

# EDUCATIONAL ATTAINMENT AND MARRIAGE AGE — TESTING A CORRELATION COEFFICIENT'S SIGNIFICANCE

## TEACHER VERSION

**Subject Level:**
High School Math

**Grade Level:**
11–12

**Approx. Time Required:**
60 minutes

**Learning Objectives:**

- Students will be able to predict and test the significance of the relationship between two quantitative variables.
- Students will be able to write a line of best fit and interpret the slope and *y*-intercept in the context of the data.
- Students will be able to assess the strength and direction of a linear association based on a correlation coefficient.
- Students will be able to compute a correlation coefficient and distinguish between correlation and causation.

# Activity Description

Students will develop, justify, and evaluate conjectures about the relationship between two quantitative variables over time in the United States: the median age (in years) when women first marry and the percentage of women aged 25–34 with a bachelor's degree or higher. Students will write a regression equation for the data, interpret in context the linear model's slope and *y*-intercept, and find the correlation coefficient (*r*), assessing the strength of the linear relationship and whether a significant relationship exists between the variables. Students will then summarize their conclusions and consider whether correlation implies causation.

**Suggested Grade Level:**
11–12

**Approximate Time Required:**
60 minutes

**Learning Objectives:**

- Students will be able to predict and test the significance of the relationship between two quantitative variables.
- Students will be able to write a line of best fit and interpret the slope and *y*-intercept in the context of the data.
- Students will be able to assess the strength and direction of a linear association based on a correlation coefficient.
- Students will be able to compute a correlation coefficient and distinguish between correlation and causation.

**Topics:**

- Correlation vs. causation
- Hypothesis testing
- Line of best fit
- Linear regression

**Skills Taught:**

- Calculating and interpreting correlation coefficients
- Distinguishing between correlation and causation
- Testing the significance of a linear relationship
- Writing a regression equation that best models the data

# Materials Required

- The student version of this activity, 9 pages
- Graphing calculators (preferably TI-84 Plus) or graphing technology

# Activity Items

The following items are part of this activity. The items, their data sources, and any relevant instructions for viewing the source data online appear at the end of this teacher version.

- Item 1: Data Table
- Item 2: Optional Instructions for Calculating *r* on a TI-84 Plus
- Item 3: Critical Values of *r* at a 5 Percent Significance Level

For more information to help you introduce your students to the U.S. Census Bureau, read "*Census Bureau 101 for Students*." This information sheet can be printed and passed out to your students as well.

# Standards Addressed

See charts below. For more information, read "*Overview of Education Standards and Guidelines Addressed in Statistics in Schools Activities*."

## Common Core State Standards for Mathematics

| Standard | Domain | Cluster |
|---|---|---|
| **CCSS.MATH.CONTENT.HSS.ID.B.6**<br>Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.<br><br>**CCSS.MATH.CONTENT.HSS.ID.B.6.A**<br>Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models. | ID – Interpreting Categorical & Quantitative Data | Summarize, represent, and interpret data on two categorical and quantitative variables. |
| **CCSS.MATH.CONTENT.HSS.ID.C.7**<br>Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data. | ID – Interpreting Categorical & Quantitative Data | Interpret linear models. |

| Standard | Domain | Cluster |
|---|---|---|
| **CSS.MATH.CONTENT.HSS.ID.C.8** Compute (using technology) and interpret the correlation coefficient of a linear fit. | ID – Interpreting Categorical & Quantitative Data | Interpret linear models. |
| **CCSS.MATH.CONTENT.HSS.ID.C.9** Distinguish between correlation and causation. | ID – Interpreting Categorical & Quantitative Data | Interpret linear models. |
| **CSS.MATH.CONTENT.HSS.IC.A.1** Understand statistics as a process for making inferences about population parameters based on a random sample from that population. | IC – Making Inferences & Justifying Conclusions | Understand and evaluate random processes underlying statistical experiments. |

## Common Core State Standards for Mathematical Practice

| Standard |
|---|
| **CCSS.MATH.PRACTICE.MP3**. Construct viable arguments and critique the reasoning of others. Students will develop, justify, and evaluate their predictions about data. They will also reason inductively about data, making plausible arguments that account for the data's context. |
| **CCSS.MATH.PRACTICE.MP4**. Model with mathematics. Students will relate population data to predictions made about the association between two variables. They will then find the correlation coefficient and assess the significance of these variables' relationship. |

## National Council of Teachers of Mathematics' Principles and Standards for School Mathematics

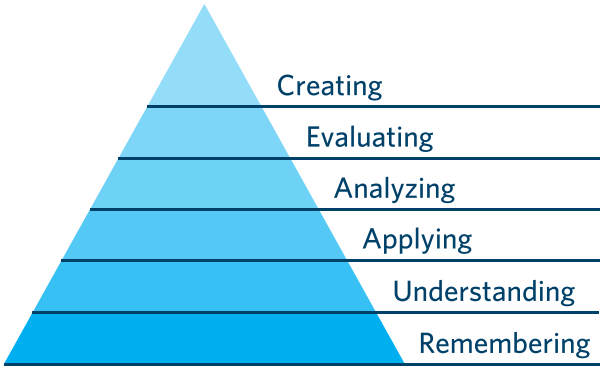
| Content Standard | Students should be able to: | Expectation for Grade Band |
|---|---|---|
| Data Analysis and Probability | Select and use appropriate statistical methods to analyze data. | For bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools. |
| Data Analysis and Probability | Develop and evaluate inferences and predictions that are based on data. | Understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference. |

## Guidelines for Assessment and Instruction in Statistics Education

| GAISE | Level A | Level B | Level C |
|---|---|---|---|
| Formulate Questions | X | | |
| Collect Data | | | |
| Analyze Data | | X | |
| Interpret Results | | X | |

## Bloom's Taxonomy

Students will ***evaluate*** data by making and testing predictions using inference.



# Teacher Notes

## Before the Activity

Students must understand the following key terms:

- **Confounding variable** – an outside variable that correlates with both the dependent and independent variables and could affect the conclusions we draw between them, possibly leading to a spurious (false) correlation
- **Correlation coefficient (*r*)** – a measure of the strength of a linear relationship between two variables — indicating how two variables vary jointly — whose absolute value indicates a stronger association when closer to 1 and a weaker association when closer to 0; the negative or positive sign of the coefficient indicates the direction of the relationship.
- **Alternative hypothesis** – a conjecture about the population that can be tested with sample data and that usually reflects a genuine association or difference in the population rather than random chance (i.e., the hypothesis that most researchers hope to establish with evidence)
- **Null hypothesis** – a conjecture about the population that can be tested with sample data and that usually reflects no association or difference in the population data (i.e., any association or difference observed in the sample reflects random variation in the data collection process)
- **Significance level** – the probability of rejecting the null hypothesis in a statistical test when it is actually true (typically 0.05)
- **Statistical significance** – when the relationship observed between the variables in the sample is unlikely to occur without a genuine relationship in the population
- **Critical value** – a point on the test distribution (typically listed in a table) that corresponds with a specified significance level and that must be less than the absolute value of the observed statistic to establish statistical significance (i.e., rejecting the null hypothesis) at that level
- **Degrees of freedom (*df*)** – the number of observations in a sample minus the number of population parameters (e.g., slope, correlation coefficient, and other measures) that must be estimated from that sample
- **Conjecture** – an opinion formed on the basis of inconclusive or incomplete evidence
- **Correlation** – a connection, including the degree and type of relationship, between two or more things
- **Regression equation** – a model of the relationship between two or more variables that predicts the value of the dependent variable for a given value of the independent variable(s)
- **Slope** – the rate of change in a linear model, or the amount by which a *y* value increases (for positive slopes) or decreases (for negative slopes) for every unit increase in an *x* value
- **y-intercept (constant)** – the value of *y* when a regression line crosses the *y*-axis (i.e., when the value of *x* is 0)
- **Residual** – the difference between the actual *y* coordinate of a data point and what the linear model predicts (actual - predicted)

Students should have a basic understanding of the following concept:

- How creating a residual plot can indicate the accuracy of a regression equation for the data

Students should have the following skills:

- Ability to create a scatter plot
- Ability to assess the strength and direction of the linear relationship between two quantitative variables based on the *r* value, a scatter plot, or a given context
- Ability to distinguish between correlation and causation
- Ability to write and interpret a line of best fit in the context of the data

Teachers should decide whether students will calculate their regression equation (question 3 of part 2) by hand or with technology. Teachers should be aware that **Item 2** provides instructions for calculating regressions and *r* values on a TI-84 Plus calculator, but that modifications may be needed if students are using other types of graphing technology.

## During the Activity

Teachers should be aware that "correlation," "association," and "relationship" are used interchangeably throughout the activity.

Teachers should pause after part 1 to lead a class discussion about students' responses and then could have students work independently or in groups of two to four for part 2, encouraging collaboration and discussion.

Teachers should caution students about interpreting the data in the activity when determining statistical significance. As with any data that are dependent over time, the observations happen neither simultaneously nor independently from year to year and may be misleading: The percentage of women with a bachelor's degree or higher in one year cannot really decrease much the next year because the population is the same. Teachers should also remind students that making predictions beyond the data points (extrapolation) risks accuracy and should be viewed with skepticism.

## After the Activity

Teachers should facilitate a class discussion in which students propose and debate potential reasons for the correlation between the median age of women when they first marry and the percentage of women aged 25–34 with a bachelor's degree or higher over time in the United States.

## Extension Ideas

Teachers could use other Statistics in Schools activities about similar topics to build on this activity.

# Student Activity

Click ***here*** to download a printable version for students.

## Activity Items

The following items are part of this activity and appear at the end of this student version.

- Item 1: Data Table
- Item 2: Optional Instructions for Calculating *r* on a TI-84 Plus
- Item 3: Critical Values of *r* at a 5 Percent Significance Level

## Student Learning Objectives

- I will be able to predict and test the significance of the relationship between two quantitative variables.
- I will be able to write a line of best fit and interpret slope and *y*-intercept in the context of the data.
- I will be able to assess the strength and direction of a linear association based on a correlation coefficient.
- I will be able to compute a correlation coefficient and distinguish between correlation and causation.

### Part 1 – Make Predictions

1. Between 2011 and 2019 in the United States, do you think the percentage of women aged 25–34 with a bachelor's degree or higher increased, decreased, or stayed the same? Explain your reasoning.

   **Student answers will vary.**

2. During the same period in the United States, do you think the median age of women when they were first married increased, decreased, or stayed the same? Explain your reasoning.

   **Student answers will vary.**

3. State the null and alternative hypotheses for whether there is a relationship between the variables in questions 1 and 2.

   Null hypothesis: **There is no relationship between the median age of women when they were first married and the percentage of women aged 25–34 with a bachelor's degree or higher in the United States between 2011 and 2019.**

   Alternative hypothesis: **There is a relationship between the median age of women when they were first married and the percentage of women aged 25–34 with a bachelor's degree or higher in the United States between 2011 and 2019.**

4. Make a conjecture predicting whether a relationship exists between the variables.

   **Student conjectures will vary.**

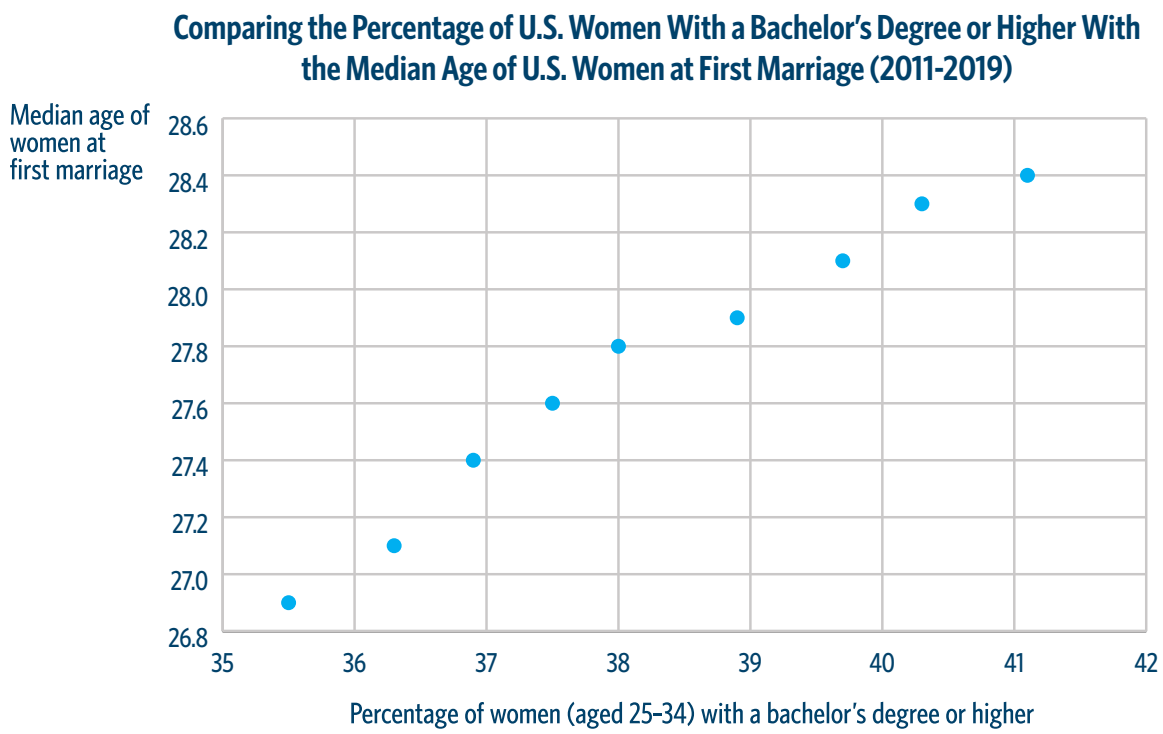
5. If you predicted a relationship in question 4, state its direction: Do you think there will be higher percentages of women with bachelor's degrees during years with *higher* or *lower* median ages of women when they were first married? Why?

   **Student responses (if any) will vary but should include either "positive" or "negative" and a justification of students' logic.**

## Part 2 – Evaluate Data to Assess Predictions

1. Use **Item 1: Data Table** to create a scatter plot on the following grid. Use the education data as the independent variable and the marriage data as the dependent variable, keeping in mind that this particular choice is arbitrary.

   **Student scatter plots should look similar to:**



2. Does a linear model appear to be a sufficient description of the relationship between the two variables in this sample? Explain your reasoning, while keeping in mind that the data are from sample estimates so they could include random error in their values.

   **Yes, a linear model appears to be a sufficient description of the relationship. Student explanations should include observations about the general pattern in the scatter plot or about the rate of change in the table.**

3. Find a regression equation (by hand or using technology; feel free to reference Item 2: Optional Instructions for Calculating *r* on a TI-84 Plus for help) that best models the data in your scatter plot. Round your values to the nearest hundredth, and explain your equation's meaning.

   **Student answers will vary but should be similar to: $\hat{y}$ = 0.27x + 17.366, meaning that the median age of women at first marriage is roughly equal to 17.366 plus 0.27 times the percentage of women aged 25–34 with a bachelor's degree or higher that year.**

4. Define the variables in the sample, and interpret the values of the regression coefficients in the context of the data.

   *x* values: **The *x* values represent the percentage of U.S. women aged 25–34 with a bachelor's degree or higher between 2011 and 2019.**

   *y* values: **The *y* values represent the median age of U.S. women when they were first married, from 2011 to 2019.**

   Slope: **The coefficient of *x*, 0.27, is the slope, or rate of change, which represents the change in the median age of women at first marriage relative to a corresponding change in the percentage of women aged 25–34 with a bachelor's degree or higher: For every 1 percentage point increase in the percentage of women aged 25–34 with a bachelor's degree or higher, the median age of women at first marriage increases 0.27 years, on average. In other words, as more women get bachelor's degrees, the age of new brides increases.**

   *y*-intercept: **17.366 is the *y*-intercept, meaning the regression equation predicts that if no 25- to 34-year-old women had bachelor's degrees or higher during a particular year, the median age at which women would first marry that year would be 17.366.**

5. How could you assess how accurately your regression equation represents the data?

   **I could plot and analyze the residuals. If the residual plot displays no pattern with respect to x, then the linear model accurately captures the main form of the association.**

6. Calculate the correlation coefficient (*r*) of your linear model from question 3 using graphing technology. (You can use Item 2 for reference.) Round your answer to the nearest thousandth.

   **Student answers may vary according to the graphing technology used but should be around 0.987.**

   a. Why would a person want to find a correlation coefficient?

      **The correlation coefficient helps you assess the strength and direction of the linear relationship between two variables.**

b. Based on the *r* value you calculated, how strong is the linear relationship between the variables?

**The correlation coefficient, 0.987, is very close to 1, which indicates a strong linear relationship between the variables.**

7. Calculate the degrees of freedom (*df*) for your equation using the formula *n* - 2, where *n* represents the number of pairs of data points. Show your work.

**$df = n - 2 = 9 - 2 = 7$**

a. Find the corresponding *df* in Item 3: Critical Values of *r* at a 5 Percent Significance Level, and determine whether your r value is greater than or equal to it.

**Critical value of $r = 0.666$**

**$r = 0.987$**

**$0.987 > 0.666$**

8. How does the critical value for the sample help us determine whether there is a significant relationship between the variables in the population?

**If the sample's *r* value is greater than or equal to its corresponding critical value, we can conclude that a significant relationship exists between the variables in the population, because the probability of finding an *r* value as extreme as, or more extreme than, the one we calculated (0.987) under the null hypothesis is very small (less than 5 percent). Such a small probability provides strong evidence for rejecting the null hypothesis and for the presence of a genuine correlation between the two variables in the population data.**

9. Based on what you found in question 8, is there a significant relationship between the median age of U.S. women at first marriage and the percentage of U.S. women aged 25–34 with a bachelor's degree or higher in the years observed? If so, explain and state the direction.

**Yes. Because the *r* value is greater than or equal to the critical value for my *df*, there is a significant correlation between the variables. The variables appear to have a positive correlation, indicating that years in which a higher percentage of women aged 25–34 have bachelor's degrees are also years with a higher median age of women at first marriage.**

10. Do the results support your initial conjecture of whether there would be a significant relationship? Explain.

**Student answers will vary.**

11. Does a significant correlation between two variables also indicate a cause-and-effect relationship? Explain, thinking about the two variables in this case.

**No, because correlation does not imply causation. Just because the two variables are related does not imply that one causes the other.**

12. Explain three possible interpretations of a significant correlation for this data set.

**Student answers will vary but could include:**

- **It is possible that the increase in the median age of women at first marriage caused the increase in the percentage of 25- to 34-year-old women with a bachelor's degree or higher, though there is no clear logic for this interpretation.**
- **It is possible that the increase in the percentage of 25- to 34-year-old women with a bachelor's degree or higher caused the increase in the median age of women at first marriage. A possible explanation for this interpretation is that women may be waiting to marry until after they graduate from college, so that as more women earn bachelor's degrees, the age at which women first marry increases.**
- **It is possible that the two variables each increased due to a confounding (lurking) third variable, such as time in this case.**

13. Which of the possible interpretations that you identified in question 12 is most likely to explain the results? Justify your theory.

**Student answers will vary but could include:**

- **The most likely interpretation is that the increase in the percentage of U.S. women aged 25–34 with a bachelor's degree or higher caused the increase in the median age of U.S. women at first marriage, because a logical rationale exists for this interpretation. Increased graduation rates for women in undergraduate degree programs may indicate that they are finishing their educations and stabilizing their careers at an older age than they might have otherwise at lower levels of education. They may prefer to wait until after graduation to get married, which would increase their age at that time. Still, because correlation does not necessarily imply causation, this rationale cannot be confirmed.**
- **The most likely interpretation is that the correlation is spurious, because time-series data often create an illusion of a causal relationship between two variables, as the observations are not fully independent.**

## Item 1: Data Table

| Year | Percentage of U.S. women (aged 25–34) with a bachelor's degree or higher | Median age of U.S. women (aged 15–54) at first marriage |
|---|---|---|
| 2011 | 35.5 | 26.9 |
| 2012 | 36.3 | 27.1 |
| 2013 | 36.9 | 27.4 |
| 2014 | 37.5 | 27.6 |
| 2015 | 38.0 | 27.8 |
| 2016 | 38.9 | 27.9 |
| 2017 | 39.7 | 28.1 |
| 2018 | 40.3 | 28.3 |
| 2019 | 41.1 | 28.4 |

*Source for education data: U.S. Census Bureau, Educational Attainment. 2011–2019. American Community Survey 1-Year Estimates.*

*https://data.census.gov/cedsci/table?q=S1501%3A%20EDUCATIONAL%20ATTAINMENT&tid=ACSST1Y2019.S1501*

*Copy and paste the link above into your browser to view the source data online.*

*Source for marriage data: U.S. Census Bureau, Median Age at First Marriage. 2011–2019. American Community Survey 1-Year Estimates.*

*https://data.census.gov/cedsci/table?q=B12007&tid=ACSDT1Y2019.B12007*

*Copy and paste the link above into your browser to view the source data online.*

## Item 2: Optional Instructions for Calculating *r* on a TI-84 Plus

Step 1: Turn on diagnostics to ensure that the *r* value will appear in the display when calculating a linear regression.

[2nd] [0] [$x^{-1}$] scroll down to DiagnosticOn [Enter] [Enter]

Should say "Done."

Step 2: Clear any previous data in $L_1$ and $L_2$, and then enter the *x* and *y* values from this activity in $L_1$ and $L_2$, respectively.

[Stat] [Enter] highlight $L_1$ [Clear] [Enter] highlight $L_2$ [Clear] [Enter]

Enter the X values in $L_1$ Enter the Y values in $L_2$

Step 3: Use the LinReg ($a + bx$) function to find the parameters for the line of best fit (linear regression), including *r*.

[Stat] Scroll to the right to Calc [8] [Enter] [Enter]

The last value listed is *r*.

## Item 3: Critical Values of *r* at a 5 Percent Significance Level

Critical values of *r* for a = .05

| *df* | *a*=.05 |
|---|---|
| 1 | .997 |
| 2 | .950 |
| 3 | .878 |
| 4 | .811 |
| 5 | .754 |
| 6 | .707 |
| 7 | .666 |
| 8 | .636 |
| 9 | .602 |
| 10 | .576 |
| 11 | .553 |
| 12 | .532 |
| 13 | .514 |
| 14 | .497 |
| 15 | .482 |
| 16 | .468 |
| 17 | .456 |
| 18 | .444 |
| 19 | .433 |
| 20 | .423 |

| *df* | *a*=.05 |
|---|---|
| 21 | .413 |
| 22 | .404 |
| 23 | .396 |
| 24 | .388 |
| 25 | .381 |
| 26 | .374 |
| 27 | .367 |
| 28 | .361 |
| 29 | .355 |
| 30 | .349 |
| 35 | .325 |
| 40 | .304 |
| 45 | .288 |
| 50 | .273 |
| 60 | .250 |
| 70 | .232 |
| 80 | .217 |
| 90 | .205 |
| ∞ | .195 |

*Note: These critical values pertain to a two-sided* t*-test.*