Statistics One

Lecture 5
Correlation

1

Three segments

- Overview
- Calculation of *r*
- Assumptions

2

Lecture 5 ~ Segment 1

Correlation: Overview

3

Correlation: Overview

- Important concepts & topics
  - What is a correlation?
  - What are they used for?
  - Scatterplots
  - CAUTION!
  - Types of correlations

4

Correlation: Overview

- 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
    - 0 is no correlation (independence)
    - -1 is a perfect negative correlation

5

Correlation: Overview

- When 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

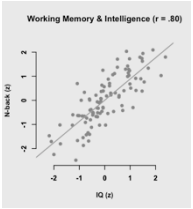
6

Correlation: Overview

- Example:
  - Working memory capacity is strongly correlated with intelligence, or IQ, in healthy young adults
  - So if we know a person's IQ then we can predict how they will do on a test of working memory

7

Correlation: Overview



8

Correlation: Overview

- CAUTION!
  - Correlation does not imply causation

9

Correlation: Overview

- CAUTION!
  - The magnitude of a correlation depends upon many factors, including:
    - Sampling (random and representative?)

10

Correlation: Overview

- CAUTION!
  - The magnitude of a correlation is also influenced by:
    - Measurement of X & Y (See Lecture 6)
    - Several other assumptions (See Segment 3)

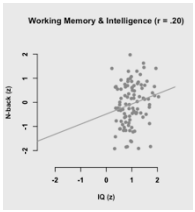
11

Correlation: Overview

- For now, consider just one assumption:
  - Random and representative sampling
  - There is a strong correlation between IQ and working memory among all healthy young adults.
    - What is the correlation between IQ and working memory among college graduates?

12

Correlation: Overview



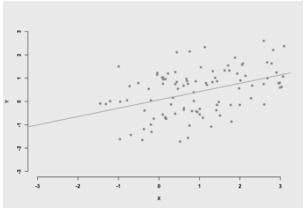
13

Correlation: Overview

- CAUTION!
- Finally & perhaps most important:
  - The correlation coefficient is a sample statistic, just like the mean
    - It may not be representative of ALL individuals
      - For example, in school I scored very high on Math and Science but below average on Language and History

14

Correlation: Overview



15

Correlation: Overview

- Note: there are several types of correlation coefficients, for different variable types
  - Pearson product-moment correlation coefficient (*r*)
    - When both variables, X & Y, are continuous
  - Point bi-serial correlation
    - When 1 variable is continuous and 1 is dichotomous

16

Correlation: Overview

- Note: there are several types of correlation coefficients
  - Phi coefficient
    - When both variables are dichotomous
  - Spearman rank correlation
    - When both variables are ordinal (ranked data)

17

Segment summary

- Important concepts/topics
  - What is a correlation?
  - What are they used for?
  - Scatterplots
  - CAUTION!
  - Types of correlations

18

END SEGMENT

19

Lecture 5 ~ Segment 2

Calculation of *r*

20

Calculation of *r*

- Important topics
  - *r*
    - Pearson product-moment correlation coefficient
      - Raw score formula
      - Z-score formula
  - Sum of cross products (SP) & Covariance

21

Calculation of *r*

- *r* = the degree to which X and Y vary together, relative to the degree to which X and Y vary independently
- *r* = (Covariance of X & Y) / (Variance of X & Y)

22

Calculation of *r*

- Two ways to calculate *r*
  - Raw score formula
  - Z-score formula

23

Calculation of *r*

- Let's quickly review calculations from Lecture 4 on summary statistics
  - Variance = SD^2 = MS = (SS/N)

24

Linsanity!

25

Jeremy Lin (10 games)

Points per game	(X-M)	$(X-M)^2$
28	5.3	28.09
26	3.3	10.89
10	-12.7	161.29
27	4.3	18.49
20	-2.7	7.29
38	15.3	234.09
23	0.3	0.09
28	5.3	28.09
25	2.3	5.29
2	-20.7	428.49
M = 227/10 = 22.7	M = 0/10 = 0	M = 922.1/10 = 92.21

26

Results

- M = Mean = 22.7
- SD^2 = Variance = MS = SS/N = 92.21
- SD = Standard Deviation = 9.6

27

Just one new concept!

- SP = Sum of cross Products

28

Just one new concept!

- Review: To calculate SS
  - For each row, calculate the deviation score
    - $(X - M_x)$
  - Square the deviation scores
    - $(X - M_x)^2$
  - Sum the squared deviation scores
    - $SS_x = \Sigma[(X - M_x)^2] = \Sigma[(X - M_x) \times (X - M_x)]$

29

Just one new concept!

- To calculate SP
  - For each row, calculate the deviation score on X
    - $(X - M_x)$
  - For each row, calculate the deviation score on Y
    - $(Y - M_y)$

30

Just one new concept!

- To calculate SP
  - Then, for each row, multiply the deviation score on X by the deviation score on Y
    - $(X - M_x) \times (Y - M_y)$
  - Then, sum the "cross products"
    - $SP = \Sigma[(X - M_x) \times (Y - M_y)]$

31

Calculation of *r*

Raw score formula:

$r = SP_{xy} / \mathbf{SQRT}(SS_x \times SS_y)$

32

Calculation of *r*

$SP_{xy} = \Sigma[(X - M_x) \times (Y - M_y)]$

$SS_x = \Sigma(X - M_x)^2 = \Sigma[(X - M_x) \times (X - M_x)]$

$SS_y = \Sigma(Y - M_y)^2 = \Sigma[(Y - M_y) \times (Y - M_y)]$

33

Formulae to calculate *r*

$r = SP_{xy}$ / **SQRT** $(SS_x \times SS_y)$

$r = \Sigma[(X - M_x) \times (Y - M_y)]$ /
SQRT $(\Sigma(X - M_x)^2 \times \Sigma(Y - M_y)^2)$

34

Formulae to calculate *r*

Z-score formula:

$r = \Sigma(Z_x \times Z_y) / N$

35

Formulae to calculate *r*

$Z_x = (X - M_x) / SD_x$
$Z_y = (Y - M_y) / SD_y$

$SD_x = SQRT\ (\Sigma(X - M_x)^2 / N)$
$SD_y = SQRT\ (\Sigma(Y - M_y)^2 / N)$

36

Formulae to calculate *r*

Proof of equivalence:

$Z_x = (X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)$

$Z_y = (Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)$

37

Formulae to calculate *r*

$r = \Sigma \{ [(X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)] \times [(Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)] \} / N$

38

Formulae to calculate *r*

$r = \Sigma \{ [(X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)] \times [(Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)] \} / N$

$r = \Sigma [(X - M_x) \times (Y - M_y)] / SQRT (\Sigma(X - M_x)^2 \times \Sigma(Y - M_y)^2)$

$r = SP_{xy} / SQRT (SS_x \times SS_y)$ ← The raw score formula!

39

Variance and covariance

- Variance = MS = SS / N
- Covariance = COV = SP / N

- Correlation is standardized COV
  - Standardized so the value is in the range -1 to 1

40

Note on the denominators

- Correlation for descriptive statistics
  - Divide by N
- Correlation for inferential statistics
  - Divide by N – 1

41

Segment summary

- Important topics
  - *r*
    - Pearson product-moment correlation coefficient
      - Raw score formula
      - Z-score formula
  - Sum of cross Products (SP) & Covariance

42

END SEGMENT

43

Lecture 5 ~ Segment 3

Assumptions

44

Assumptions

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
  - Linear relationship between X and Y
  - Homoscedasticity

45

Assumptions

- Assumptions when interpreting *r*
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

46

Assumptions

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
    - How to detect violations?
      - Plot histograms and examine summary statistics

47

Assumptions

- Assumptions when interpreting *r*
  - Linear relationship between X and Y
    - How to detect violation?
      - Examine scatterplots (see following examples)

48

Assumptions

- Assumptions when interpreting *r*
  - Homoscedasticity
    - How to detect violation?
      - Examine scatterplots (see following examples)

49

Homoscedasticity

- In a scatterplot the vertical distance between a dot and the regression line reflects the amount of prediction error (known as the "residual")

50

Homoscedasticity

- Homoscedasticity means that the distances (the residuals) are not related to the variable plotted on the X axis (they are not a function of X)
- This is best illustrated with scatterplots

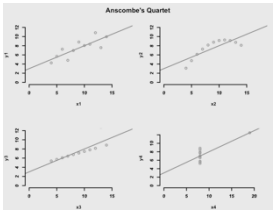
51

Anscombe's quartet

- In 1973, statistician Dr. Frank Anscombe developed a classic example to illustrate several of the assumptions underlying correlation and regression

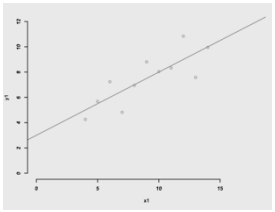
52

Anscombe's quartet



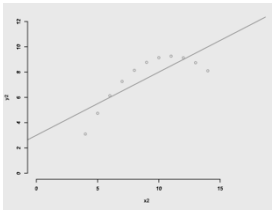
53

Anscombe's quartet



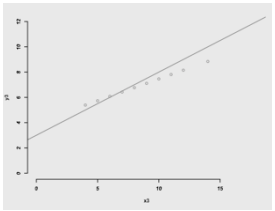
54

Anscombe's quartet



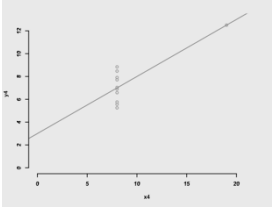
55

Anscombe's quartet



56

Anscombe's quartet



57

Segment summary

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
  - Linear relationship between X and Y
  - Homoscedasticity

58

Segment summary

- Assumptions when interpreting *r*
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

59

END SEGMENT

60

END LECTURE 5

61