**Unit 7: Investigation 1 (2 Days)**

**Inference on Correlation and Regression**

**Common Core State Standards**

IC.A1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

IC.A.2 Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

**Overview**

This investigation introduces students to statistical inference on population correlation coefficients and population regression line slopes. Students compute and interpret sample correlation coefficients and least-square regression lines for small data sets, and then use the samples to perform randomization tests by constructing randomization distributions. The randomization distributions are obtained by permuting data values in the original samples (under the assumption that no correlation exists between the two variables). Students then use the randomization distributions to evaluate the likelihood of observing sample statistics as extreme as the ones observed. This activity involves the use of hands-on activities and simulated randomization distributions.

**Assessment Activities**

**Evidence of Success: What Will Students Be Able to Do?**

* Compute and interpret sample correlation coefficients and least-square regression lines for bivariate data
* Construct a randomization distribution of sample correlation coefficients using data from a random sample under the assumption that there is no correlation between the variables
* Construct a randomization distribution of sample regression lines slopes using data from a random sample under the assumption that there is no correlation between the variables
* Perform a randomization test to test a claim about a population correlation coefficient
* Perform a randomization test to test a claim about a population regression line slope

**Assessment Strategies: How Will They Show What They Know?**

* **Exit Slip 7.1.1** asks students to interpret a randomization distribution of sample correlation coefficients, identify a *P*-value, determine statistical significance, and draw a conclusion about a population correlation coefficient.
* **Exit Slip 7.1.2** asks students to interpret a randomization distribution of sample regression line slopes, identify a *P*-value, determine statistical significance, and draw a conclusion about a population regression line slope.
* **Journal Prompt 1** Explain how to construct a randomization distribution of sample correlation coefficients under the assumption that there is no correlation between the variables.
* **Journal Prompt 2** Describe why small probabilities (small *P*-values) lead us to reject assumptions made about population parameters.
* **Journal Prompt 3** Describe two plausible explanations that explain why a sample statistic would vary from an assumed value of the population parameter.
* **Activity 7.1.1 Introduction to Statistical Inference** is a hands-on activity that introduces students to the process of using a sample and sample correlation coefficient to make an inference about an unknown population correlation coefficient.
* **Activity 7.1.2 Inference on the Slope of a Linear Relationship** is a hands-on activity that introduces students to the process of using a sample and sample regression line slope to make an inference about an unknown population regression line slope.
* **Activity 7.1.3** **Inference Problems on Correlation and Regression** provides students additional opportunities to practice performing hypothesis tests for population correlation coefficients and regression line slopes.

**Launch Notes**

This investigation introduces statistical inference as the process of using a sample and sample statistic to make an inference about an unknown population parameter. This activity focuses on the prior learned concepts of correlation and regression in Algebra 1. You could begin this investigation by providing a quick refresher of these concepts by showing a few scatterplots and discussing their correlation coefficients and least-square regression lines. Two scatterplots are provided in **Activity 7.1.1 Supplement.** The scatterplots and sample statistics are shown below.

The first graph shows the total number of offensive plays and the total points scored of 15 National Football League (NFL) teams in the 2014 season. The second graph shows the grams of sugar per serving and ratings of 25 breakfast cereals.

|  |  |  |
| --- | --- | --- |
|  |  | |
| *r* = 0.46 | | *r* = -0.81 |

For each scatterplot, ask students to describe the correlation, estimate the slope of the regression line, and interpret the slope of the regression line. You may consider displaying different scatterplots that contain data that are of particular interest to students in your class. To reinforce linear function concepts, ask students to calculate the slope of the line of best fit by hand using two data points that would appear to lie on the line.

**Teaching Strategies**

1. **Activity 7.1.1 Introduction to Statistical Inference** is a hands-on activity that introduces students to the process of using a sample and sample correlation coefficient to make an inference about an unknown population correlation coefficient. Students explore foot length and arm length data from a small sample of high school students.

*Note to Teacher*: If time permits, consider using foot length and arm length data from students in your class. You could ask students to measure their foot and arm lengths at home or allow them an opportunity to do so in class.

Students are introduced to the terms: statistic, parameter, and statistical inference. Students learn the key components of a hypothesis test: making an assumption about a population parameter (hypothesis), collecting sample data, using probability to assess the likelihood of obtaining a sample statistic as extreme as the one observed under the assumption about the parameter, and making a decision about the population parameter.

The activity focuses on randomization tests using randomization distributions. Students create randomization distributions in class by permuting arm-length values to create new samples, and then calculating the resulting sample correlation coefficients. Students permute values by shuffling index cards. The randomization distribution is formed under the assumption that there is *no correlation* between the two variables. Students then use the distribution to assess the likelihood of observing a sample correlation coefficient as extreme as the one observed. Students are introduced to the term *P*-value and statistical significance.

*Notes about Randomization Tests* – A randomization test consists of four basic steps:

1. Make an assumption (hypothesis) about the value of a population parameter
2. Construct a randomization distribution of sample statistics under the assumption that the population parameter is equal to the hypothesized value
3. Find the probability of observing a sample statistic as extreme as the one found
4. Make a decision about the population parameter

On page 3 of the activity, students create a randomization distribution of sample correlation coefficients with a small number of simulated statistics. If time permits, present the creation of a simulated distribution of sample correlation coefficients using Statkey.

To create a simulated distribution of sample correlation coefficients:

* Go to Statkey (<http://lock5stat.com/statkey/index.html>)
* Click on Test for Slope, Correlation in the Randomization Hypothesis Tests section
* Click Edit Data and paste in two columns containing the sample data. The sample data can be copied from two columns in an Excel spreadsheet.
* Select Correlation where it says “Randomization Dotplot of”.
* Click Generate 1 sample (this will create a randomization sample, calculate the sample correlation coefficient, and plot the correlation coefficient), repeat nine times, and discuss the results
* Click Generate 10 samples and Generate 100 samples until the distribution has 500 sample correlation coefficients

Make sure students understand what the distribution represents, how it was constructed via randomization, and why the mean of the distribution equals 0 (since we assume no correlation exists between the two variables).

You can assign **Exit Slip 7.1.1** after students complete **Activity 7.1.1.**

**Group Activity**

Students are encouraged to work in pairs to complete **Activity 7.1.1** and **Activity 7.1.2**. Students could partition the responsibilities as follows: one student shuffles the index cards and the other student records the values in the sample. All students should practice calculating sample correlation coefficients and regression line slopes for the randomization samples.

**Differentiated Instruction (For Learners Needing More Assistance)**

Assist students to understand that sample correlation coefficients vary since samples vary. This may help to clarify the distinction between a population data set (and parameter) and a sample data set (and statistic). (1) Provide students with a small “population” data set and have them calculate the population correlation coefficient. (2) Randomly obtain multiple samples from the population and determine the sample correlation coefficients. (3) Compare the sample statistics to the population parameter and emphasize that the statistics are estimates of the parameter.

**Journal Prompt 1**  Explain how to construct a randomization distribution of sample correlation coefficients under the assumption that there is no correlation between the variables.

Students should state that we must randomly permute the values of one of the variables many times to create many randomization samples of paired data values. Each randomization sample has its own scatterplot and sample correlation coefficient. The many sample correlation coefficients form a distribution of sample correlation coefficients.

1. **Activity 7.1.2 Inference on the Slope of a Linear Relationship** is a hands-on activity that introduces students to the process of using a sample and sample regression line slope to make an inference about an unknown population regression line slope. The activity has a similar form to **Activity 7.1.1** and uses the same sample data.

Students create randomization distributions in class by permuting arm-length values to create new samples, and then calculating the resulting sample regression line slopes. Students permute values by shuffling index cards. The randomization distribution is formed under the assumption that there is *no correlation* between the two variables. Students then use the distribution to assess the likelihood of observing a sample regression line slope as extreme as the one observed.

On page 3 of the activity students create a randomization distribution of sample regression line slopes using the statistics computed in class. If time permits, present the creation of a simulated distribution of sample regression line slopes using Statkey.

To create a simulated distribution of sample regression line slopes:

* Go to Statkey (<http://lock5stat.com/statkey/index.html>)
* Click on Test for Slope, Correlation in the Randomization Hypothesis Tests section
* Click Edit Data and paste in two columns containing the sample data. The sample data can be copied from two columns in an Excel spreadsheet.
* Select Slope where it says “Randomization Dotplot of”.
* Click Generate 1 sample (this will create a randomization sample, calculate the sample regression line slope, and plot the regression line slope), repeat nine times, and discuss the results
* Click Generate 10 samples and Generate 100 samples until the distribution has 500 sample regression line slopes

Make sure students understand what the distribution represents, how it was constructed via randomization, and why the mean of the distribution equals 0 (since we assume no correlation exists between the two variables).

You can assign **Exit Slip 7.1.2** after students complete **Activity 7.1.2.**

**Differentiated Instruction (Enrichment)**

Encourage students to obtain a sample of bivariate quantitative data from the Internet, calculate and interpret the sample statistics (correlation coefficient, regression line), and use Statkey to perform a randomization test for the population parameters.

**Journal Prompt 2**  Describe why small probabilities (small *P*-values) lead us to reject assumptions made about population parameters.

Students should state that small *P*-values indicate that the observed sample statistic (actual sample correlation coefficient or actual sample regression line slope) is very unlikely to occur by chance alone. Or, students could state that a *P*-value is the probability of observing a sample statistic as extreme as the one observed, assuming the population parameter is a certain value, so a low *P*-value leads us to conclude that this assumption about the parameter is false.

**Activity 7.1.3** **Inference Problems on Correlation and Regression** provides students additional opportunities to practice performing hypothesis tests for population correlation coefficients and regression line slopes. This is intended as an out-of-class activity. The problems require students to have access to a computer to use Statkey to construct and explore randomization distributions.

**Journal Prompt 3**  Describe two plausible explanations that explain why a sample statistic would vary greatly from an assumed value of the population parameter.

The key term here is *assumed*, which means that we don’t know the actual value of the population parameter. Students should state that: (1) random chance could lead a sample statistic to vary greatly from the assumed population parameter since sample statistics vary, some a lot more than others; or (2) the population parameter could actually be different than the assumed value, in which case the sample statistic could be representative of the true population parameter value.

**Closure Notes**

On the final day of this investigation, have students complete problems that reinforce their understanding of the inference process introduced in this investigation. Emphasize that statistical inference is the process of using samples to draw conclusions about populations. Encourage students to explore problems in **Activity 7.1.3**, or to develop their own research question and obtain sample data from the Internet. Students should use Statkey to perform a randomization hypothesis test on the sample data and could prepare a short PowerPoint presentation to share their results with the class.

**Vocabulary**

Categorical variable

Correlation

Correlation coefficient

Data

Distribution of sample correlation coefficients

Distribution of sample regression line slope

Hypothesis

Hypothesis test

Intercept

Least-squares regression line

Linear association

P-value

Parameter

Population

Quantitative variable

Randomization distribution

Randomization sample

Randomization test

Regression line

Sample

Slope

Statistic

Statistical inference

Statistically significant

**Resources and Materials**

**All activity sheets 7.1.1. – 7.1.3 should be completed. Note: Activity 7.1.3 is intended as an out-of-class activity.**

Activity 7.1.1 Introduction to Statistical Inference

Activity 7.1.1 Supplement

Activity 7.1.2 Inference on the Slope of a Linear Relationship

Activity 7.1.3 Inference Problems on Correlation and Regression

Census at School (<http://www.amstat.org/censusatschool/index.cfm>)

Statkey (<http://lock5stat.com/statkey/>)

Index cards (10 per student)

Graphing calculator