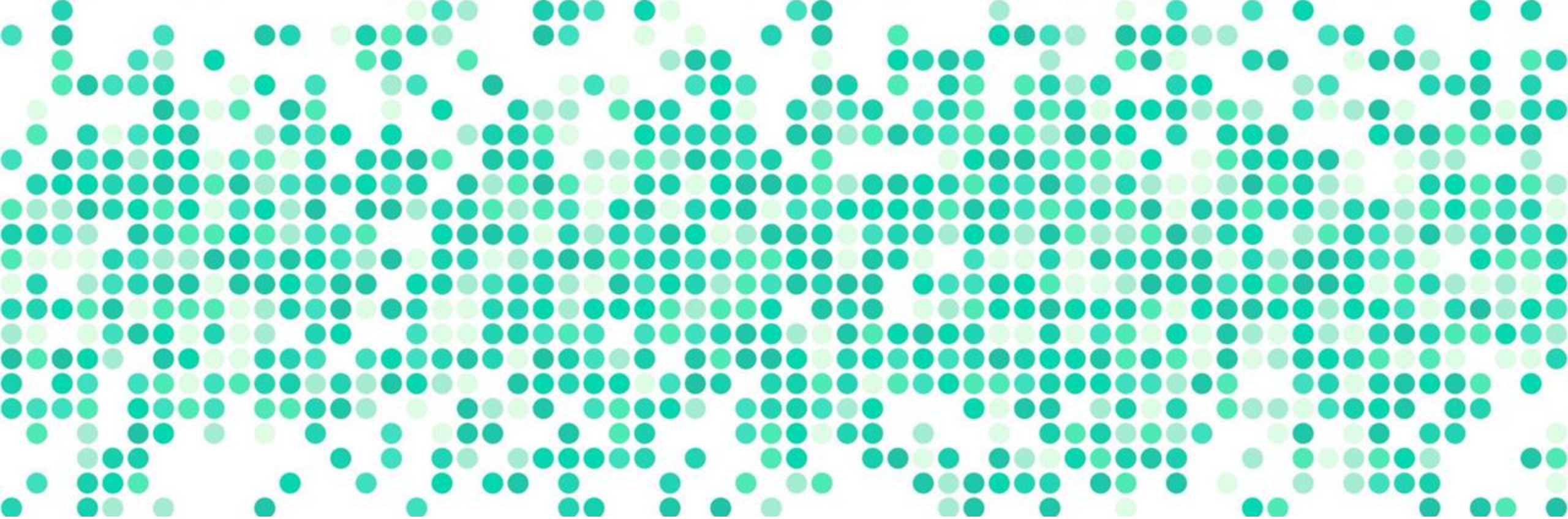
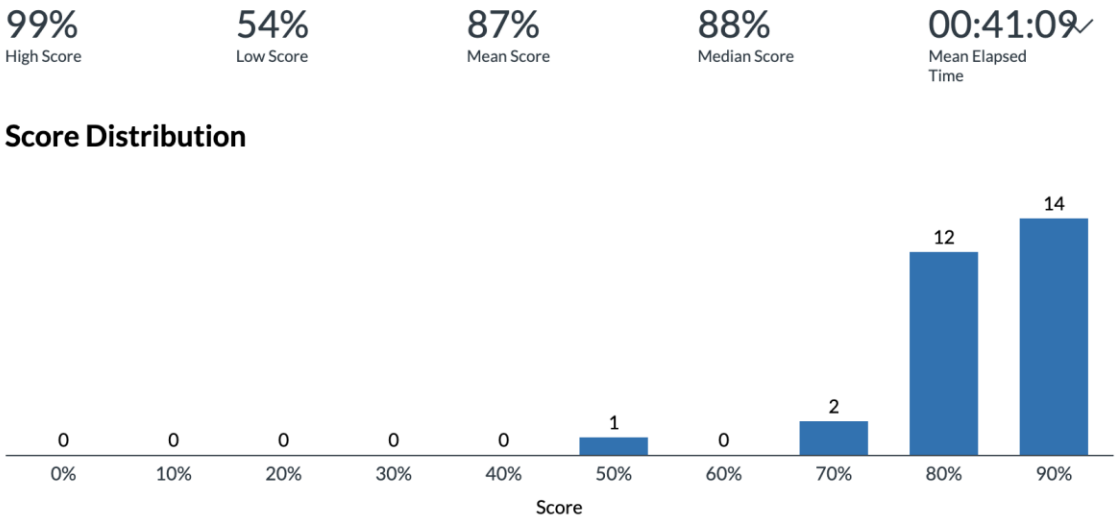


# DATA ANALYSIS

Week 6: Midterm 1 review!

# logistics




- practice midterm:
  - conceptual has been fully graded
  - computational graded for completion + answer key available (make sure to review)
- office hours and review sessions:
  - *Tuesday*: Prof. Kumar,  4.30 pm – 5.30 pm (Kanbar 217)
  - *Wednesday*: LA review, 5 pm -7.30 pm (Kanbar 101)
  - *Thursday*: Prof. Kumar, 4.30 pm – 5.30 pm (Kanbar 217)
  - *Friday*: Prof. Kumar, 9 am – 1 pm (Kanbar 217)
  - no LA office hours this weekend, they will not help with exam Qs
  - *Monday*: Prof. Kumar (TBD, will send email)

# today’s agenda

- midterm 1 format
- lingering questions
- review of difficult questions/concepts
- help and formula sheet

# midterm 1 (15%) format

- in-class conceptual: 40% of first midterm grade
  - multiple choice, matching, short answer
  - **closed** book
  - one help sheet + one formula sheet (on Canvas) + blank paper
  - be careful about true/false order – options will be shuffled!
- take-home computational: 60% of first midterm grade
  - short math + actual data analysis
  - submissions will involve: (1) PDF of solution sheet + (2) downloaded worksheet (.xlsx)
  - open book but NOT open person

# lingering questions

- What specifically do we need to know about percentiles?
- For the exam do we need to know the mathematical proof for things (like why the sum of the mean is always 0)

## Why take this course? a.k.a. learning goals

My hope as an instructor is to **empower** you with an analytical toolkit that will not only prepare you for other psychology courses you may take in your academic career, but also apply to other areas of your life. At the end of this course, you will be able to:

1. Describe the *conceptual* principles behind statistical thinking and uncertainty [Department Goal #1]
2. Apply a *computational and statistical* toolkit to test specific claims and questions [Department Goal #5]
3. Communicate effectively through numbers, graphs, and scientific writing [Department Goal #7]

# short answer questions

Siobhan is interested in the total food intake of her new hamster Holly. For a week, she tracks how many grams of pellets Holly eats. She finds that Holly eats an average of 10 grams of pellets every day. She then wants to compare Holly's intake to her turtle Tom's, whose mean consumption is 5 grams. So, she z-scores Holly's and Tom's food consumption. What will the sum of Holly's z-scored food consumption be? What will the sum of Tom's z-scored food consumption be? Explain why.

Sum will be 0, because the sum of z-scores is 0 always

Both sums of the z scores will be equal to 0. This is becasue z scores act as a standerdized form of measurement to classify how far away a point is from the mean. Thus, there are bith data points below and above the mean, which will cancel each other out when added together. This also makes sense since the z score of the mean is 0, and the z scores represent deviation from the mean. The positive and negative devaitions balance out from one another.

The sum of Z-scores is always zero because the z-score is a measurement of deviations form the mean so the mean of z-scores is always zero.

Because Z-scores for a population or sample are always equal to zero, because they act similarly to the mean with 0 being the balancing point, Holly's z-scored food consumption will sum to zero. All of the points above the mean and below the mean will balance out, The sum of Tom's Z-scores food consumption will also be zero for the same reason.

# short answer questions

When we calculate the standard deviation of a population, the denominator is n but when we calculate the standard deviation of a sample, the denominator is n-1. Explain why the denominators are different.

When calculating the sample you are trying to estimate the population standard deviation. It is not possible with a smaple to calulate the true population standard deviation, so you have to apply a correction factor called Bessel's correction. The n-1 compesates for the population variance and the bias in the estimation.

Because when we select samples we tend to get the most typical values in a population, therefore samples tend to underestimate values in a population. TO fix this, we penalize the sample by subtracting 1 from the denominator

To account for a specific bias a group of people out of the population may have.

This is because the standard deviation of a sample will always underestimate the standard deviation of a population. Thus, n-1 will make the value bigger to account for this underestimation.

# lingering questions

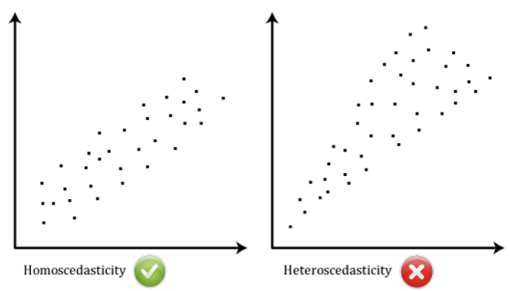
- I did not understand the second and forth questions in the quiz #5, plus even though I did 6th and 8th, I still have trouble understanding the reasoning behind it. Thanks!

A student records the number of ounces of caffeinated beverages consumed 24 hours before an exam and the exam score (an interval scale). In deciding whether it's appropriate to calculate a Pearson r, she wants to make sure that the data set

- ◯ is not homoscedastic
- ◯ is curvilinear
- ◯ does not have a truncated range on either variable
- ◯ has outliers

## Pearson's r assumptions

- interval/ratio scale: variables should be on interval / ratio scale: if the distance between the values is not equal, estimates of variability are difficult
- homoskedasticity: dispersion of Y remains relatively similar across the range of X
- no significant outliers
- variables should be approximately normally distributed



# Pearson's *r* vs. other correlations

- Pearson's *r*
  - interval/ratio scale: variables should be on interval / ratio scale
  - homoskedasticity: dispersion of Y remains similar across X
  - no significant outliers
  - variables should be approximately normally distributed
  - variables should have a *linear* relationship
- spearman's *rho*
  - clear non-linear pattern
  - both variables ordinal
- ***point biserial:*** one nominal variable with 2 levels, one interval/ratio variable
- ***phi***: both variables nominal with 2 levels

Ahana takes a Math and English midterm and wants to find out the relationship between her performance in the two courses. If the scores on both tests are standardized (i.e., z-scored), then what would the slope of the line that predicts Math z-scores using English z-scores if they are perfectly correlated?

- ○ b = -1
- ○ $b = \frac{s_{math}}{s_{english}}$
- ○ b = 1
- ○ $b = \frac{s_{english}}{s_{math}}$

$$b = r\,\frac{s_y}{s_x}$$

$$a = M_y - bM_x$$

The Netflix research team wants to evaluate whether there is any relationship between the viewership of two of their shows: *Love is Blind* and *Money Heist*. They assign the task to two new interns, who attempt to fit a linear model to the viewing data from 2017-2020 as follows: *Love is Blind* ~ a + b (*Money Heist*). Neither of the interns have taken a statistics course, though, so they merely estimate the values of *a* and *b* using guesswork. Which quantitative estimate could the research team use to assess which intern's models provides a better fit to the data?

- ◯ Sum of squared errors between actual viewing time for *Love is Blind* and predicted viewing time for *Love is Blind* based on the values of a and b from both interns
- ◯ Mean of viewing time for *Love is Blind* - mean of viewing time for *Money Heist*
- ◯ Sum of squared errors between actual viewing time for *Money Heist* and predicted viewing time for *Money Heist* based on the values of a and b from both interns
- ◯ Mean of viewing time for *Money Heist* - mean of viewing time for *Love is Blind*

I am a small business owner and I am interested in finding out whether displaying a thumbnail for my handmade oven mitts right next to the thumbnail for the shot glasses on my website is a good business strategy. I decide to try this strategy for a week and look at whether I find any relationship between the number of people buying shot glasses and oven mitts. I compute the correlation and find that $r$ = .5. If I wanted to report to my investors how much variance in oven mitts sales was explained by the sales of shot glasses, what would be an accurate statement?

- ◯ About 25% of the variance in oven mitt sales is explained by shot glass sales
- ◯ More than 50% of the variance in oven mitt sales is explained by shot glass sales
- ◯ About 2.5% of the variance in shot glass sales is explained by oven mitt sales
- ◯ 100% of the variance in oven mitt sales is explained by shot glass sales

# practice midterm Qs

Which of the following actions will always change the value of the mean?

- removing a score from the distribution
- All 3 of the other choices are correct
- changing the value of one score
- adding a new score to the distribution

## some properties of the mean

$$\mu = \frac{\sum X}{N}$$

- the calculation of the mean includes **all** values, so changing a score will change the mean
- adding a new score or removing a score will ***usually*** change the mean
  - unless the new score is the mean itself
- adding/subtracting/multiplying/dividing a constant value from each score will lead to applying the same operation to the mean

# when to use which measure?

- mean
  - most common, includes all scores, generally our “best” bet if we have no other variables
- median
  - extreme scores / skewed distribution and wanting to get a sense of “typical” or “most” values
  - undetermined values / open-ended distribution
- mode
  - nominal scale, only the mode can be used
  - if the “most typical case” is to be identified
  - mean and median often produce fractional values (that may not be interpretable)

# key concepts

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

# more lingering questions

- How does truncated range affect correlation? I don't remember going over it in class
- Could you please conceptually explain the difference between $SE_{model}$ and $SE_r$?
- what's the second way of calculating "*How much total variability in science scores is history scores able to explain*?" I only know R^2

# W6 activity: computational review

- is there a relationship between work experience and salary?
- what percentage of variability in salaries can be explained by work experience?
- can you predict salaries based on work experience?
- lingering question

# Salary ($) vs. Experience (Years)

# Salary ($) vs. Experience (Years)

# Salary ($) vs. Experience (Years)

# next time

- midterm 1
- good luck!!!
- bring a fully charged laptop
- bring a **handwritten** help sheet + formula sheet

## After Tuesday

- Review Answers/Feedback to Practice Midterm (Provided in comments)
- Do more ungraded practice for conceptual exam (new questions each time)
- Work on your help sheet + review all content from Week 1-5
- Attend office hours/review sessions!

## Thursday

- Complete Midterm 1 (Conceptual): IN CLASS