Exploration 10.1-2: Tracking Activity and Physiological Parameters

PART 1: Two Quantitative Variables: Scatterplots and Correlation

LEARNING GOALS

- Recognize that a scatterplot is the appropriate graph for displaying the relationship between two quantitative variables and create a scatterplot from raw data.
- Summarize the characteristics of a scatterplot by describing its direction, form, and strength, as well as whether there are any unusual observations.
- Recognize that a correlation coefficient of 0 means that there is no linear association between the two variables and that a correlation coefficient of –1 or 1 means that the scatterplot is exactly a straight line.
- Estimate the value of the correlation coefficient within ± 0.3 by looking at a scatterplot.
- Recognize that the correlation coefficient is appropriate only for summarizing the strength and direction of a scatterplot that has linear form.
- Understand that the correlation coefficient is not resistant to extreme observations.
- Recognize how the association between two variables may change when data are split into smaller groups.

At one time most heath indicators such as blood pressure, blood oxygen level, or heart rate were just collected in a physician's office or a hospital, perhaps only on an annual basis. Now these indicators can be collected almost continuously by wearing one of many smart watches or fitness trackers that are on the market. Does this give a person, or their doctor, useful information to help them catch diseases earlier? Does tracking these health indicators along with tracking physical activity help people lead healthier lives? Researchers (Li et al., 2017) tracked 43 people ages 35 to 70 years old for an average of about 5 months collecting more than 1.7 million various health measurements in the process. For 38 of these individuals, the researchers analyzed their average resting heart rate and average number of steps taken per day.

1. Are the variables of average resting heart rate and average number of steps per day categorical or quantitative?

2. In your own words, explain what it would mean for there to be an association between the explanatory and response variables in this study.

Scatterplots

This study is different from those we've looked at before because both of the variables of interest are quantitative. For this reason, we need new graphical and numerical techniques to summarize the data.

Key Idea

A ***scatterplot*** is a graph showing a dot for each observational unit, where the location of the dot indicates the values of the observational unit for both the explanatory and response variables. Typically, the explanatory variable is placed on the *x*-axis and the response variable is placed on the *y*-axis.

Paste the **StepsHR** data set into the **Corr/Regression** applet. The variable HR is the average resting heart rate (in beats per minute), and the variable Steps is the average daily steps in thousands. We want the explanatory variable to be the average steps per day and the response to be the average resting heart

rate. You can reverse those roles by toggling the **Explanatory, Response** button or using the pull-down menus next to **Response Variable** and **Explanatory Variable.**

3. Click on the dot farthest to the left and highest up on the graph. You should see information about this dot show up in red below the graph. What specifically do these numbers mean?

When describing a scatterplot we look for three aspects of association: direction, form, and strength. The ***direction*** of association between two quantitative variables is either positive or negative, depending on whether the response variable tends to increase (positive association) or tends to decrease (negative association) as the explanatory variable increases.

4. **Direction.** Is the association between steps and heart rate positive or negative? Interpret what this means in context.

The ***form*** of association between two quantitative variables is described by indicating whether a line would do a reasonable job summarizing the overall pattern in the data or some other pattern such as a curve would be better. It is important to note that, especially when the sample size is small, you don't want to let one or two points on the scatterplot change your interpretation of whether or not the form of association is linear. In general, assume that the form is linear unless there is compelling (strong) evidence in the scatterplot that the form is not linear.

5. **Form.** Does the association between steps and heart rate in this sample appear to be linear or is there strong evidence (over many observational units) suggesting the relationship is nonlinear?

In describing the ***strength*** of association revealed in a scatterplot, we see how closely the points follow the form: that is, how closely do the points follow a straight line or curve. If all of the points fall pretty close to a straight line or curve, we say the association is strong. Weak associations will show little pattern in the scatterplot, and moderate associations will be somewhere in the middle.

6. In your opinion, would you say that the association between steps and heart rate appears to be strong, moderate, or weak?

It is also important to consider any ***unusual observations*** that do not follow the overall pattern.

7. Are there any observational units (dots on the scatterplot, representing individual subjects) that seem to fall outside of the overall pattern? If yes, identify it/them. If not, provide a hypothetical example of an average number of steps and average resting heart rate combination which would represent an unusual observation.

Note: There are two kinds of unusual observations seen in scatterplots: influential observations and outliers. An observation is *influential* if removing the observation from the data set dramatically changes our perception of the association. Often, influential observations tend to fall on the far right or left on the scatterplot with a response that is in the opposite direction of the association. *Outliers* are observations that don't follow the overall pattern of the relationship. Outliers may or may not be influential and they may or may not be extreme in either variable individually but are unusual in terms of the combination of values.

Key idea

The association between quantitative variables can be described with direction, form, and strength.

- If above-average values of one variable tend to correspond to above-average values of the other variable, the ***direction*** is positive. If, however, above-average values of one variable are associated with below-average values of the other, the ***direction*** is negative.
- The ***form*** of the association is whether the data follow a linear pattern or some more complicated pattern.
- The ***strength*** of an association refers to how closely the data follow a particular pattern. When an association is strong, the value of one variable can be accurately predicted from the other.
- Also look for, and investigate further, any ***unusual observations*** that do not fit the pattern of the rest.

Numerical summaries

Describing the direction, form, and strength of association based on a scatterplot, along with investigating unusual observations, is an important first step in summarizing the relationship between two quantitative variables. We can also use a statistic to summarize the association. One of the statistics most commonly used for this purpose is the correlation coefficient, which measures the strength and direction of the *linear* association.

The ***correlation coefficient***, often denoted by the symbol ***r***, is a single number that takes a value between –1 and 1, inclusive. Negative values of *r* indicate a negative linear association, whereas positive values of *r* indicate a positive linear association. The stronger the linear association between the two variables, the closer the value of the correlation coefficient will be to either –1 or 1, whereas weaker linear associations will have correlation coefficient values closer to 0. Moderate linear associations will typically have correlation coefficients in the range of 0.30 to 0.70 or –0.30 to –0.70.

8. Will the value of the correlation coefficient for the Steps-HR data be negative or positive? Why?

Key idea

The correlation coefficient uses a rather complex formula that is rarely computed by hand; instead, people almost always use a calculator or computer to calculate the value of the correlation coefficient. But you should be able to apply the above properties to interpret the correlation coefficient that is found.

9. Without using the applet, give an estimated value of the correlation coefficient between HR and Steps based on the scatterplot.

10. Now, check the **Correlation coefficient** box in the applet to reveal the actual value of the correlation coefficient. Report the value.

Key idea

The correlation coefficient is only applicable for data which has a linear form; nonlinear data are not summarized well by the correlation coefficient.

The correlation coefficient is also sensitive to influential observations. Earlier we said that removing an influential observation dramatically changes our perception of the association. In particular, removing an influential observation can substantially change the value of the correlation coefficient.

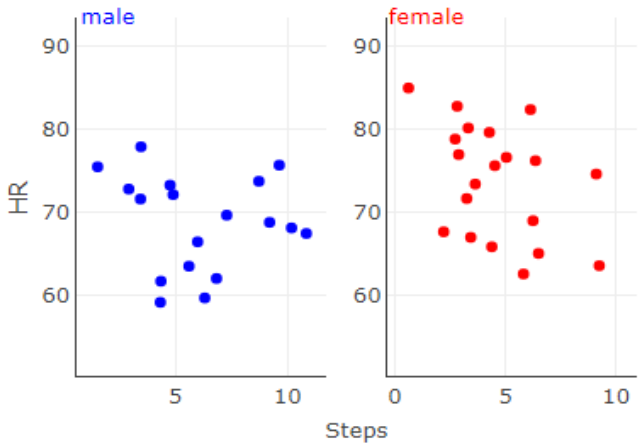
11. Check the box for **Show data options** then check the box for **Move observations**. Now put the cursor on the point farthest to the right in the graph and slowly slide it upwards. What happens to the value of the correlation coefficient as you slide this point up? Can you slide it up far enough so that the correlation coefficient is positive? Would you consider this new moved point an influential observation?

Press the **Revert** button to revert the data to its original form. You will notice that there is a third variable of sex included in the dataset. Under the **Change variable selections**, in the pull-down menu for **Color by**, select Sex. This will include a third variable (this time categorical) into the graph by making each male observational unit blue and each female observational unit red.

12. How do the female heart rates tend to compare to the male heart rates?

13. Does the association between heart rate and steps seem to be similar or different between males and females? Explain.

The two graphs below show separate scatterplots for the male and female observations. This might make it easier to see and compare the two associations.



14. Both the male and female data have negative associations. Which one appears to be stronger? What do you predict for the value of the correlation coefficient between HR and Steps for females only? Why? What do you predict for the males only?

The correlation coefficient for females is -0.423 and for males is -0.114. The correlation coefficient for both groups combined is -0.350 (a number between the two individual correlations). This won't always happen. You could have two data sets (both with negative correlation coefficients) that when combined result in a data set with an even stronger negative association than when they were split. The correlation could also even become positive. Almost anything is possible which is why it is helpful to explore the data both combined and separated to gain a better understanding.

PART 2: Inference for the Correlation Coefficient: Simulation-Based Approach

LEARNING GOALS

- Apply the 3S strategy when evaluating the hypothesis of linear association using the correlation coefficient as the statistic.
- Articulate how to conduct a tactile simulation to implement the 3S strategy for testing a correlation coefficient.
- Define the p-value in the context of the 3S strategy using simulated correlation coefficients under the null hypothesis of no association.

15. If average number of steps per day and average resting heart rate were not associated in the population, what should be the value of the correlation coefficient between these two variables.

16. Remember that the sample correlation coefficient for the data set is $r = -0.350$. Let's think about how we would complete a test of significance for the population correlation coefficient.

    a. Suggest two possible explanations (hypotheses) which could have generated the nonzero value of the correlation coefficient. (*Hint*: These are very similar to the two possible hypotheses you've seen many times before.)

    b. In your own words, how could we go about determining whether random chance is a plausible explanation for the observed correlation value between average number of steps per day and average resting heart rate? Explain how the 3S strategy could be applied here; in particular, identify a simulation strategy you could conduct "by hand." *Note:* You do not need to actually carry out a simulation analysis.

        i. What is the statistic?

        ii. How would you simulate could-have-been results?

        iii. How would you evaluate the strength of evidence against the null hypothesis?

The null hypothesis to be tested here is that there is *no* association between average number of steps per day and average resting heart rate. The alternative hypothesis is that there is an association between average number of steps per day and average resting heart rate.

How can we assess whether the observed correlation coefficient of $r = -0.350$ is far enough from zero to provide convincing evidence that there is an association in the population? Like always, we ask how unlikely it would be to have a random sample produce a correlation coefficient value as far from zero as –0.350 if there is no association between our variables in the population. Our simulation approach will generate a large number of sample results assuming no underlying association (any *y*-value in the dataset can be paired with any *x*-value) , calculating the correlation coefficient for each one, and seeing how often we obtain a correlation coefficient as or more extreme (as far from zero) as –0.350.

17. Continuing with the **Corr/Regression** applet (you can return to No color), check the **Show Shuffle Options** box and select **Correlation** from the pull-down menu by **Choose statistic**. Then press **Shuffle Y-values** to simulate one result assuming there is no association between the two variables.

a. Describe how the scatterplot of the simulated results reflects no association between the two variables.

b. Record the value of the correlation coefficient between the shuffled heart rates and steps.

c. Press **Shuffle Y-values** four more times to generate results of four more random shuffles of HR values to Steps values. Record the values of the shuffled correlation coefficients in the table below.

Repetition	**1**	**2**	**3**	**4**	**5**
Correlation coefficient					

d. Did any of your simulated statistics from assuming no association produce a correlation coefficient as extreme (far from zero) as the observed –0.350?

e. Change the **Number of Shuffles** from 1 to 995 and press **Shuffle Y-values** to produce 995 more simulated results. Look at the null distribution of these 1,000 simulated correlation coefficients. Approximately where is this null distribution centered? Why does this center make sense?

f. Next to the **Count Samples** box choose **Beyond** from the pull-down menu. Specify the observed correlation coefficient (–0.350) and press **Count**. What proportion, of the 1,000 simulated random results produced a correlation coefficient as extreme (as far from zero in either direction) as –0.350? Report the approximate p-value.

g. Interpret this p-value: This is the probability of what, assuming what?

h. What conclusion would you draw from this p-value? Do you have strong evidence that there is an association between average number of steps per day and average resting heart rate? Explain the reasoning behind your conclusion.

18. Generalization and Causation.

a. To what population are you willing to generalize these results?

b. Can you state that a change in the average daily number of steps causes a change in the average resting heart rate?

Note: Keep in mind that the correlation coefficient measures the strength of the linear association. This test may not find significance when a different type of association exists.

Reference

Li X, Dunn J, Salins D, Zhou G, Zhou W, Schüssler-Fiorenza Rose SM, et al. (2017) Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. PLoS Biol 15(1): e2001402. https://doi.org/10.1371/journal.pbio.2001402