PLTW COMPUTER SCIENCE

Activity 3.2.1

Inferential Statistics

Introduction

Patterns become visible when you display data in a visualization. Is the pattern just random chance, or is it scientific knowledge that you have discovered in the data? Inferential statistics calculate the likelihood that you just have a coincidence on your hands.

The proof is in the pudding.

Materials

- Computer with Canopy distribution of Python® programming language and access to Internet

Resources

3.2.1 source Files

Procedure

Part I: Comparing Proportions

In this activity you will use *Python* to visualize simulated data and make scientific observations of the model's data.

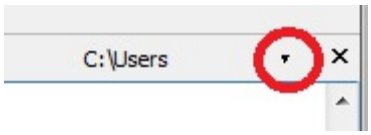
1. Form pairs as directed by your teacher. Meet or greet each other to practice professional skills. Set team norms.
2. A model is a mathematical description of something real. The model is an abstraction, losing some of the details of the real thing. For example, the Pew Research Center found that 72% of both high school sophomores and juniors sometimes text via a mobile device. If you look at the real poll results, you might discover additional patterns: a difference in texting between

males and females, between students who work and students who don't, and so on. But a simple model could abstract away those details and keep only the pattern that 72% of students text and 28% do not.

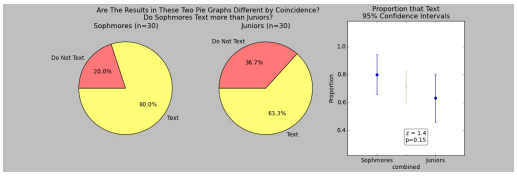
What is another pattern that might have been lost by abstracting the real data to a model in which 72% of students text?

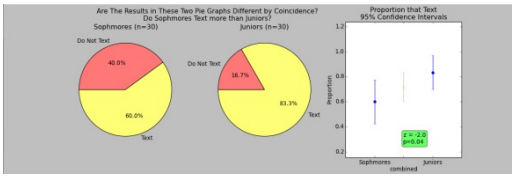
3. In a simulation, an algorithm implements a model. The simulation is executed to generate meaningful but fictional data. We will use modeling and simulation in this activity to understand how data are related to reality. We can simulate results from the poll about texting by using a model in which each student has a 72% chance of saying they text and a 28% chance of saying they don't text. Because of the randomness built into this model, the result of each run of the simulation will be different.

   In Canopy open the file `differenceBetweenProportions.py`. Change the working directory of Canopy to the directory containing this file. This can be done by clicking the arrow in the upper right corner of the iPython window and selecting **Change to Editor Directory**.



   Execute the program two or three times and examine the matplotlib window that appears each time. Two examples are shown here.





   A model's **parameters** are adjustable values that control how the model works. The parameters of this model are shown here.

```
number_group1 = 30 # Data points in sample 1
number_group2 = 30 # Data points in sample 2
p1_hat, p2_hat = [0.72, 0.72] # Portion answering "yes" in each grc
```

   Using pseudorandom numbers in a simulation is sometimes called **Monte Carlo simulation**. The following two statements are both true:

   - The model assumes that sophomores and juniors are both 72% likely to text.
   - The simulation often generates data in which a much higher or much lower percentage than 72% of students text.

   The code for `two_categories()` in `cse.montecarlo.py` (in the cse folder in 3.2.1 sourceFiles) acts a lot like flipping a coin multiple times. A coin lands heads 50% of the time, but not in any given sample. How can the two statements above simultaneously be true?

4. Just because a survey found that a greater percentage of sophomores or juniors text doesn't mean that there is enough evidence to think that the groups are actually different. Inferential

statistics calculates the likelihood of various claims about reality based on a set of measurements. Scientists report inferences about the real value of something that was measured using a 95% **confidence interval**, an interval with enough wiggle room that such inferences will be correct 95% of the time. The **p-value** tells the probability that an observed pattern has arisen because of chance and not because of an actual pattern in reality. Most scientific journals will only publish results with p<0.05.

Run `differenceBetweenProportions` repeatedly until you see evidence that more sophomores text than juniors, or vice versa. The evidence will be indicated by a p-value less than 0.05, in which case the program will shade it green. Run the program and view the results repeatedly until you (or any pair near you– it will take 20 times, on average) obtain the green-shaded result. Examine that result and record the data below.

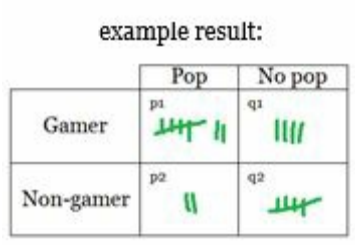
	Sophomores	Juniors
% of sample who text	_____	_____
% of population who text (95% confidence interval)	_____ to _____	____ to ____

5. Computing is a powerful tool to automate data collection, visualization, and analysis. The code `differenceBetweenProportions`, for example, can automatically visualize and analyze real data about proportions on any topic. As an example, follow these steps.
   - Choose two binary questions as a class that you think might be associated. One pair of questions you could use is given here.

     Do you drink pop/soda once per week or more?

     Do you play video games once per week or more?
   - Have each person in the class ask themselves the two questions picked by the class and mark one tally per person in a 2x2 grid similar to this one.

	p1	q1
	p2	q2



   - Modify `differenceBetweenProportions` using your questions and results following this example. Note the line numbers.

```
p_and_q = ['Drink Pop', 'Drink No Pop'] # [p-label, q-label]
```

```
treatments = ['Gamers','Non-Gamers'] # Sample 1 and 2

simulate = True # Change to False if providing your own data

number_p1 = 7 # number in group 1 who said "p"
number_q1 = 4 # number in group 1 who said "q"
number_p2 = 2 # number in group 2 who said "p"
number_q2 = 5 # number in group 2 who said "q"
```

- Execute the program and save the resulting figure as directed by your teacher. Describe what you think can be inferred from your data.

6. All statistics tests make certain assumptions that must be met. To compare proportions, there must be at least 5 tally marks in each of the cells in the 2×2 grid. The inferences about gamers and pop would therefore not be valid using the example data above, but computation makes it easy to collect larger samples. Brainstorm as a class: How could networked computing allow you to collect data from a large sample of people answering your two questions?
7. Consider all the forms of automated and crowd-sourced data collection, including information mined from debit cards, mining social media, etc. Describe how a person with access to all already-existing data might be able identify data for the two questions you chose above.
8. Patterns are detected by **disaggregating data**, which means taking measurements of multiple variables and then pulling the data set apart based on one or more of those variables. By disaggregating a data set, associations can be discovered between the variable(s) used to pull apart the sample, and variables that you then look at within each sample.

   This part of the activity gave one example of disaggregation. If the simulated people surveyed about texting were each represented on one row of a text file, then the top few rows of the text file might look like this.

   ```
   Grade,Texts per day
   11,0
   10,7
   10,4
   ```

   The first row of data shown here says an 11th grader does not text. Describe the algorithm you would use to read this text file and generate the four numbers needed for the frequency table shown below. You could use natural language, pseudocode, or a programming language.

	Sophomore	Junior
Texts	p1	q1
Does Not Text	p2	q2

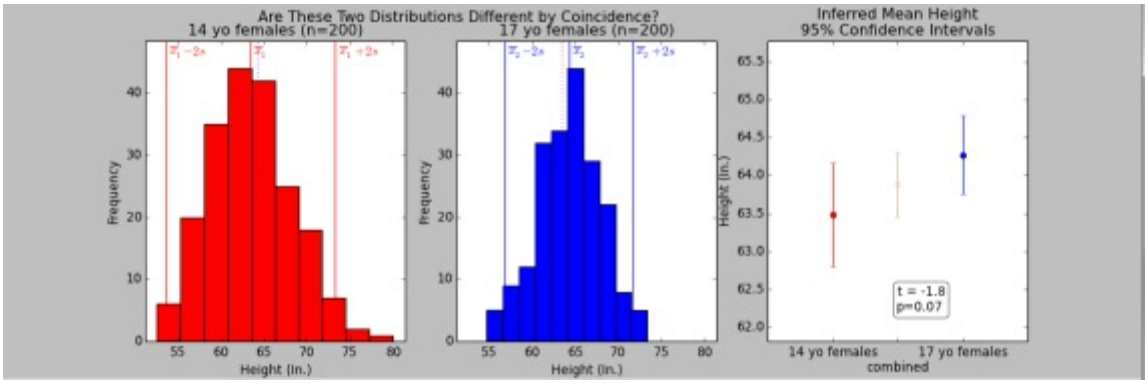
9. Just because two variables are associated does not mean there is a causal relationship between them. In the example we used, even if data led us to claim that drinking pop is

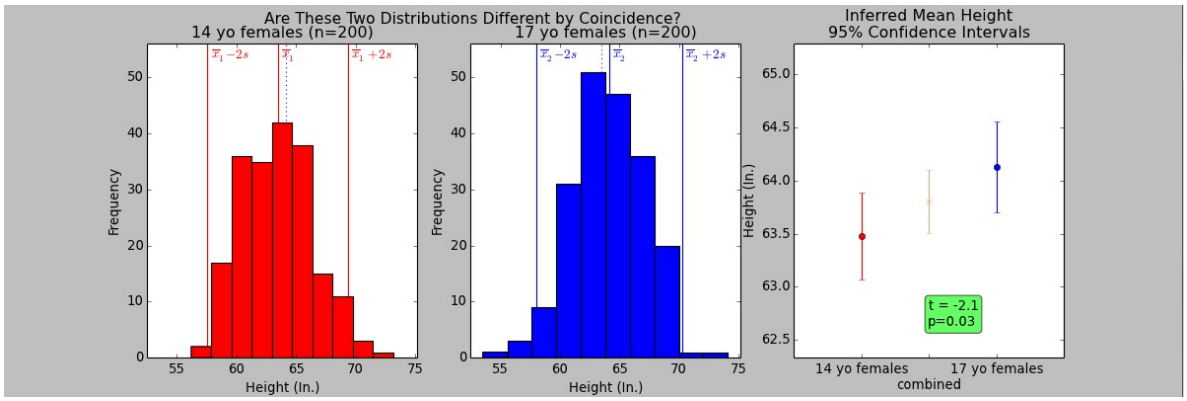
associated with playing video games, we must not infer that either one causes the other. Instead, a third variable might cause both. What would be a possible **common causality** in this case – a third characteristic that would cause people to drink pop and play video games?

Part II: Comparing Means

10. In the last part, you analyzed categorical variable vs. categorical variable. We made texts per day and grade level into variables with only two categories: text/no text, sophomore/junior. If we used either grade level or texts per day as a quantitative variable, we would have had a quantitative variable vs. a categorical variable. For that, we would use histograms or box plots.

    Open `differenceBetweenMeans.py` in Canopy. This visualization compares the heights of 14- and 17-year old females. The sample data is randomized, and therefore produces different results whenever it is run.





    Run `differenceBetweenMeans` a few times and observe the different results you get. Save a copy of one visualization as directed by your teacher and describe its abstract meaning (i.e., comparison of measures of center and spread, rather than detailed data) as well as you can. Recall that x-bar and s ($\bar{x}$ and s) represent the sample mean and standard deviation.

11. The simulation uses model parameters for teenage girls' heights based on U.S. government measurements. The parameters are shown below.

```
mu1, sigma1 = [63.6, 2.9] # 14 yo females
mu2, sigma2 = [64.2, 3.3] # 17 yo females
```

How much difference is there between the heights of 14- and 17-year old females in the model? Recall that mu and sigma (μ and σ) represent the population mean and standard deviation.

12. What is the pattern embedded in the model?
13. How much difference is there between the heights of 14- and 17-year old females in the sample produced by the simulation?
14. Run the simulation repeatedly. From the runs of the simulation you observed, what portion of them detected evidence ($p<0.05$, shaded green) for the pattern that was embedded in the model?
15. In lines 37-38 of the program, increase by a factor of 10 the sample sizes `n1` and `n2`, which are the number of simulated people in each group whose heights are being modeled.

    - The sample sizes were 200 in each group. After increasing `n1` and `n2` by a factor of 10, how many data points will be in each group?

    - Describe the increase in time you think will be required to run the simulation and visualization with the larger sample size.

    - Run the simulation with the larger sample size several times. What portion of them detected evidence ($p<0.05$, shaded green) for the pattern that was embedded in the model?

    - Consider the following two facts.

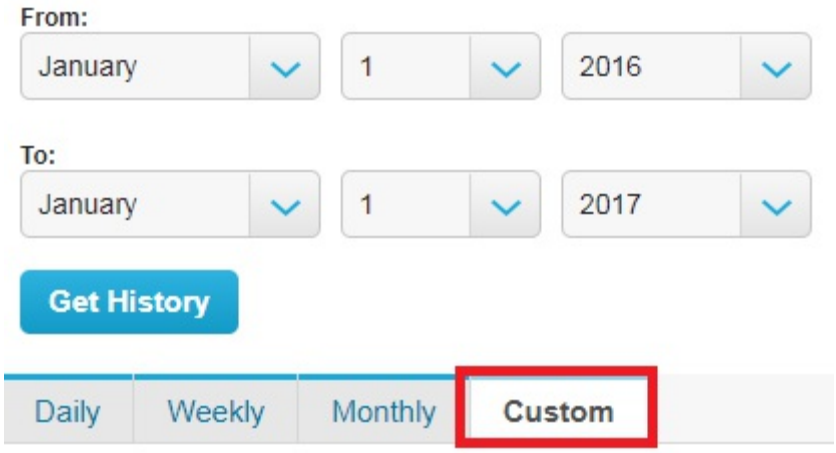
        - The **Law of Large Numbers** says that larger sample sizes are more likely to have the same mean as the population being sampled.
        - Larger samples make it easier to detect a pattern.

      Compare your answer to Step c here with your answer to Step 14 above. Describe what you observe in that comparison, and explain how the two facts above are related to your observation.

16. Computing is a powerful tool to automate data collection, visualization, and analysis. The code `differenceBetweenMeans`, for example, can automatically visualize and compare two distributions of any quantitative variable. As an example, follow these steps.

    - Spend a few moments familiarizing yourself with data offered at a weather data webiste: **http://www.wunderground.com/history/**

        - **Submit** a zip code.
        - Select the **Custom** tab as shown below.
        - Enter a date range as shown below and select **Get History**.



- Automating the collection of data presented on web pages is web scraping. The Python libraries `requests` and `lxml` can automate web scraping. We'll settle for some data points copied manually (by typing). Pick a quantitative variable from the web site and record many values of the variable for two categories (e.g. precipitation amounts for July days with high temperatures above and below 80°F.).

- Modify `differenceBetweenMeans` using the data you scraped from the web. Follow the example provided below. Note the line numbers.

```
measurement_variable = 'Precipitation'
measurement_unit = 'In.'

treatments = ['July Days > 80F','July Days <= 80F']

simulate = False # Change to False if providing your own data

sample1 = [0.25, 0.125, 0.875, 0, 0, 0]
sample2 = [0,0, 0.25, 0,0.75, 0.25, 0, 0]
```

- Execute the program and save the resulting figure as directed by your teacher. Your data likely do not allow inference from the statistics as depicted in the matplotlib figure because the statistics require 30 values in each category or normally distributed data. Our focus, however, is to understand patterns that can be discovered and how and to understand the power that automating data collection can provide. Describe what you think could be inferred from your data if the confidence intervals and p-value were valid.

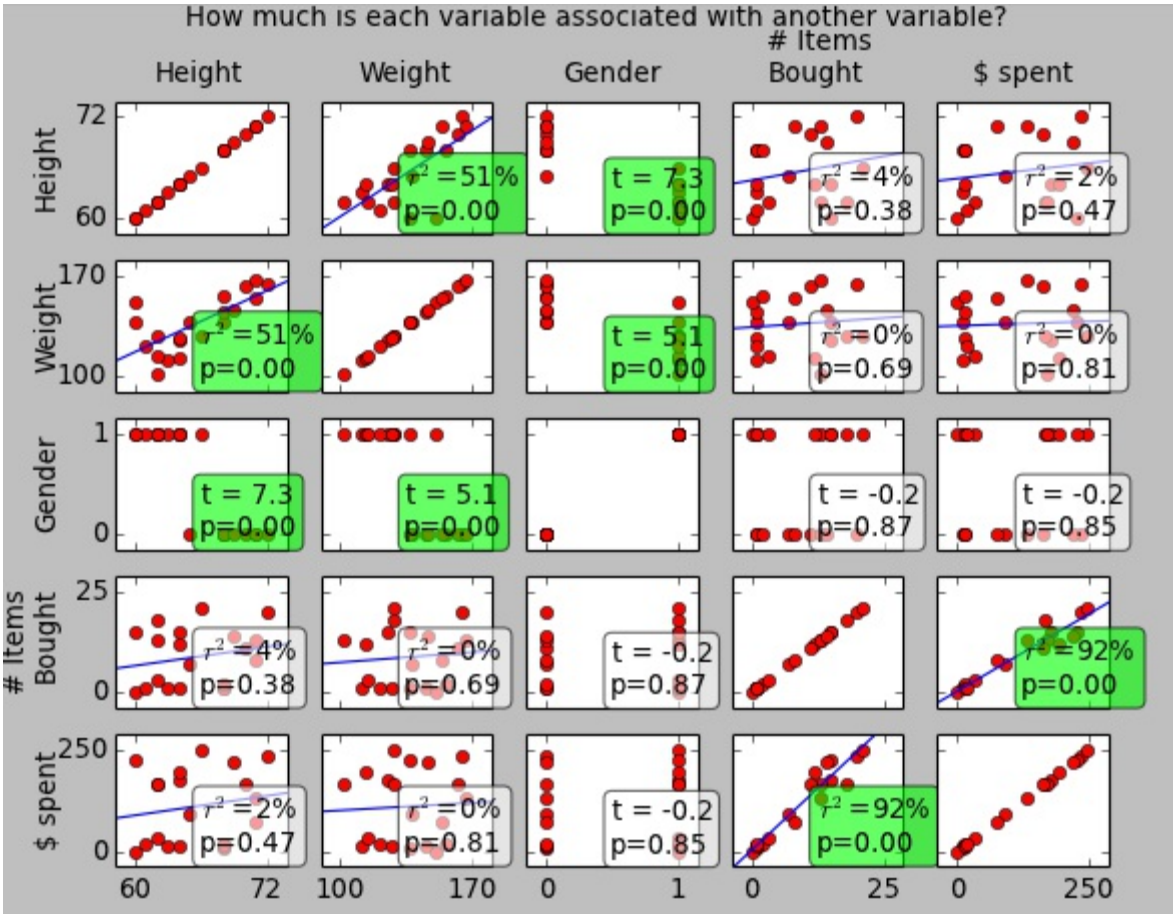
Part III: Linear Correlation

17. In the last two parts of this activity, you examined data with only two variables and asked if the two variables were associated. If there were 10 variables in the data set, how many associations could you ask about?
18. When comparing age and height, you treated age as a categorical variable with two categories: 14-year-old and 17-year old. We compared two histograms to analyze a

quantitative variable vs. categorical variable. If *both* variables are quantitative, we use a scatter plot and ask if the variables are associated by a linear correlation, which is when the points showing one variable vs. the other lie at least roughly along a line.

Open `linearCorrelation.py` in Canopy and execute it. The graphic that is produced visualizes the relationships among five variables in a simulated data set: the height, weight, and gender of customers along with the number and dollar value of items purchased. An example is below. Save the graphic produced from your run of the simulation as directed by your instructor.

The five-by-five grid asks if each variable is associated with each of the other variables. One of the variables is a categorical variable with two categories, and the visualization is comparing two means. Which variable is categorical, and which other variables (according to your data) is it associated with?



19. In the graphic above, there is evidence ($p<0.05$) that taller people weigh more. There is no evidence ($p>0.05$) that taller people buy more items in this store. Describe one other correlation for which there is evidence, and describe one other correlation for which there is not evidence.
20. In the graphic above, 51% of the variation in people's weight can be predicted if you know their height, but 92% of a person's purchase in dollars can be predicted by knowing how many items they buy. How is this related to the appearance of the weight vs. height and dollars vs. items plots?
21. Open `plot_age_income.py` in Canopy and execute it. Can you tell if there is evidence that household income is linearly correlated with householder age in the United States? Explain.
22. Open `plot_age_income_subset.py` in Canopy and execute it. Is there evidence that

income is linearly correlated with age for this subset of householders? Explain.

Conclusion

1. During exploratory data analysis, we search for patterns in a data set. Why are statistics necessary to confirm patterns that are observed?
2. If a pattern is discovered in data and then confirmed by statistics, what additional evidence, if any, is needed to count the pattern as “knowledge?”