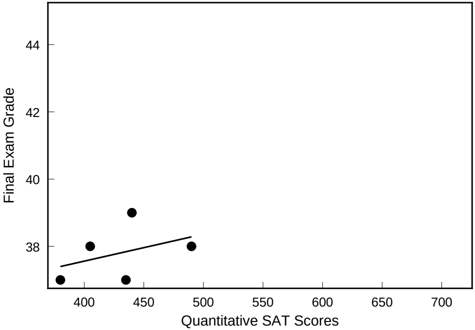
HOMOGENEITY

- suppose we looked at the relationship among only the best students
- (those with final exam scores above 44)
- $r = 0.62$



HOMOGENEITY

- or worst students
- (those with final exam scores below 40)
- $r = 0.62$



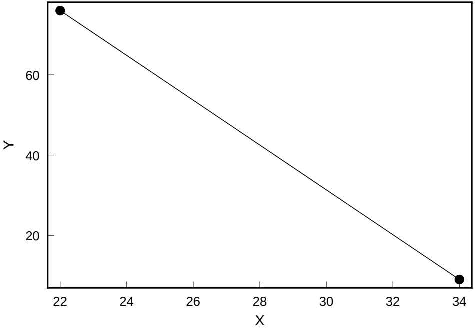
- correlation drops!

SIGNIFICANCE

- if you have $r \approx 0$, it may be because there is not enough variation in your data set
- e.g.
  - IQ and problem solving is probably unrelated among a group of geniuses
  - IQ and problem solving is probably unrelated among a group of idiots
  - IQ and problem solving is probably strongly related among a mix of geniuses, idiots, and normals

SIZE OF GROUP

- suppose you have only two data points
- you can always draw a straight line connecting them
- which implies perfect correlation
- $r = -1.0$



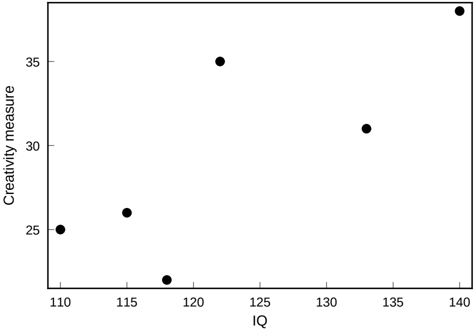
- (correlation doesn't tell us anything useful!)

SIZE OF GROUP

- if you have enough data points for correlation to be meaningful (> 2), and you have enough variation in the data, then
- size of group is not important in determining the **value** of r
- we will see later that it is important in determining the **accuracy** of the relationship (hypothesis testing)

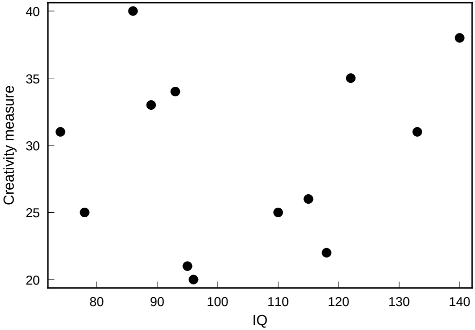
RESTRICTED RANGE

- if you sample data from a limited range you may not be able to trust the correlation values in general
- e.g., suppose you want to study relationships between IQ and creativity
- if you sample college students you will probably get IQ's between 110 and 140
- perhaps you find a strong correlation, e.g. $r = 0.78$



RESTRICTED RANGE

- if you sample from the general population (not just college students) you would get a larger range of IQs
- you may find a much weaker correlation, e.g. $r = 0.12$



RESTRICTED RANGE

- of course, it could be that you fail to find a large r over a restricted range, but a larger range finds a large r (this is slightly different from the issue of homogeneity)
- in general
- a correlation measure applies **only** to the range of values used to compute it
- you **cannot** extend the correlation value to other ranges

INTERPRETATION OF r

- if we calculate a value of r
- How do we know what it means?
- How do we compare r values for different data sets?
- Rule of thumb

$\|r\|$	Interpretation
0.9 to 1.0	Very high correlation
0.7 to 0.9	High correlation
0.5 to 0.7	Moderate correlation
0.3 to 0.5	Low positive correlation
0.0 to 0.3	Little if any correlation

SCALE OF r

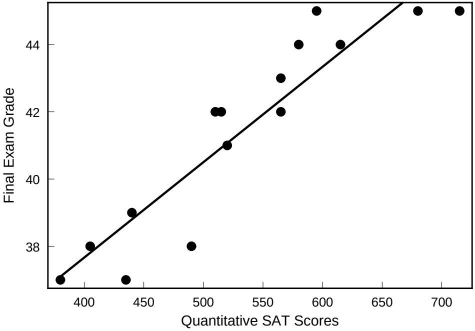
- values of r are **ordinal** measures of correlation
  - higher r values indicate larger correlation
  - equal spacings of r values may not indicate equal spacings of correlation
- thus, $r = 0.90$ is **not** twice as correlated as $r = 0.45$
- the difference in correlation between $r = 0.90$ and $r = 0.75$ is **not** the same as the difference in correlation between $r = 0.60$ and $r = 0.45$.

VARIANCE

- we can interpret r in terms of variance
- correlation coefficient indicates relationships between variables
- also indicates proportion of **individual differences** that can be associated with individual differences of another variable

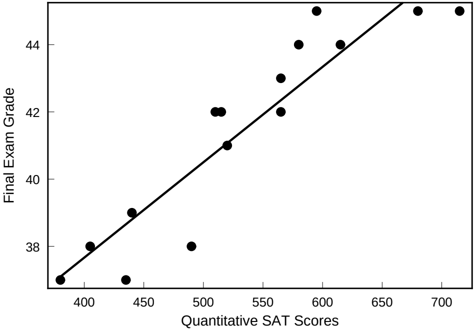
VARIANCE

- the idea is embedded in mathematical models
- assume you want to **predict** the final exam score when you know the SAT score
- line predicts score (could go in reverse too)



VARIATION

- deviation of a final exam score from the mean value can be due to deviation accounted for by SAT scores, or due to something else



VARIATION

- it turns out that

$$r^2 = \frac{s_a^2}{s_y^2}$$

- where:
  - s_y^2 = the total variance in y
  - s_a^2 = the variance in Y associated with variance in X
- thus, r^2 is the **proportion** of variance in Y accounted for with variance in X
- we are skipping the mathematical details (thank you!)
- called the coefficient of determination

CONCLUSIONS

- Pearson r
- size
- interpretation

NEXT TIME

- probability
- rules
- significance

Why casinos make money.