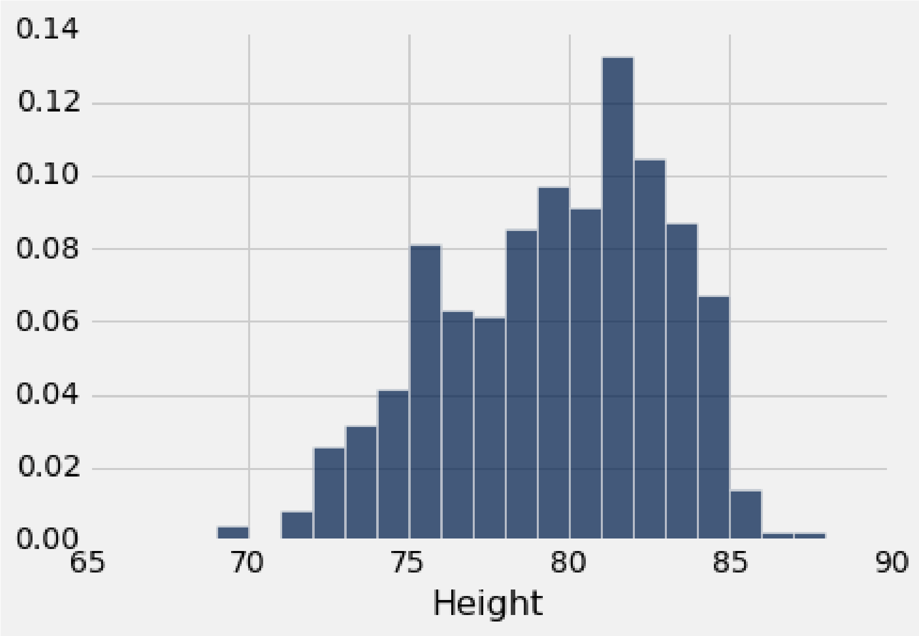
DATA 8 FALL 2016: Some Problems from the Fall 2015 Final

We haven't included all the problems because the course contents were different last year.

1. The histogram below shows the distribution of heights of all NBA (National Basketball Association) players in 2013. The heights are measured in inches.



a) Stephen Curry is 75 inches tall. He was part of this dataset. In 2013, the proportion of players in the NBA who were taller than Stephen Curry was closest to:

(i) 20% (ii) 50% (iii) 80%

Pick one option and explain your choice.

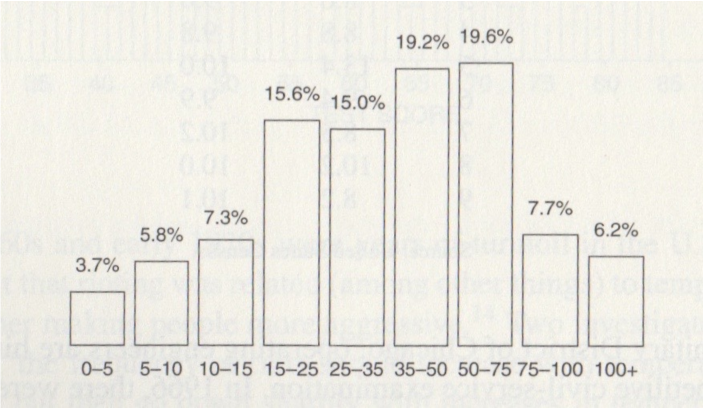
b) For the heights of NBA players in 2013, which was larger: the average or the median? Explain your answer.

2. A mutliple choice test consists of 100 questions. Each question has five possible answers, one of which is correct.

A student knows the correct answers to 70 questions. For each of the remaining 30 questions, she picks one of the five possible answers uniformly at random, independently of her answers to all other questions.

One of the questions is picked at random. Given that the student got the correct answer, what is the probability that she knew the answer? Show your work, and leave the answer as an arithmetic expression; don't simplify.

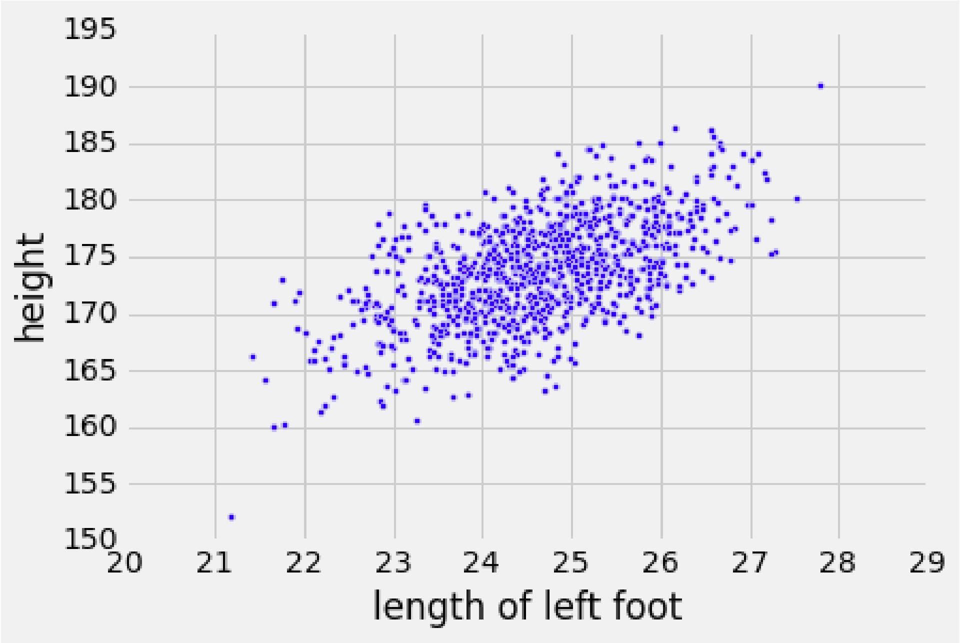
3. The figure below shows the distribution of incomes in a city. Incomes have been measured in thousands of dollars.



a) Is the figure a histogram? Explain your answer.

b) Say whether the following statement is true or false, and explain your answer.
"Assuming that the incomes are evenly distributed within each interval, the percent of incomes in the 40-50 range is pretty close to the percent of incomes in the 65-75 range."

4. The figure below shows the scatter plot of the height versus the length of the left foot for 1,020 men in a part of southern India. Both variables are measured in centimeters.



One of the men is on the 30th percentile of left foot lengths. What would be your best estimate for his percentile rank in height? Pick one of the three choices below as your answer, and explain your choice.

(i) below the 30th percentile of heights

(ii) on the 30th percentile of heights

(iii) above the 30th percentile of heights

5. According to a genetics model, each plant in a species has a 75% chance of being pink-flowering, independently of all other plants. Among 200 such plants, 170 turn out to be pink-flowering. Do the data support the model, or not? Develop an answer in the following steps.

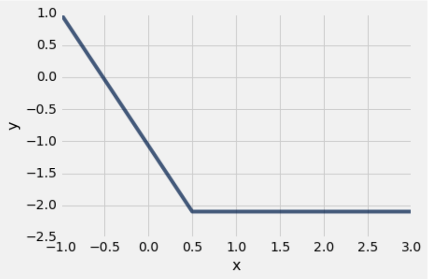
a) State precise null and alternative hypotheses that reflect the question as posed above.

b) Test your null hypothesis by writing Python code such that the final line evaluates to either an exact or an approximate P-value; say whether the value you are computing is exact or approximate.

6. This question is about interpreting Python code and plotting a function using the Table method `plot`. First let us give an example. When the following code is executed:

```
  Table().with_columns(
    'x', make_array(-1, 0.5, 3),
    'y', make_array(1, -2.1, -2.1)
).plot('x')
```

. . . this graph is produced:



Sketch the plot produced when the following code is run. Don't worry about exactly what numbers `plot` will show on the horizontal and vertical axes. Just be clear about what you have drawn, by providing the numerical coordinates of the points that you think are important.

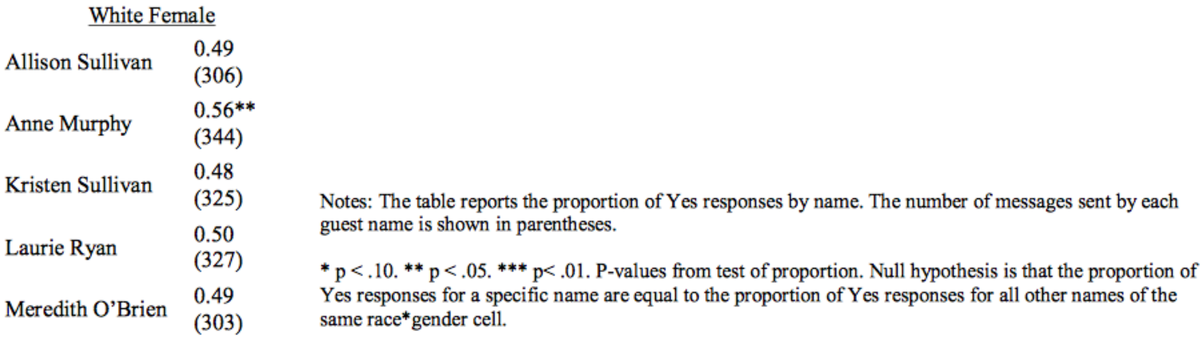
```
def plot_function(f, min_input, max_input):
    NUM_TICKS = 30
    tick_spacing = (max_input - min_input) / NUM_TICKS
    inputs = np.arange(min_input, max_input+tick_spacing, tick_spacing)
    tbl = Table().with_column('input', inputs)
    function_table = tbl.with_column('f(input)', tbl.apply(f, 'input'))
    function_table.plot('input')

def proportion_above(x):
    data = make_array(-1.0, 5.5, -3.5, 6.5)
    return np.count_nonzero(data >= x) / len(data)

plot_function(proportion_above, -5.0, 10.0)
```

7. A survey organization wants to estimate the proportion of voters who would like Donald Trump to win the Presidential election. [Note from Prof. A., 11/28/2016: Wow. Last year I probably thought this was funny.]

Assume that the organization is using the methods we have learned in this class, based on a large sample drawn by a method that is essentially equivalent to sampling at random with replacement. How large must the sample be so that an approximate 95% confidence interval for the proportion will have a total width of no more than 0.04?

Please leave your answer as an unsimplified numerical expression.

8. In a recent (very recent: 9 December 2015) report, researchers at the Harvard Business School studied the effect of a guest's name on the response of Airbnb hosts. They created 20 guest profiles, identical except for the name. Hosts in the Boston area received availability queries from a randomly assigned guest name. The chosen guest names were among those known from previous studies to be commonly associated with a particular race and gender. The main result of the study was that hosts were more likely to respond "Yes" about availability if the guest's name is associated with being white. However, the study also looked for differences in response within each race/gender group. Here are some extracts from their working paper.



	White Female
Allison Sullivan	0.49 (306)
Anne Murphy	0.56** (344)
Kristen Sullivan	0.48 (325)
Laurie Ryan	0.50 (327)
Meredith O'Brien	0.49 (303)

Notes: The table reports the proportion of Yes responses by name. The number of messages sent by each guest name is shown in parentheses.

* p < .10. ** p < .05. *** p< .01. P-values from test of proportion. Null hypothesis is that the proportion of Yes responses for a specific name are equal to the proportion of Yes responses for all other names of the same race*gender cell.

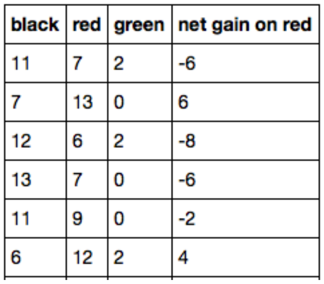
a) Read the two short paragraphs in the section called Notes, next to the displayed data. Each guest name has a corresponding null hypothesis. For the row of the table corresponding to Anne Murphy, is the null hypothesis saying that 0.56 is equal to the overall Yes proportion observed for the other four names? If not, what is the null hypothesis saying?

b) Suggest a reasonable statistic to test the null hypothesis in part **(a)**, and compute its observed value. Just give an arithmetic expression using the numbers in the table; don't simplify.

9. A roulette wheel has 38 pockets, of which 18 are black, 18 are red, and 2 are green. On each spin, all pockets are equally likely to win, independently of all other spins. In what follows, it might help to know that 18/38 is approximately 0.474 and 2/38 is approximately 0.0526.

The bet on red pays 1 to 1. As described in class, this means that if you bet \$1 on red on one spin of the wheel, and a red pocket wins, you make a net gain of \$1; if a black or green pocket wins, you make a net gain of −\$1 (that is, you lose a dollar).

In a Data Science class of 110 students, each student bets \$1 on red on 20 different spins of the wheel, independently of all other students. The table `roulette` contains the results. There is one row for each student. In each row, the columns `black`, `red`, and `green` contain the number of times the student's winning pocket was black, red, or green respectively. The final column contains the dollar amount of the student's net gain. Here are the first few rows of `roulette`.



black	red	green	net gain on red
11	7	2	-6
7	13	0	6
12	6	2	-8
13	7	0	-6
11	9	0	-2
6	12	2	4

Assume that the function `correlation(table_name, x_column_label, y_column_label)` returns the correlation between the two specified columns of the table `table_name`.

In each part of this question, pick the number from the list below that you think best approximates the output of the given Python code. Explain your choice briefly.

$-1.27 \qquad -1 \qquad -0.9 \qquad -0.0495 \qquad 0 \qquad 0.0495 \qquad 0.9 \qquad 1 \qquad 1.27$

a) `correlation(roulette, 'black', 'red')`
b) `np.mean(roulette.column('green')/20)`
c) `correlation(roulette, 'red', 'net gain on red')`

10. This question is about the conclusions made in statistical tests of hypotheses.

a) Say whether the following statement is true or false, and give a brief explanation.
"If the *P*-value of a test is 0.02%, then there is only about a 0.02% chance that the null hypothesis is true."

b) Suppose you are performing a test of hypotheses using the 5% cutoff for the *P*-value. Fill in the blank using one of the values in the list below, and explain your choice.
"If the null hypothesis is true, then the chance that the conclusion of the test favors the alternative hypothesis is closest to ________.

0   0.05   0.5   0.95   1

11. In this question you will construct a confidence interval using the bootstrap.

a) Write Python code to define a function `boot1_median` that returns **one** bootstrapped median, as follows. The function should take just one argument: an array consisting only of the data to be bootstrapped. The function should perform **one** bootstrap replication and return the median of the resulting data.

For example, suppose the array `income` contains a sample of incomes. The call `boot1_median(income)` will bootstrap the values in `income` just once and return the median of the bootstrapped sample.

b) The array `berkeley` contains the monthly rents of a random sample of rental units in Berkeley. The array `oakland` contains the monthly rents of a random sample of rental units in Oakland. You can assume that each sample is like draws made at random with replacement from all rental units in the appropriate city, but you cannot assume that the two sample sizes are equal.

Construct an approximate 95% bootstrap confidence interval for the difference between the median rent of Berkeley rental units and the median rent of Oakland rental units, by defining a function `ci_diff_medians` that takes as its argument the number of bootstrap replications and returns an array consisting of the left endpoint and right endpoint of the interval. Thus the call `ci_diff_medians(5000)` should return an array consisting on the left end and right end of an approximate 95% confidence interval for the difference between the median rent in the two cities, based on 5,000 bootstrap replications.

In your code you may use the function `boot1_median` defined in part **(a)** (even if you didn't do that part of the question), and also the function `middle95` that takes a numerical array as its argument and returns an array consisting of the 2.5th and 97.5th percentiles of the argument array.

Note: You can calculate differences either as "Berkeley − Oakland" or as "Oakland − Berkeley", but be consistent throughout your code.

12. In the United States, the proportion of boys among live-born babies is just about 0.512. Assume that the sex of each baby is random and independent of all other babies.

The following events are about live births in the United States. Arrange them in increasing order of probability, by filling in the table with the letters corresponding to the events.

A: Fewer than 500 boys among 1,000 live births
B: More than 500 boys among 1,000 live births
C: Fewer than 5,000 boys among 10,000 live births

	least likely	**next least likely**	**most likely**
letter			

Explain your choices. No, you don't need a computer or outstanding algebra skills.