EDCI 913 - Statistical Analysis in Curriculum and Instruction

Spring 2026 class notes

Chanh B. Lam, Ph.D.

College of Education
Kansas State University

January 16, 2026

EDCATS
Educators Powered by Purpose

1. Week 10. Correlation
    - I. Correlation coefficient
    - II. Inside the black box
    - III. Hypothesis testing with r
2. Week 11. Regression
    - I. Modeling a relationship
    - II. Ordinary Least Squares (OLS)
    - III. Model fitness and Inference
3. Week 12. Application

1. Week 10. Correlation
    - I. Correlation coefficient
    - II. Inside the black box
    - III. Hypothesis testing with r
2. Week 11. Regression
3. Week 12. Application

Correlation

- Correlation is understood as a *statistical* relationship between two variables.
- If two variables are correlated, once we know the change in one variable, we can hope to make an educated guess about the other.
- Here are some examples:
  - Net force and acceleration: $F = ma$ (holding the mass constant)
  - Judicial rulings and time since meal breaks[1]
  - Ice cream sales and deaths by drowning

[1] Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889-6892. https://doi.org/10.1073/pnas.1018033108

Ice cream can kill?

Let's check out this video:

https://www.youtube.com/watch?v=VMUQSMFGBDo

Takeaway:

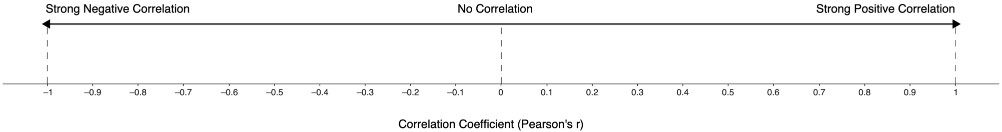
*Correlation is **not** causation.*

Measuring correlation

- Correlation is often measured by the *Pearson's correlation coefficient.*
- The correlation coefficient, ***denoted r*** (called "small" r), is a statistical measure that quantifies the strength and direction of a *linear* relationship between two variables.
- Usually, we will be able to observe one of these cases:
  - Positive correlation: The two variables tend to increase and decrease together.
  - Negative correlation: When one variable increases, the other tends to decrease.
  - No correlation: The two variables go wild and show no pattern at all.

Correlation coefficient

- The correlation coefficient, “small” r, ranges from -1 to 1.
- The closer the coefficient to either ends, the stronger the relationship. The closer the coefficient to zero, the weaker the relationship.
- Specifically, the correlation coefficient, r, indicates:
  - Positive correlation: r is close to 1,
  - Negative correlation: r is close to -1,
  - No correlation: r is close to 0.



Scatter plots

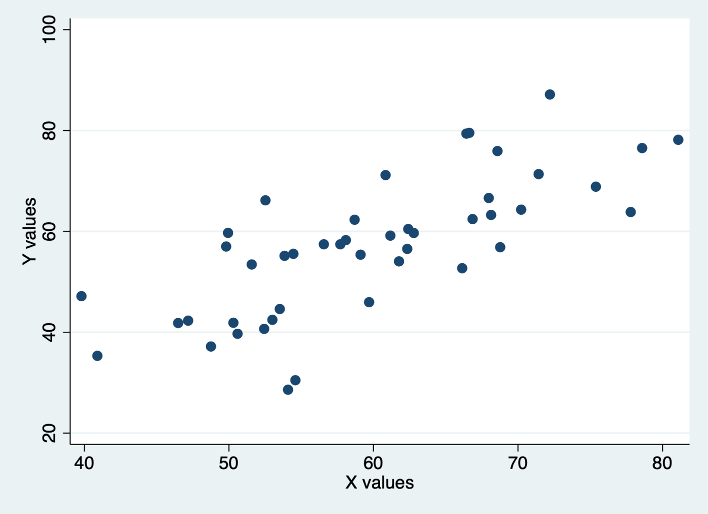
- A scatter plot is a type of graph used to visually display the relationship between two variables.
- The horizontal x-axis represents the values of one variable. The verticle y-axis represents the values of the other variable. Each point on the scatter plot represents a single observation with a pair of coordinates: x and y.
- Examining the scatter plot is usually the first step of observing a relationship.

A strong positive correlation

Y tends to ***increase*** when X increases.

$$r = 0.7458.$$

(r is close to 1)

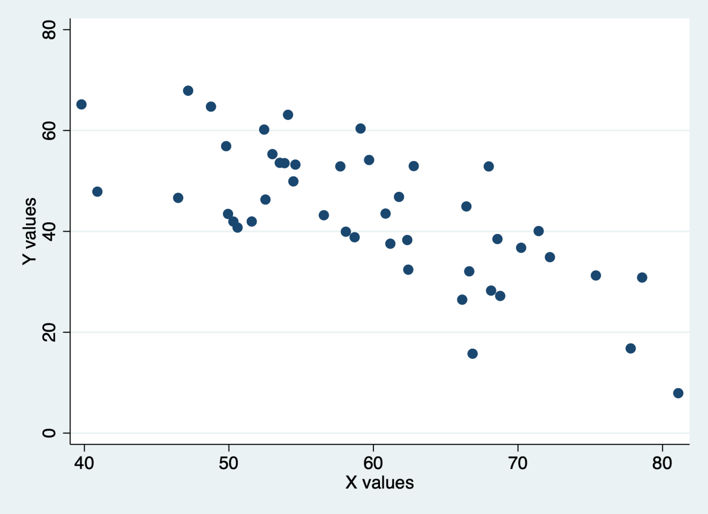


A strong negative correlation

Y tends to ***decrease*** when X increases.

$$r = -0.7326.$$

(r is close to -1)

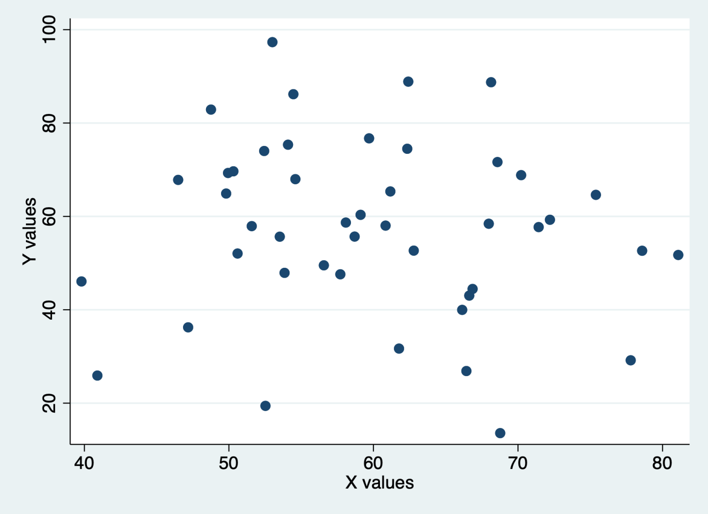


No correlation

The observations scatter around, showing **no pattern**.

$$r = -0.0980.$$

(r is close to 0)

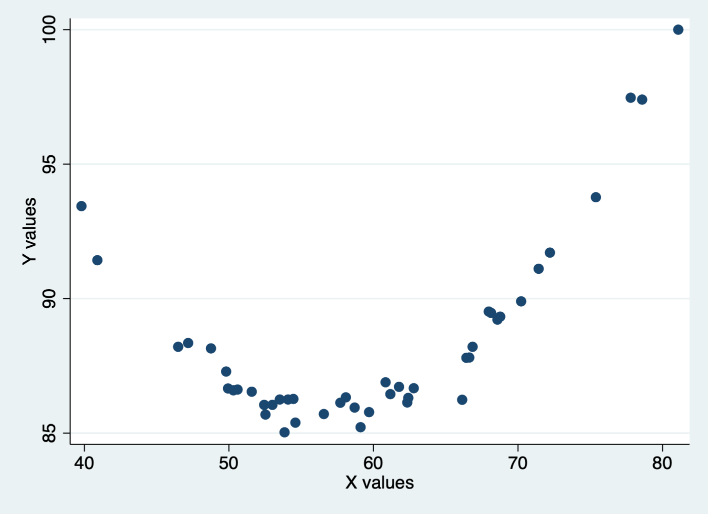


Non-linear relationship

The coefficient r detects some sort of positive correlation.

$$r = 0.5452.$$

But it's not always positive. Take a look at X from 45 to 55.



Blind spots of Pearson's r: Warning before the spell

- A strong r doesn't mean X is causing Y, or vice versa.

  *Correlation is **not** causation.*

- Whatever value of r, it doesn't tell you what relationship there is. Outliers or non-linear relationships are the coefficient's blind spots.
- Always examine your data thoroughly, such as by looking at the scatter plot, or several different plots, before running the software calculation.

1. Week 10. Correlation
   - I. Correlation coefficient
   - II. Inside the black box
   - III. Hypothesis testing with r
2. Week 11. Regression
3. Week 12. Application

What does r actually measure? Let's talk math

- Mathematically, the correlation coefficient is defined as

$$r = \frac{\text{Covariance}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)}.$$

- Recall the variance from Week 3? It measures how much the values of a variable vary *relative to its mean* by averaging these values' deviations (i.e., differences) from the mean, $x_i - \overline{x}$, such as $x_1 - \overline{x}$, $x_2 - \overline{x}$, ..., $x_{100} - \overline{x}$, etc.
- Covariance, as named, measures how much two variables *vary together* relative to their respective means by averaging the *products* of those deviations:

$$\text{Covar}(X, Y) = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y})}{N - 1}$$

Hands-on

Let's take a look at some data and see how this actually works:

week10_correlation.xlsx

- The first tab, "covariance," presents the manual calculation of the correlation coefficient. Try running the calculation before class.
- Observe how the product of deviations (i.e., differences from the mean), "(X-MeanX)*(Y-MeanY)," changes when the values of X and Y move relative to the mean of X and the mean of Y, respectively.
- You can also try different value sets of Y from the second tab, "more data," to see how the correlation coefficient changes. These were the data used to generate the scatter plots earlier.

The next page provides the explanation. Please be sure you have tried applying these calculations before moving forward.

Co-variation: The beauty of mathematics

- Each deviation product, "(X-MeanX)*(Y-MeanY)," returns a value with a sign when X and Y *co-move* relative to their respective means.
- If X moves above the mean of X, and Y moves above the mean of Y, the product yields a *positive* sign. If X moves below the mean of X, and Y *also* moves below the mean of Y, the product *still* has a positive sign.
- When X moves above mean X, but Y moves below mean Y, the product yields a *negative* sign. This reflects the relatively opposite movements.
- Moreover, the product itself gets larger in magnitude when X and Y move far from their means, and smaller when one or both are close to their means.
- Together, these are understood as *co*-variations. Averaging these (co)deviation products over $N - 1$ (because we have sample data), we have a measure that does the job of the variance, but for two variables, the *co*-variance.

Correlation coefficient: Again, it's standardization

- When mathematicians think they can standardize something, they just do it:

$$r = \frac{\text{Covariance}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)}.$$

- Covariance is a wild measure. It can be extremely large or extremely small, depending on the scale of the data. Using covariance isn't convenient!
- Rescaling the covariance using the variables' standard deviations, we obtain a nice, scale-free piece of number that stays within -1 to 1. That's our r. As you have been familiar, this is called standardization.

  r is simply *a standardized version* of the covariance.

  It tells you how much and to which direction the two variables co-move relative to their averages.

Hands-on, but not really

One last look at the dataset:

week10_correlation.xlsx

- Certainly, in practice we don't calculate through that many steps to arrive at the correlation coefficient. We use these built-in functions in Excel:
  - COVARIANCE.S() returns the (sample) covariance—the average of deviation products—from two variables.
  - CORREL() returns the correlation coefficient—the standardized, scale-free version of the covariance.
- You may try calculating these statistics both manually and using Excel functions, then compare the results.

1. Week 10. Correlation
   - I. Correlation coefficient
   - II. Inside the black box
   - III. Hypothesis testing with r
2. Week 11. Regression
3. Week 12. Application

Didn't we get rid of hypothesis testing?

- The short answer is: No, we didn't.
- In fact, any statistic is just a not-so-useful number without inference.
- What we need to know is:

  If we were to repeat our research, should it happen again that we find a correlation coefficient that is different from zero?

  The correlation coefficient equal to zero implies no correlation. A non-zero coefficient implies some sort of correlation.

Sampling distribution of the correlation coefficient

Assuming X and Y are randomly and normally distributed, the below test statistic

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

follows a Student's t distribution with $N-2$ degrees of freedom.

Of course, you don't have to remember the formula. But based on this, we have two key takeaways:

- The sampling distribution of the correlation coefficient is a t-distribution.
- The degree of freedom needed to perform hypothesis testing with this is $N-2$.

Hypotheses for the significance of r

Let's examine the correlation between early-career teachers' Praxis II scores and effectiveness in the classroom. We begin by stating our hypotheses:

- H_0: There is *no* correlation between teachers' Praxis II scores and effectiveness
- H_a: There *is* a correlation between teachers' Praxis II scores and effectiveness

This can be mathematically expressed as

- $H_0 : \rho = 0.$
- $H_a : \rho \neq 0.$

The Greek letter ρ ("rho") is the population notation of the correlation coefficient. (Recall a hypothesis is an assertion on the value of a population parameter? And yes, mathematicians love Greek letters.)

Hypothesis-testing coefficient r

- There are multiple ways to do hypothesis testing, as you have been familiar with over the past several weeks.
- With the correlation coefficient, we can do this even more conveniently.
- Instead of calculating the test statistic, then the critical value *for the test statistic*, we can get the critical value ***directly*** *for* r:

$$r_{\text{critical}} = \frac{t_{\text{critical}}}{\sqrt{N - 2 + t^2_{\text{critical}}}}.$$

You can either calculate your t-critical (e.g., using T.INV in Excel), then plug that in to obtain the r-critical. *Even better*, because this has been so common, you can in fact get r-critical values from most online calculators.

Hypothesis-testing with r

Here is the data set on the early-career teachers' Praxis II scores and their effectiveness evaluated on an interval scale from 0 to 100:

week10_praxis.xlsx

Download the data and try

- Generating a scatter plot to investigate what you're provided with.
- Calculating the Pearson's r coefficient of correlation.
- Use this online calculator to obtain the critical value and see if your correlation is significant. Make sure you select r-value, and take the two tailed criticals.

Praxis II and teacher effectiveness

- The correlation coefficient is calculated by Excel's CORREL:

$$r\text{-coefficient} = 0.4502.$$

- The online calculator shows the critical value for a 0.05-alpha, two-tailed test at this degree of freedom, $DF = N - 2 = 23$, is

$$r\text{-critical} = 0.3961.$$

- Since our r-coefficient is way larger than the r-critical, we have found a significant correlation.

Note: If you chose the manual way, you should get the t-critical of 2.0686, which returns the r-critical of 0.39607.

Interpreting the Pearson's correlation coefficient

- What we have found is the correlation coefficient, $r = 0.4502$, at its degree of freedom, is statistically significant. We can infer there is some sort of correlation between Praxis II scores and teacher effectiveness.
- Nevertheless, whether this is a strong correlation is a *different story*. It's not zero doesn't mean it's a strong correlation.
- The conventional wisdom says $r = 0.4502$ indicates a "moderate" correlation. However, today's research no longer interprets the correlation coefficient this way. Instead, most researchers interpret correlation coefficients in the respective contexts of their studies. Meanwhile, many studies use multiple approaches to their research question(s), where correlation is just a part of their work, and thus those r-coefficients are interpreted together with other pieces of evidence instead of as stand-alone statistics.

Other types of correlation (optional)

Besides Pearson's correlation, which works best for linearly related interval/ratio types of data (i.e., continuous variables), we have some other types of correlation:

- Point-biserial correlation: Adapted version of Pearson's r for one continuous (i.e., interval/ratio) and one dichotomous (e.g., yes/no) variables.
- Phi coefficient (ϕ): Another adapted version of Pearson's r that works for two dichotomous variables.
- Spearman's correlation: An approach different from Pearson's r that uses *orders* of variable values instead of the values per se. This works better for monotonic relationships, and can be used to handle ordinal data.
- Kendall's τ ("tau") coefficient: Also deals with ranking systems, but focuses specifically on agreement/disagreement (concordant/discordant) among the rankings. The Kendall's τ-coefficient in particularly popular in interrater reliability analysis, which you might have heard in assessment fields.