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

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.

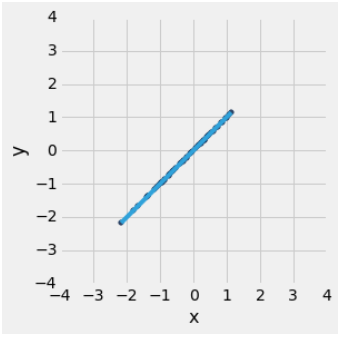
Problem 1 Correlation Constellation

Here is a list of numbers:

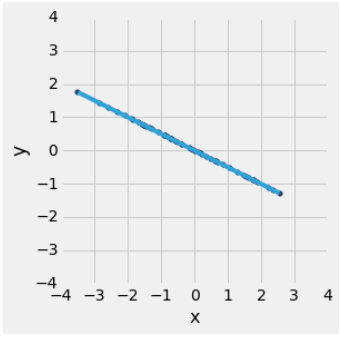
(i) -2	(iii) -0.9	(v) 0	(vii) 0.9	(ix) 2
(ii) -1	(iv) -0.5	(vi) 0.5	(viii) 1	

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.

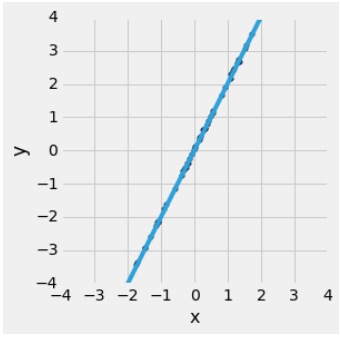
(a)



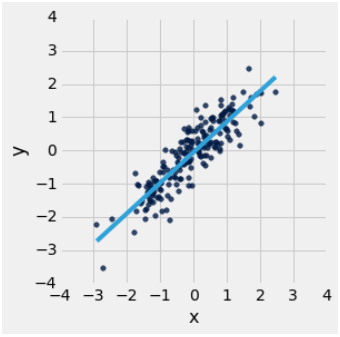
(d)



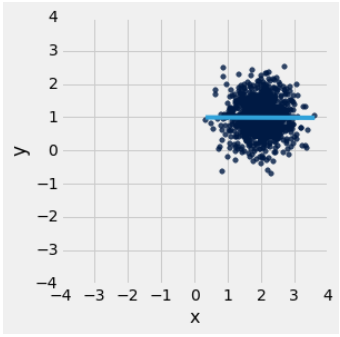
(g)



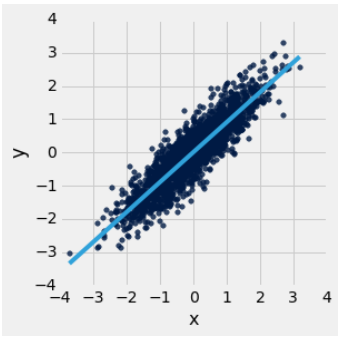
(b)



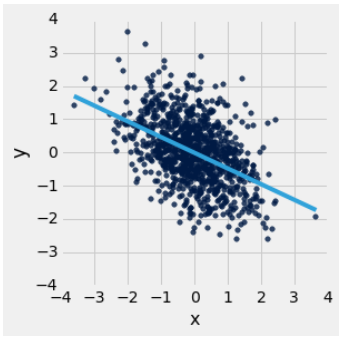
(e)



(c)



(f)



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.

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.

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.

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 students 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.

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

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.

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

(a)

(b)

(c)

(d)

(e)

(f)

(g)

Problem 2 Cocoa Kudos?

(a)

(b)

Problem 3 Incubator Indicators

(a) (i)

(ii)

(iii)

(iv)

(b)

Problem 4 Predictive Poundage

(a)

(b)

(c)

Problem 5 Groovy Grades

(a)

(b)

Problem 6 Gambler's Gaffe?

(a)

(b)

(c)

Problem 7 Infant Inference