Data 8 Mastery Section

April 17, 2017

Hypothesis Testing

1. Rival researchers claims that the data have been falsified! They believe that in fact every student gave at least one answer on Piazza, but the researchers flipped a fair coin for each student and only recorded his/her number of answers if it came up tails. For heads, they just wrote 0 answers.

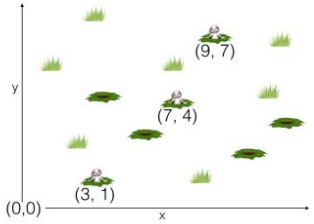
How would you determine whether the evidence in the students table is consistent with this controversial claim?

a) What hypothesis would you evaluate?

b) What test statistic would you use?

c) How would you compute the probability distribution of the test statistic under the hypothesis?

d) What observation would you compare to this probability distribution?



2. Suppose we are scientists who study rabbits (bunny biologists), and we're interested in the burrowing behavior of the rabbits in a field. Specifically, we would like to know the smallest distance between any of the rabbit holes, among all pairs of holes. The holes are hidden, so we can't see them directly, but we can see the rabbits when they poke their heads out. By spraying the field with a carrot-based perfume, we entice some of the rabbits to poke their heads out of their holes, and we note the locations of those holes. Suppose these locations constitute a random sample without replacement of 30 of the hole locations, out of 100 total holes. We record these locations in a table called hole location sample. It has two columns: x, the x coordinate of each hole, and y, the y coordinate of each hole. The distance between two holes can be calculated using their coordinates and the Pythagorean theorem.

(a) Using only our sample, what would be an appropriate statistic to use as an estimate of the smallest distance between any two rabbit holes in the field? (For example, "the average distance between holes in the sample" would be a (poor) estimate.)

(b) If you wanted to know how well your estimate worked, you could repeat our sampling process 1,000 times and make a histogram of the resulting 1,000 estimates. What would that histogram be called?

(c) Suppose the smallest distance between any two holes in the field is 10 meters. Would you expect the histogram in (b) to have its mean at roughly 10 meters, or less than 10 meters, or more than 10 meters? Draw what you think it would look like.

(d) As Patrick Star once said, "We're not cavemen! We have technology!" Suppose we have access to all 100 of the hole locations in a table called hole locations, and a function called estimate smallest distance that computes your proposed estimate on 1 sample. Write code that will generate the histogram in (b).

P-values

1. Fill in the blanks for the following statements.

   The p-value is the _____, under the _____ _____, that the _____ _____ is equal to the value that was observed in the data or is even further in the direction of the _____.

2. If you use a p% cutoff for the P-value, and the _____ _____ happens to be true, then there is about a _____% chance that your test will conclude that the _____ is true.

3. If I constructed a 95% confidence interval, what should be the p-value cutoff for significance?

   i. What can you say when your p-value is below this number?

   ii. What can you say when your p-value is above this number?

Experimental Design

1. A *stent* is a kind of medical device – a tube implanted in a person's artery to shore it up when it is in danger of becoming blocked. A certain stent is designed to have a very precise thickness to match an artery's width, but the stent's manufacturing process makes small errors, so every stent has a slightly different thickness. These errors in thickness can make the device work less well. You're evaluating the manufacturing process, and you want to know the average amount of error in thickness among all the stents the manufacturer has made so far.

You have a machine for measuring the thickness of a stent, but when it measures a stent, it renders it unsuitable for implantation. So you can't just measure all the stents. Instead, you choose 10,000 stents uniformly at random from among all the manufactured stents, and you measure the errors in their thicknesses. (Error is measured as the absolute value of the difference between each stent's actual thickness and its designer-specified thickness.) Then you compute the average of those 10,000 numbers. For the purpose of the questions below, that average is the *statistic* you're working with in this scenario.

(a) What population parameter are you trying to estimate with this statistic?

(b) What is the population?

(c) You're worried that this statistic isn't a good enough estimate of the population parameter. Describe how you would use an inferential technique you've learned to make a quantitative claim about the population parameter that somehow conveys your uncertainty. Assume you have access only to the 10,000 measurements you've made.

2. Air pollution is a serious health concern in many cities around the world. Suppose that last year, before you took this course, you lived in Beijing and wanted to measure the average amount of fine particulate matter (known as PM2.5) across the city on March 30. We'll call that the "average PM2.5" for short. You couldn't get to every place in the city, so you measured the PM2.5 on March 30 at the 40 street corners nearest your apartment. You decided to use the average of those measurements as an estimate of the average PM2.5.

You knew that your sample didn't include all the locations in the city, so your estimate was prone to error. To reflect your uncertainty, you decided to compute an interval of estimates you might reasonably have seen instead. You took 10,000 resamples (uniform random samples with replacement) of size 40 from your sample, computed the average of each resample, and claimed that the 2.5th and 97.5th percentiles of those resample averages formed a 95% confidence interval for the average PM2.5.

(a) Did you use a randomized control experiment or some other kind of experiment in this study?

(b) If you had repeated this study many times, would your reported interval have contained the true average PM2.5 in roughly 95% of the repetitions? Why or why not?

(c) Describe a problem with the design of the study and recommend a fix.

3. For each study below, describe the treatment variable, the outcome variable, and whether or not the study provides evidence of a causal relationship between the treatment and outcome.

(a) (Study A: Soda) A team analyzed data on daily soda intake and obesity from 622 participants. Each participant recorded how much soda they drank as well as their weight. The study found that people who drank soda regularly (i.e., they drank at least one can daily) were 33% more likely to be obese. study found that people who drank soda regularly (i.e., they drank at least one can daily) were 33% more likely to be obese.

(b) (Study B: Artificial Sweetener) A study randomly assigned rats into two groups. They exposed each group of rats to the same amount of food, exercise, and living conditions, but fed one group artificially sweetened yogurt, and fed the other group unsweetened yogurt. They found that the first group had significantly more body fat on average after 12 weeks than the other group.

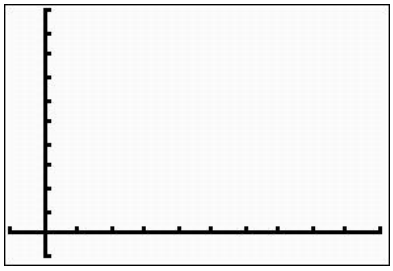
(c) (Study C: Chocolate) A study examined the eating habits of 8,000 men over 65 years, selected at random from the population of a particular state. They found that men who ate chocolate lived on average 1 year longer than those that did not eat any chocolate.

Study	Treatment	Outcome	Evidence of Causality? (yes/no)
Study A: Diet Soda			
Study B: Artifical Sweetener			
Study C: Chocolate			

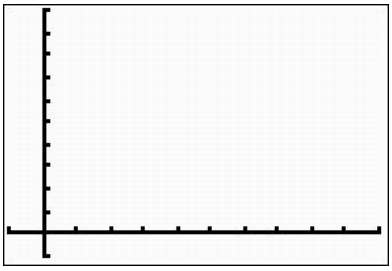
Correlation

1. Draw graphs with the following correlation coefficients.

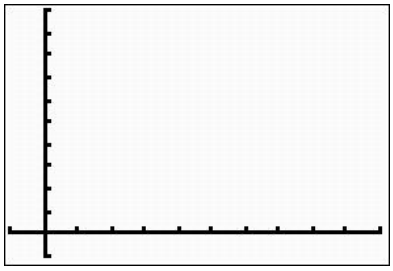
r = -0.9



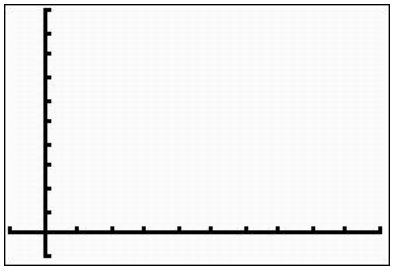
r = -0.1



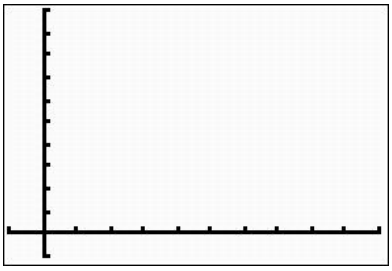
r = 0



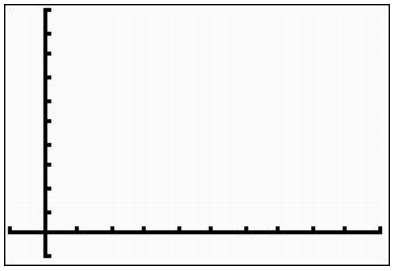
r = 0.2



r = 0.5



r = 0.8



Which two graphs have the strongest correlation?

2. In the StudentLife Study at Dartmouth College, researchers studied a class of 48 Dartmouth students over a 10 week term to assess their mental health (e.g., depression, loneliness, stress), academic performance (grades across all their classes, term GPA and cumulative GPA) and behavioral trends. Part of that study included Piazza usage, which was recorded for 30 of the 48 students in the class. Among the 30 students, there are 14 who answered at least one question; 16 did not.

a. Assuming these 30 were selected uniformly at random from all 48 students in the class, how would you compute a 95% confidence interval for the standard deviation of the GPA for all students in the class?

b. Among these 30 students, how would you determine whether their GPA and their number of Piazza contributions are linearly related?

c. Assuming these 30 were selected uniformly at random from all 48 students in the class, how would you determine whether or not their GPA and their number of Piazza contributions are positively linearly related for all students in the class?

Regression/Prediction

Central Limit Theorem

Definition:

Consider a normal histogram of the number of people in each household in the Bay Area. Say that the sample size that this histogram was generated from is huge. Hint: When does the *Central Limit Theorem* apply?

1. Say we took a sample of 20 households, and counted the number of people living in each one, and plotted the results on a histogram. What, if anything, could we predict about this histogram?

2. Say we took a sample of 1000 households, and counted the number of people living in each one, and plotted the results on a histogram. What, if anything, could we predict about this histogram?

3. Say we took 500 samples of 20 households each, and for each sample we took the average number of people living in a household, and then plotted the result of each sample on a histogram. What, if anything, could we predict about this histogram?

Normal Curve/Standard Units

1. Say we have a normal distribution of exam scores, where the mean is 71 points and the standard deviation is 7 points.

   a. What is the mean and standard deviation of this distribution is standard units?

   b. Say that Matheno scored q points. Provide a formula to write an arbitrary score q in standard units. How would you express Matheno's score using your formula?

   b. Oliver scored -3 on the standard units scale. What score did he receive?

   c. What can you say about the proportion of scores above 78 points? What can you say about the proportion of scores below 57 points?

   e. Under what condition does your previous answer hold (Think about the assumptions made at the beginning of the question).

Regression

1. Laurel has a table *t* with two columns, *x* and *y*. The x values have a mean of 12 and a standard deviation of 3. The y values have a mean of 23 and a standard deviation of 5. She wants to find the slope of the regression line. Assume she has a function *regress* which returns an array [slope, intercept] given a table and two column names. She tries three different code snippets to achieve this. Which one is correct and why? Why are the others considered wrong?

```
(a)   np.mean(t.column('x') * t.column('y'))

(b)   regress(t, "x", "y")

(c)   np.mean(((t['x'] - 12)/3)*((t['y']-23)/5))
```