**Activity 7.1.1 – Introduction to Statistical Inference**

*Statistics* is the science of collecting, analyzing and interpreting data. *Data* are pieces of information that are collected on variables of interest. Variables can be divided into two types: *quantitative* and *categorical*. A quantitative variable has values that are numerical counts or measurements. A categorical variable has values that can be separated into categories or groups.

**Examining a Bivariate Relationships: Arm length vs. Foot Length**

The table and scatterplot below display the arm length and foot length data for a random sample of nine high school students in a local high school.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Foot length (cm) | Arm length (cm) | | 24 | 164 | | 24 | 166 | | 24 | 171 | | 25.5 | 179 | | 24 | 175 | | 22 | 156 | | 21.5 | 161.5 | | 28 | 181 | | 20.32 | 172 | |

1. Summarize the sample data by calculating the correlation coefficient and the slope and

*y*-intercept of the least-squares regression line. Round your answers to two decimal places.

1. Interpret the sample statistics. What do they tell about the relationship between these two variables?

Inference question:Does the sample provide evidence that a linear relationship exists between foot length and arm length for *all* students in the high school?

*Statistical inference* is the process of using sample statistics to draw conclusions about population parameters. *Sample statistics* are numerical descriptions of sample characteristics. *Population parameters* are numerical descriptions of population characteristics. When population parameters are unknown we use sample statistics to make inferences about population parameters.

The population correlation coefficient, (Greek letter rho, pronounced as “row”) is a parameter. It is the correlation coefficient for foot-length and arm-length data from *all* students in the high school. This parameter is unknown. By conducting a *randomization test*, we can use the sample correlation coefficient *r* to make an *inference* about the unknown population correlation coefficient . We do so as follows.

**Population:**

All students in the high school

**Parameter:**

**Sample:**

9 high school students

**Statistic:**

Randomization Test for Population Correlation Coefficient

* We assume . That is, we assume there is no linear association between foot length and arm length in the population.
* We treat the sample like a population, assuming there is no relationship between foot length and arm length, and generate *randomization samples*.
* We find the likelihood of getting a sample correlation coefficient as extreme as the one we found by chance alone.

1. Your instructor has given you nine index cards. Write the arm length values on the cards. Write one value on each card. The arm lengths from the random sample are {164, 166, 171, 179, 175, 156, 161.5, 181, 172}.
2. Shuffle the index the cards well and stack the cards into a single pile. Then, in the random order that the arm lengths appear, copy them into the table below. Notice that the arm lengths have been shuffled while the foot lengths have remained in the same order. Calculate *r* for this randomization sample. Round *r* to one decimal place.

Randomization Sample # 1

|  |  |
| --- | --- |
| Foot length (cm) | Arm length (cm) |
| 24 |  |
| 24 |  |
| 24 |  |
| 25.5 |  |
| 24 |  |
| 22 |  |
| 21.5 |  |
| 28 |  |
| 20.32 |  |

1. Repeat the process. Shuffle the index the cards well and stack the cards into a single pile. In the order that the arm lengths appear, copy them into the table below. Calculate *r* for this randomization sample. Round *r* to one decimal place.

Randomization Sample # 2

|  |  |
| --- | --- |
| Foot length (cm) | Arm length (cm) |
| 24 |  |
| 24 |  |
| 24 |  |
| 25.5 |  |
| 24 |  |
| 22 |  |
| 21.5 |  |
| 28 |  |
| 20.32 |  |

1. Create a dotplot of all sample correlation coefficients obtained from randomization samples in the class.

The dot plot above is a *randomization distribution.* It was formed under the assumption that there is *no relationship* between foot length and arm length. The *variability* in the sample correlation coefficients is due solely to random chance.

1. What do you notice about the distribution of correlation coefficients from the randomization samples?
2. Does the distribution of sample statistics appear to come from a population in which there is no association between foot length and arm length?
3. Assuming there is no association between foot length and arm length (i.e. ), what is the probability of getting a sample correlation coefficient greater than or equal to the one we found? Use the randomization distribution to answer this question. This probability is called a *P*-value (probability value).

A *P*-value is the probability of obtaining a sample statistic as extreme as the one observed assuming the population parameter is equal to a specific value.

* When a *P*-value is less than 5%, we say the sample statistic is *statistically significant*. We *reject* the assumption about the population parameter. When this occurs we find that the sample statistic did not occur solely due to chance.
* When a *P*-value is greater than or equal to 5%, we say the sample statistic is *not statistically significant*. We *do not reject* the assumption about the population parameter. When this occurs we find that the sample statistic could have occurred solely due to chance.

1. Is the observed sample correlation coefficient statistically significant? Explain.
2. What inference can we make about the population correlation coefficient ? Does the sample provide evidence that a linear relationship exists between foot length and arm length for *all* students in the high school?