STAT/CS 94 Fall 2015 Adhikari HW10, Due: 11/13/15

NAME: SID:

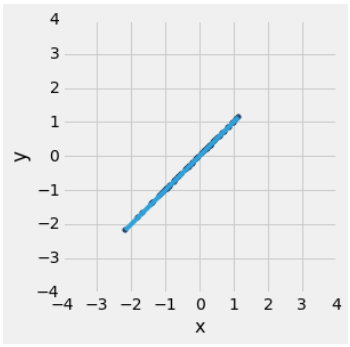
Please write your answers on the (double-sided) printed answer sheet, in the space provided. Note the due date of this homework: 11/11/15 is Veterans' Day, so the homework is due by the end of lecture (11AM) on 11/13.

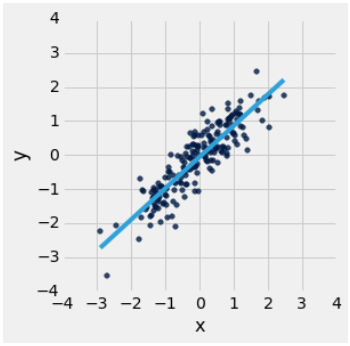
Problem 1 Correlation Constellation

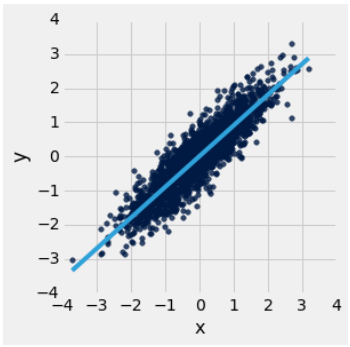
Here is a list of numbers:

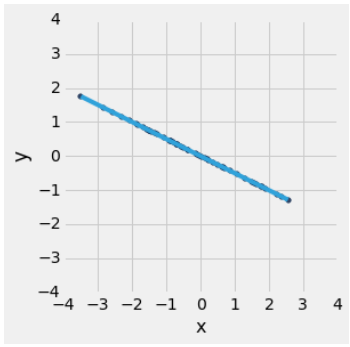
(i) -2	(iii) -0.9	(v) 0	(vii) 0.9	(ix) 2
(ii) -1	(iv) -0.5	(vi) 0.5	(viii) 1	

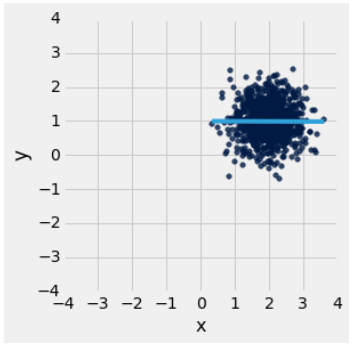
Each of the following scatter plots depicts a different football-shaped dataset; not all have the same number of data points. The least-squares fit line has been drawn over each scatter plot. For each plot, the correlation coefficient of the plotted data is one of the above numbers. (Several datasets may share a correlation coefficient.) Identify the correlation coefficients.

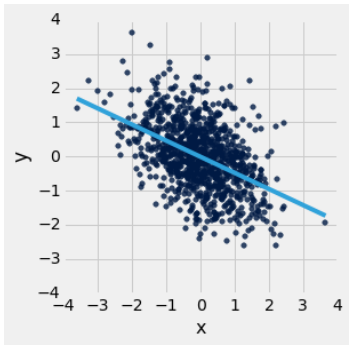
(a) 

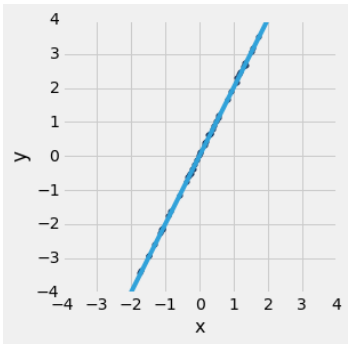
(b) 

(c) 

(d) 

(e) 

(f) 

(g) 

Answer:

(a) 1 (viii)

(b) .9 (vii)

(c) .9 (vii). It's not immediately obvious whether the answer should be .5 or .9. Note that the number of data points *does not directly matter* in determining the correlation. What matters is the spread of the cloud. The spread here is similar to that in part (b) and narrower than that in part (f).

(d) -1 (ii)

(e) 0 (v)

(f) -0.5 (iv)

(g) 1 (viii). The slope of the best-fit line is 2, but when we put the data in standard units, the slope of the best-fit line (which is the correlation coefficient) is 1. There are several ways to see this: the vertical axis in the original units is more spread out than the horizontal axis; the correlation coefficient is always between -1 and 1.

Problem 2 Cocoa Kudos?

A 2012 paper in the New England Journal of Medicine studied the relation between chocolate consumption and the number of Nobel Prizes received in 23 countries. No, I'm not kidding; the reference will appear in the textbook. You can see if you think the article was intended to be taken seriously.

The correlation between chocolate consumption per capita and the number of Nobel laureates per 10 million persons was 0.791 and the scatter diagram was fairly linear.

(a) Do the data show that consuming chocolate is a reason for getting the Nobel Prize, or that getting the Nobel Prize is a reason for consuming chocolate, or neither?

(b) Give **one** substantive and plausible explanation for the correlation.

Answer:

(a) Neither. Nonzero correlation does not imply causation; below we'll see several plausible explanations for the correlation in which neither thing is a cause of the other. (We'll use the terminology "A causes B" and "A is a reason for B" interchangeably in our answers.)

(b) You only had to give one explanation, but here is a list of *potentially true* explanations we came up with. There could be others. Note that these are just stories; the data we've been given are insufficient to tell us whether any of these are true.

   (1) The Nobel judges are predominantly Swedish, so they are biased toward European cultures. People in European cultures also eat more chocolate. So there is a *common causal factor* (being European) that causes both, but neither thing causes the other.

   (2) Higher education or wealth cause both higher chocolate consumption and higher rates of Nobel Prizes. Again, this is an explanation with a common causal factor.

   (3) In the whole population of nations, there is no (or a smaller, or a negative) correlation between chocolate consumption and Nobel Prize winning. The researchers chose a random sample of 23 nations and happened to see a positive correlation. So there is no causation going on anywhere.

   (4) In the whole population of nations, there is no (or a smaller, or a negative) correlation between chocolate consumption and Nobel Prize winning. The researchers cherry-picked a dataset of 23 nations in which there was a positive correlation, so they could publish an article in a famous journal.

Here are some possible explanations that **don't** seem plausible to us.

(1) Chocolate consumption increases energy (since it contains calories and caffeine), which makes people more productive, which leads some people to win Nobel Prizes. (In this story, chocolate consumption is a reason for getting the Nobel Prize.) This story could be true, but the effect couldn't possibly be strong enough by itself to see a correlation as high as 0.791.

(2) People celebrate winning the Nobel Prize by eating chocolate, so Nobel Prizes cause higher chocolate consumption. Again, this story could be true, but it wouldn't result in a correlation of 0.791 by itself.

Problem 3 Incubator Indicators

The diameter (measured at the widest part) of the eggs of a species of bird have an average of 55 mm and an SD of 0.8 mm. The weights of the chicks that hatch from these eggs have an average of 10 grams (yes, they're tiny) and an SD of 0.5 grams. The correlation between the two variables is 0.6.

(a) Write Python expressions that evaluate to the following:

(i) The regression estimate (in grams) of the weight of a chick that hatches from an egg with diameter 55 mm.

(ii) The regression estimate (in grams) of the weight of a chick that hatches from an egg with diameter 47 mm.

(iii) An array consisting of the regression estimates (in grams) of the weights of chicks that hatch from eggs of diameters (in cm) in an array `diameters`. Element `i` of your array should be the regression estimate corresponding to element `i` of `diameters`.

(iv) The SD (in grams) of the residuals of the regression of weight on diameter.

(b) Do any of your answers to part (a) depend on the scatter diagram being football shaped or linear in some other way? Explain.

Answer:

(a) (i) `0.6*(0.5/0.8)*(55 - 55) + 10`

(ii) `0.6*(0.5/0.8)*(47 - 55) + 10`

(iii) `0.6*(0.5/0.8)*10*(diameters - 55) + 10`. Note that the diameters were given in centimeters, so we had to multiply by 10 to get millimeters.

(iv) `math.sqrt(1 - 0.6**2) * 0.5`. The general formula, given in the textbook, is:

$$\text{SD(residuals)} = \sqrt{1 - r^2} \times \text{SD(weights}$$

.

Perhaps it is more intuitive to look at this formula as:

$$\frac{\text{SD(residuals)}}{\text{SD(weights}} = \sqrt{1 - r^2}$$

.

Let us give some very rough intuition behind this formula. Think of the right hand side as being like $1 - |r|$. Since the residuals in a least-squares linear regression have mean 0, and the SD is a measure of the typical distance of data from their mean, the SD of the residuals is a measure of the typical size of the residuals. If it is 0, then the residuals are all 0, and the correlation coefficient should be 1 or -1. If it is large, then the correlation coefficient should be closer to 0. But "large" is relative to the original scale of the data – multiplying all the weights by 2 shouldn't change our estimate of the correlation between weights and diameters. So we divide the SD of the residuals by the SD of the weights to normalize it.

(b) No, none of them depend on the scatter diagram at all, except insofar as the scatter diagram determines the quantities we were given (the mean and SD of weights and diameters, and the correlation between the two). We were able to calculate our answers using just those numbers, so our answers couldn't have depended on other things. This should be intuitive for (i), (ii), and (iii); we can calculate regression predictions if we have the information necessary to calculate the regression line. It's less obvious why this should be true for (iv); you should consult the explanation above, or the textbook, if you find it confusing.

Note: If you said that the answer depends on the shape of the scatter diagram *but only insofar as that affects the mean and SD of the two variables and their correlation*, you received some credit.

Problem 4 Predictive Poundage

The scatter diagram of weights and systolic blood pressures of a population is football shaped; so you can assume that all the variables relevant to this exercise are normally distributed to an excellent approximation. The correlation between the two variables is 0.4.

One of the people is on the 80th percentile of weights. Write Python expressions that evaluate to the following:

(a) The person's weight, in standard units.

(b) The regression estimate of the person's systolic blood pressure, in standard units (of blood pressures).

(c) The regression estimate of the percentile rank of the person's systolic blood pressure.

Answer:

(a) `stats.norm.ppf(0.8)`

(b) `0.4 * stats.norm.ppf(0.8)`

(c) `100*stats.norm.cdf(0.4 * stats.norm.ppf(0.8))`

Problem 5 Groovy Grades

The scatter diagram of the midterm scores and final exam scores in a class is football shaped (hence normal assumptions as in the previous exercise).

(a) Fill in the blanks with the best bounds you can come up with, and explain your choices.

A student is on the 40th percentile of midterm scores. The regression estimate of the student's percentile rank on the final exam will be between the _____ percentile and the _____ percentile.

(b) Pick one option and explain your choice.

Of the students who are on the 70th percentile of midterm scores,

(i) fewer than half (ii) about half (iii) more than half

were above the 70th percentile of final exam scores.

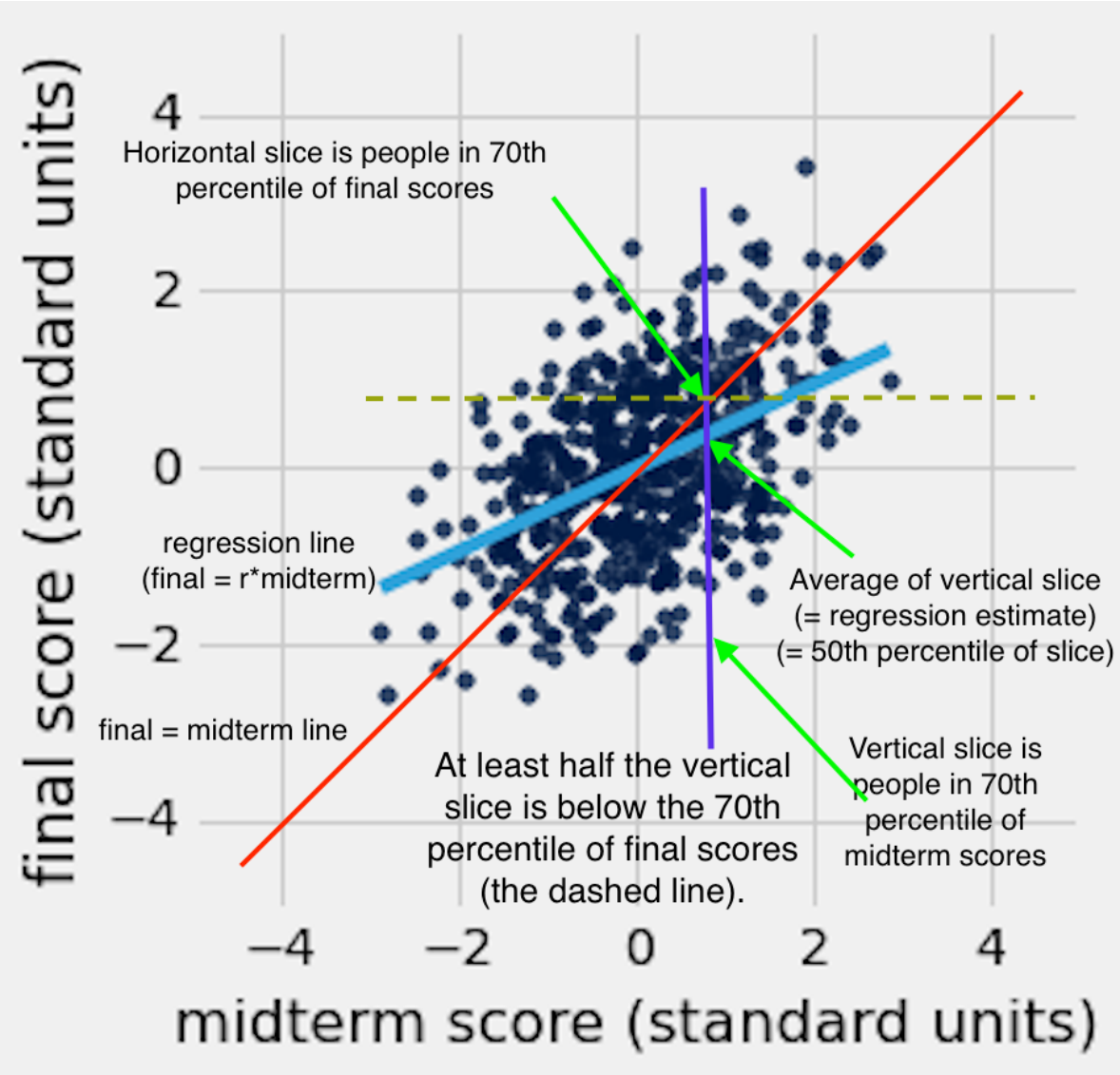
Answer:

(a) The regression estimate of the student's percentile rank on the final exam will be between the 40th percentile and the 60th percentile. Informally, this is regression to the mean. More formally, imagine the student's midterm score in standard units is x. We could calculate x as `x = stats.norm.ppf(.4)`, which is around -.25. (Here we are relying on the assumption, noted in the question, that the midterm scores are normally distributed.) Then the regression estimate of the student's final score is $r \times x$. We don't know r, but we know it's always between -1 and 1. So, correspondingly, the regression estimate of the student's final score in standard units is between x and $-x$. Now, again relying on the assumption of normality, the range of possible percentiles of those estimates is given by the normal CDF: `stats.norm.cdf(x)` and `stats.norm.cdf(-x)`. Those are 40% and 60%, respectively.

(b) (i), fewer than half. The story is basically the regression to the mean phenomenon again, but it's actually a little complicated. This answer, like the one in part (a), would not necessarily be true if the data weren't football-shaped. So the best way to get intuition for this answer is to draw a football-shaped dataset with a best-fit regression line and see what happens at the 70th percentile on the midterm axis.

If the correlation between final exam scores and midterm scores were 1, then everyone on the 70th percentile of midterm scores would be on the 70th percentile of final scores, and nobody would be above that.

Otherwise, the correlation is strictly less than 1. Though the answer is still true if the correlation is negative, we'll focus on the positive-correlation case (that is, r between 0 and 1); you can work out an example with negative correlation if you're not convinced. So we have a picture like the following:



Problem 6 Gambler's Gaffe?

A bet on red at roulette pays 1 to 1 and there are 18 chances in 38 to win. A gambler visits a casino once a week for 20 weeks. On each visit, he bets 10 times on red, on 10 different spins of the roulette wheel. He keeps track of the number of bets won each time, resulting in a list of 20 numbers. The number of bets won has an average of 3.2 and an SD of 1.1.

If possible, find numerical answers (no code) for the following. If this is not possible, explain why.

(a) the average number of bets lost

(b) the SD of the number of bets lost

(c) the correlation between the number of bets won and the number of bets lost

Answer:

(a) 6.8. There is a mechanical relationship between the number of bets lost and the number of bets won in a single visit to the casino: They sum to 10. Equivalently, the number of bets won in a single visit equals 10 minus the number of bets lost. That means the total number of bets won in the 20 visits is $20 \times 3.2 = 64$, the total number of bets is $20 \times 10 = 200$, the total number of bets lost is $200 - 64 = 136$, and the average number of bets lost per visit is $136/20 = 6.8$.

(b) 1.1. Informally, any time the number of losses on one visit changes by 1, then the number of wins must also change by 1, so the numbers of losses and wins have the same variability. More formally, it is true in general that, if `X` is an array of numbers, and `a` and `b` are two numbers, then `np.std(X + b) == np.std(X)` (adding something to each member of a list of numbers doesn't change its variability) and `np.std(a*X) == abs(a)*np.st` (scaling each member of a list of numbers scales its variability by the same magnitude). (Neither of those are very hard to prove, but hopefully they're intuitive anyway.) So if `wins` is an array of 20 win counts, then `losses` is equal to `(-1)*wins + 10`, so `np.std(losses) == np.std((-1)*wins + 10)`, which equals `np.std((-1)*wins)`, which equals `abs(-1)*np.std(wins)`, which equals `np.std(wins)`.

(c) -1. Imagine drawing a scatter plot with 20 data points, one for each week. On the x-axis is the number of wins for that week, and on the y-axis is the number of losses. Since the number of losses is always equal to 10 minus the number of wins, the points all fall exactly on the line $y = 10 - x$. In other words, there is a perfect negative correlation.

Problem 7 Infant Inference

A team of medical researchers records the weights and the head circumferences of a large random sample of newborns. The data are stored in a table called `newborns`, with the weights in the column `'weight'` and the head circumferences in the column `'head'`.

You can assume that the sampling scheme is essentially equivalent to random sampling with replacement. The scatter plot of the two variables is football shaped.

The researchers would like to construct a bootstrap confidence interval for the correlation between the weights and head circumferences of newborns in the entire population. Define a function that will do this, as follows.

Assume that the function `corr(table, column_name_x, column_name_y)` returns the correlation between the arrays `table[column_name_x]` and `table[column_name_y]`, just as in class.

Define a function `r_ci` that takes the following 5 arguments:

1. `table`: the table containing the data
2. `column_name_x`: the label (string) of the column containing variable x
3. `column_name_y`: the label (string) of the column containing variable y
4. `L`: a floating-point number, strictly between 0 and 100, specifying the level of confidence
5. `rep`: the number of repetitions of the bootstrap resampling procedure

The function should return an array consisting of the two endpoints of an approximate L% confidence interval, constructed using the bootstrap percentile method, for the correlation between the two variables in the population.

Answer:

```
def r_ci(table, column_name_x, column_name_y, L, rep):
    correlations = []
    for rep_index in np.arange(rep):
        resample = table.sample(with_replacement=True)
        correlations.append(corr(resample, column_name_x, column_name_y))
    lower_percentile = (100-L)/2
    lower_limit = np.percentile(correlations, lower_percentile)
    upper_percentile = (100+L)/2
    upper_limit = np.percentile(correlations, upper_percentile)
    return [lower_limit, upper_limit]
```