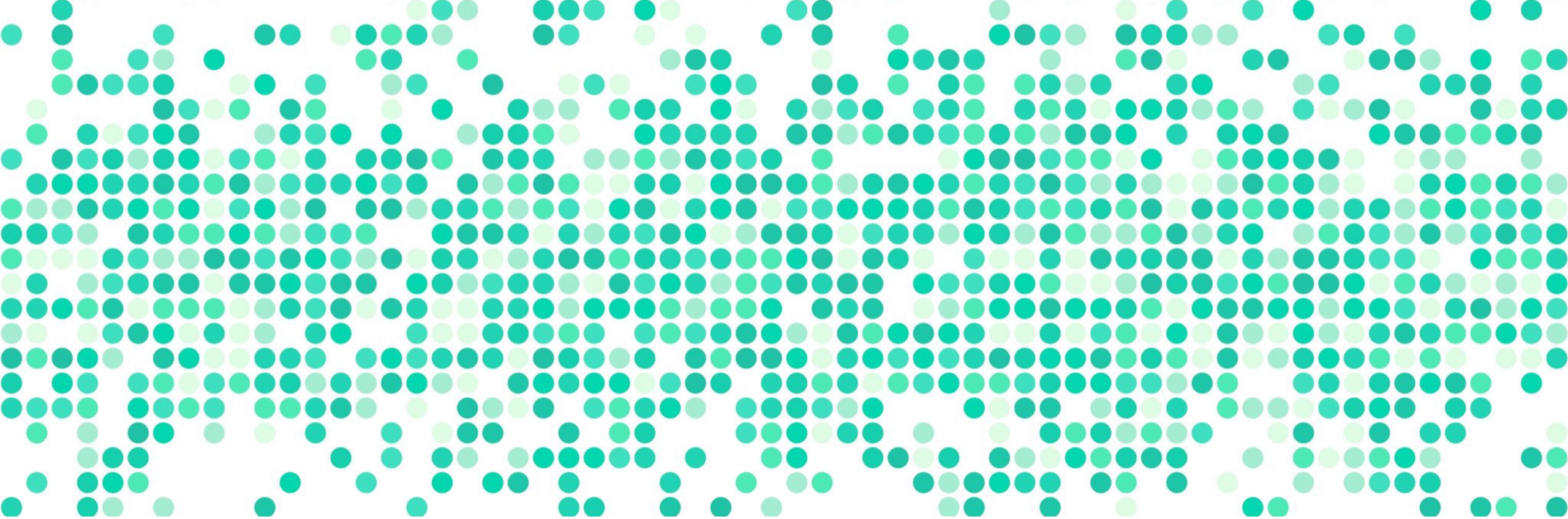


DATA ANALYSIS

Week 5: Midterm 1 review!

today’s agenda

- midterm 1 format
- correlations contd.
- conceptual + computational review
- questions

midterm 1 format

- in-class conceptual (on Canvas): 40% of first midterm grade
  - multiple choice, matching, short answer (quiz-like)
  - one help sheet + one formula sheet (on Canvas) + blank paper
  - closed book
  - very similar to practice quiz on Canvas!
- take-home computational: 60% of first midterm grade
  - short-answer + data analysis (problem set and worksheet like)
  - very similar to review activity on Canvas!
  - submissions will involve: (1) PDF of solution sheet + (2) downloaded worksheet
  - open book but NOT open person

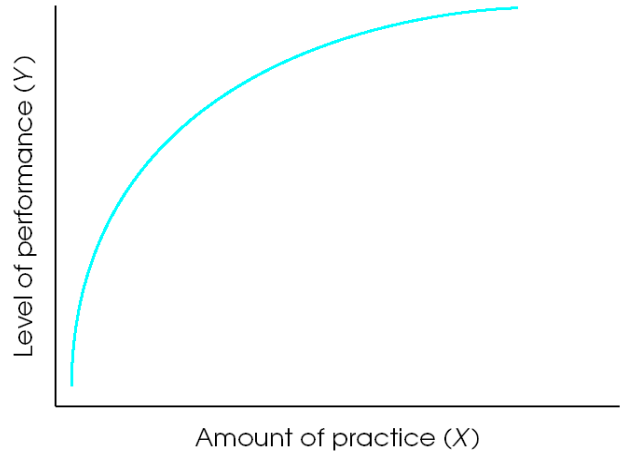
canvas walkthrough

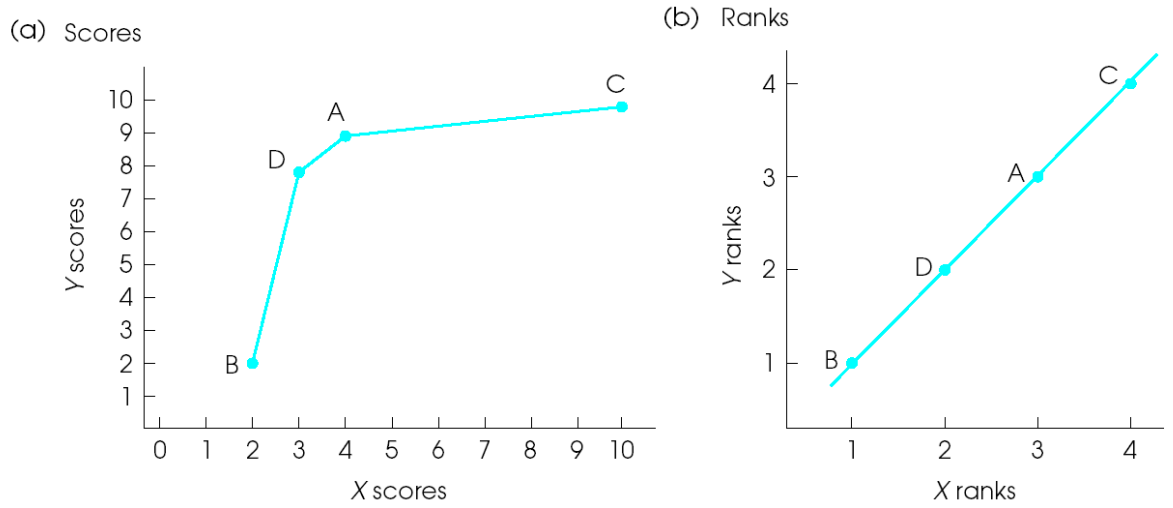
correlations recap

- when data are not interval/ratio, Pearson's r is not appropriate
- other alternatives exist
  - both variables ordinal: spearman's *rho*
  - one variable dichotomous (binomial): point biserial
  - both variables dichotomous: phi
- all alternatives are simply variations/extensions of Pearson's r

spearman's *rho*

- typically used for ordinal scales, non-linear relationships, or when outliers may need to be included
- uses ranks / ordering of scores instead of the raw scores themselves
- Pearson's r may underestimate the relationship but ranks may reveal a strong relationship
- if r is higher than rho, that typically means there is more of a linear trend in the data OR there are outliers that are exaggerating the pattern



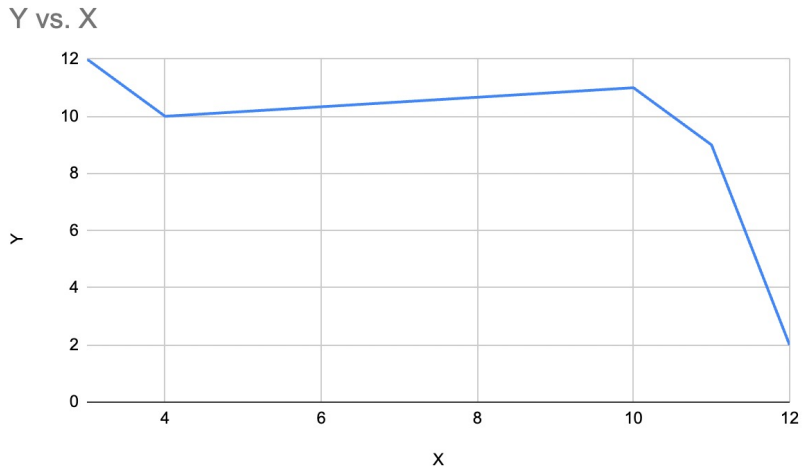


example

- a set of scores
- we first calculate Pearson's *r*

  =CORREL(X,Y)
- then we compute ranks
  - lowest numbers get lower ranks
- compute the pearson's *r* for ranks!

  =CORREL(rank_x, rank_y)



Person	X	Y
A	3	12
B	4	10
C	10	11
D	11	9
E	12	2

rank_x	rank_y
1	5
2	3
3	4
4	2
5	1

pearson
-0.6485442507

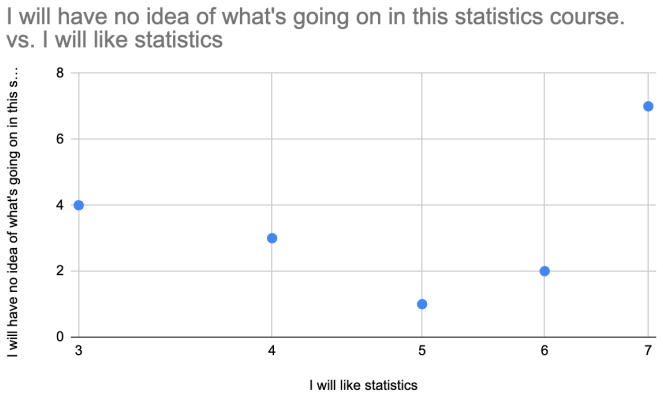
spearman
-0.9

activity: calculate spearman's rho

- calculate the correlation between two items from the statistics survey from class
- sheet (fake data)

Student	I will like statistics	I will have no idea of what's going on in this statistics course.
1	6	2
2	5	1
3	3	4
4	7	7
5	4	3

activity: calculate spearman's rho



Student	I will like statistics	I will have no idea of what's going on in this statistics course.	rank_like	rank_idea	rho	r
1	6	2	4	2	0.1	0.3434014099
2	5	1	3	1		
3	3	4	1	4		
4	7	7	5	5		
5	4	3	2	3		

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different

score
7
8
2
7
4
2
4

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different

score	initial_ranks
7	6
8	7
2	2
7	5
4	4
2	1
4	3

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different
- proceed as before: Pearson correlation for the ranks instead of raw scores

score	initial_ranks	final_ranks
7	6	5.5
8	7	7
2	2	1.5
7	5	5.5
4	4	3.5
2	1	1.5
4	3	3.5

point biserial and phi

- similar idea as Pearson's r but now our variables are not interval/ratio
- just converting the dichotomous variable to 0/1 numeric representations
  - point biserial : one variable dichotomous
  - phi : both variables dichotomous
- convert to numeric representations
- proceed as before

puzzle score	group
11	0
9	0
4	0
5	0
6	0
7	0
12	0
10	0
7	1
13	1
14	1
16	1
9	1
11	1
15	1
11	1
meanX	meanY
10	0.5

point biserial and phi

- similar idea as Pearson's r but now our variables are not interval/ratio
- just converting the dichotomous variable to 0/1 numeric representations
  - point biserial : one variable dichotomous
  - phi : both variables dichotomous
- convert to numeric representations
- proceed as before

puzzle score	group	sqx	sqy	z_x	z_y	z_x*z_y
11	0	1	0.25	0.2901905	-1	-0.2901905
9	0	1	0.25	-0.2901905	-1	0.2901905
4	0	36	0.25	-1.741143	-1	1.741143
5	0	25	0.25	-1.4509525	-1	1.4509525
6	0	16	0.25	-1.160762	-1	1.160762
7	0	9	0.25	-0.8705715001	-1	0.8705715001
12	0	4	0.25	0.5803810001	-1	-0.5803810001
10	0	0	0.25	0	-1	0
7	1	9	0.25	-0.8705715001	1	-0.8705715001
13	1	9	0.25	0.8705715001	1	0.8705715001
14	1	16	0.25	1.160762	1	1.160762
16	1	36	0.25	1.741143	1	1.741143
9	1	1	0.25	-0.2901905	1	-0.2901905
11	1	1	0.25	0.2901905	1	0.2901905
15	1	25	0.25	1.4509525	1	1.4509525
11	1	1	0.25	0.2901905	1	0.2901905
meanX	meanY	SSx	SSy			r
10	0.5	190	4			0.5803810001
		sd_x	sd_y			
		3.446012188	0.5			

review: key concepts

- frequency distributions
- mean / median / mode
- variance and standard deviation
- z-scores
- correlation
- regression
- assessing model fit

conceptual review

- questions to answer for each concept:
  - what is it?
  - how do we calculate/obtain it?
  - what information does it give us beyond raw scores?
- discuss with your partner
- come back and ask questions!

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> In a study of the dynamics of aggression, a researcher observed 20 children playing and recorded the number of acts of aggression over a 20-minute period. The researcher wants to know the spread between the most and least acts of agression. Which statistic will best allow him to answer this question?

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> To what extent is there a relationship between text messaging and student performance? A research asks 100 students to report how often they send a text message in a week. She wants to know whether those who send more text messages are less likely to have high GPAs? Which statistic will best allow her to answer this question?

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> Sally wants to know whether her daughter's performance on the SAT makes her a good candidate for a top college. She has obtained a frequency distribution for SAT scores, but does not have access to a mean nor a standard deviation. Which statistic can she calculate to best determine her daughter's relative SAT performance?

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> A college's Dean is concerned that grades are inflated at her school. As a first step, she asks the Registrar to send her the GPAs of all the students at her school. She now wants to create a visual display of the frequency distribution for the GPAs. Which display will best allow her to present a visualization of the distribution at a faculty meeting?

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> A researcher is interested in using a parent's level of optimism to predict the number of times a parent will speak in a positive manner to his/her child during a stressful situation. Which statistic will best allow him to predict, as accurately as possible, number of positive speech acts based on optimism?

practice quiz questions

- choose the appropriate statistic from: percentile rank, median, mode, mean, correlation coefficient, range, regression, histogram, z-scores, polygon, standard deviation, scatterplot, variance

> What is a typical annual income for households? A researcher obtains income data from the U.S. census bureau. The modal income is $20,000, the median is $50,000, and the average is $85,000. Which statistic best represents typical income?

computational review

- walk through review activity data

next time

- **before** class
  - *try*: practice quiz + review activity
  - *attend*: office hours with questions!!
- **during** class
  - midterm 1