Data 8 Spring 2016 lab14, Due: *Not collected*

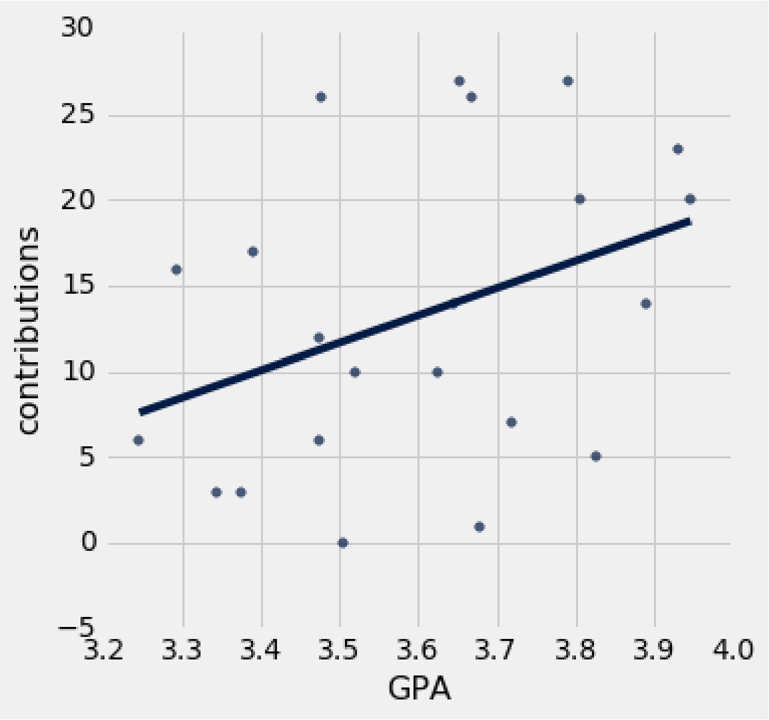
In the StudentLife Study at Dartmouth College, "passive and automatic sensing data from the phones of a class of 48 Dartmouth students over a 10 week term [were collected] to assess their mental health (e.g., depression, loneliness, stress), academic performance (grades across all their classes, term GPA and cumulative GPA) and behavioral trends (e.g., how stress, sleep, visits to the gym, etc. change in response to college workload – i.e., assignments, midterms, finals – as the term progresses)."

Part of that study included Piazza usage, which was recorded for 30 of the 48 students in the class.

A table called `students` has one row for each of these 30 students and the following columns:

- `GPA`: grade point average of the student
- `contributions`: The number of contributions, including posts and follow-up discussions
- `answers`: The number of answers posted

Analysis 1. A scatter plot of students with GPA above 3.1 versus their contributions appears below.



Analysis 2. Among the 30 students, there are 14 who answered at least one question; 16 did not.

Note: If you're interested, the StudentLife.ipynb file included with lab14 contains the data.

Problem 1 Estimation

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?

Answer: A bootstrap confidence interval would apply: resample with replacement 30 rows from the table many times, compute the standard deviation of the GPA for each resampled sample, then use the 2.5% and 97.5% percentiles of these statistics to construct the confidence interval.

Problem 2 Regression

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

Answer: Compute the correlation (r) and determine if it's not zero. Also, draw the residual plot for the regression line and look for structure that's different from just a random cloud, which would indicate a non-linear relation.

Problem 3 Regression Inference

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?

Answer: After verifying that the sample appears to be linearly related by looking at the residual plot (above), perform a hypothesis test to determine whether the slope of the regression line is greater than 0 (a positive linear relation).

The null hypothesis in this test is that the slope of the regression line for the whole class is less than or equal to 0. To evaluate it, resample with replacement 30 rows from the table many times and compute r for each resampled sample. The P-value for this test is the fraction of resampled samples with an r value that is less than or equal to zero.

An alternate approach to evaluate this null hypothesis is to compute a 95% confidence interval for r of the population. The procedure is very similar: resample with replacement 30 rows from the table many times and compute r for each resampled sample. Take the middle 95% of this distribution as the confidence interval and check whether or not that interval contains 0.

Problem 4 Comparison

You're interested in comparing the GPAs for two groups: those who did and did not answer at least one question on Piazza. Which of the following hypotheses could you test using the `students` table and how?

a) The difference between the average GPAs of the two groups among these 30 students is at least 0.2.

b) The difference between the average GPAs of the two groups among these 30 students is due to chance.

c) Answering questions on Piazza increases the GPA of students in this class.

Answer:

a) This difference can be computed directly; no inference technique is required to prove or disprove the hypothesis.

b) A permutation test would suffice. The null hypothesis is that the GPAs of the two groups were drawn from the same distribution. To evaluate it, first compute the observed absolute difference in GPAs between groups. Then, many times, shuffle the GPAs at random and compute the resampled absolute difference in GPAs between groups. The P-value is the proportion of resampled absolute differences that are at least as large as the observed statistic.

A bootstrap A/B test would also suffice. The null hypothesis is the same as above. The procedure differs only in how random sampling is performed: Many times, *resample 30 GPAs with replacement* and compute the resampled absolute difference in GPAs between the first 14 and last 16. The P-value is the proportion of resampled absolute differences that are at least as large as the observed statistic.

c) There is no way to evaluate causality using this table.

Problem 5 Randomness

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?

Answer:

a) Whether or not each student answered at least one Piazza question was determined by flipping a fair coin.

b) The number of students among 30 who answered at least one question.

c) Many times, simulate flipping 30 coins and record how many of the outcomes were heads.

d) 14

Problem 6 Median (Optional)

If you don't know how the 30 students were chosen from the 48 in the class, could you compute an interval that certainly contained the median GPA of the class as a whole? If so, how?

Answer: Yes. Sort the GPAs of the 30 observed students. All 18 other students could have even higher GPAs than these 30, in which case the average of the 24th and 25th highest observed GPA would be the median among the 48. That's the upper bound of the interval. Alternatively, all 18 other students could have even lower GPAs than the 30, in which case the average of the 6th and 7th highest observed GPA would be the median among the 48. That's the lower bound of the interval. Therefore, the median of all 48 must fall between the 6th largest and 25th largest values in the 30-person sample.

This reasoning is not so much about statistical inference as it is about a property of the median. If you know most of the elements of a set, then you know something about the set's median, because the median is a statistic that isn't affected very much by adding a few more values to the set.