

Name: Solutions

1. (12 points) Define the following:

2 (a) The method of comparison

Used to study the effect of a treatment.

2 (b) The treatment group

The treatment group is compared with the control group.

2 (c) The control group

Group that gets treatment you are studying

2 (d) A confounding variable

Group that gets no treatment at all.

2 (e) A placebo

A difference between the treatment and control group (other than the treatment) that can affect the results.

2 (f) A double blind experiment

A "fake treatment" used to trick the control group into not knowing which group they are in.

2 (12 points) Observational and controlled experiments.

2 (a) Briefly explain the major differences between observational and controlled experiments.

In a controlled experiment the investigators choose treatment & control groups and administer the treatment. In an observational study they do not intervene, they just watch.

- 2 (b) Suppose that Alice wants to determine if being taller can make you more intelligent. She takes data on a sample of students (from 1st grade to 12th grade) by measuring their height and giving every student the same exam. She finds that the taller students score much higher on the exam, so she concludes that they are more intelligent! Is there a problem with her experiment? What is the problem?

YES! The students age is a confounding variable. The taller students are also older (on average) so that is why they do better on the exam.

- 2 (c) How could Alice redesign her experiment to account for this problem?

Just compare students in the same grade.  
(ie, break the data into 12 different data sets)

- 2 (d) Bob is wondering if attending Harvard will make you more successful in life. He compares the students who graduate from Harvard to the national averages and notices that the Harvard students are much more successful. He concludes that attending Harvard makes students more successful. Do you think Bob made the right conclusion from this experiment?

No! There are many confounding variables. The students who attend Harvard had to first be qualified to be accepted to Harvard. He should compare Harvard graduates to people who were accepted but decided not to go.

- 2 (e) Why is a randomized controlled double blind experiment more reliable than an observational study?

The control & treatment groups are very similar. There will be no confounders.

- 2 (f) If a randomized controlled double blind experiment is so much better, why does anybody ever conduct an observational study?

Sometimes you have no choice. For example, to study the effects of long term life decisions (such as smoking or going to college) you cannot choose a control & treatment group. Also, there is no way to offer a "placebo" which makes patients think they attended college!

3. (9 points) Consider the list of numbers: 1, 1, 1, 1, 2, 2, 2, 3, 3, 4.

(a) What is the average of these numbers?

$$\frac{1+1+1+1+2+2+2+3+3+4}{10} = 2$$

(b) What is the median of these numbers?

$$\underbrace{1 \ 1 \ 1 \ 1}_{\text{4 ones}} \ 2 \ 2 \ \underbrace{2 \ 3 \ 3 \ 4}_{\text{4 numbers}}$$

Median is average of 2, 2.

Thus Median = 2

(c) What is the standard deviation of these numbers?

deviation from average = -1, -1, -1, -1, 0, 0, 0, 1, 1, 2

$$\text{rms} = \sqrt{\frac{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 2^2}{10}}$$

$$= \sqrt{1} = 1$$

$$\boxed{SD = 1}$$

4. (4 points) Clara measured the height of the adults in her neighborhood and notices that on average the older people were shorter. She says that this means you shrink as you get older. Is Clara right? Why or why not?

NO! She conducted a cross-sectional study.

A single person does not get shorter, but people born in the 40's are shorter.

5. (6 points) Average and Standard Deviation.

(a) What happens to the average of a list if you add 7 to each element? Why?

It increases by 7



(b) What happens to the standard deviation of a list if you add 7 to each element? Why?

Stays the same.

Shifting does not change spread



(c) Suppose there are 9 people in a room, with an average age of 10. Someone who is 20 walks into the room. What is the average age now?

Average age = 10 implies

$$10 = \frac{\text{sum of ages}}{9}$$

so sum of ages = 90.

$$\text{New average} = \frac{\text{new sum}}{10} = \frac{90 + 20}{10} = \frac{110}{10} = \boxed{11}$$

6. (8 points) (a) What is the difference between chance error and bias?

Chance error cannot be avoided and is equally likely to be + or - (so on average it will cancel). Bias is from some measuring mistake and is always in the same direction.

(b) How does an investigator deal with bias?

Try to eliminate it with careful measurements.

(c) How does an investigator deal with chance error?

Take many measurements and average.

(d) If a student scores a 155 on an IQ test (remember the average IQ is 100 with SD 15) is it more likely that chance error worked in his favor or against him? Why?

In his favor.

He got an unusually high score, so it is likely that he got at least a little lucky.

7. (10 points) Compute the correlation coefficient ( $r$ ) for the following data:

| $x$ | $y$ |
|-----|-----|
| 1   | 1   |
| 1   | 2   |
| 1   | 1   |
| 1   | 3   |
| 2   | 1   |
| 2   | 4   |
| 2   | 1   |
| 3   | 2   |
| 3   | 2   |
| 4   | 3   |

$x$   
Same data as #3.

So we know  $\text{Ave } x = 2$ ,  $\text{SD } x = 1$ .

$y$  In fact, the  $y$  data is also the same numbers but in a different order so  $\text{Ave } y = 2$  ~~#~~  $\text{SD } y = 1$ .

Now convert to standard units and multiply.

| $x$ | $y$ | $z$ (su) | $y$ (su) | $xy$ |
|-----|-----|----------|----------|------|
| 1   | 1   | -1       | -1       | 1    |
| 1   | 2   | -1       | 0        | 0    |
| 1   | 1   | -1       | -1       | 1    |
| 1   | 3   | -1       | +1       | -1   |
| 2   | 1   | 0        | -1       | 0    |
| 2   | 4   | 0        | +2       | 0    |
| 2   | 1   | 0        | -1       | 0    |
| 3   | 2   | +1       | 0        | 0    |
| 3   | 2   | +1       | 0        | 0    |
| 4   | 3   | +2       | +1       | 2    |

Take Average

$$\frac{1+0+1-1+0+\dots+0+2}{10}$$

$$= \frac{3}{10} = 0.3$$

$$\boxed{r = 0.3}$$

8. (5 points) What is the difference between the SD line and the regression line? Which is more useful?

The SD line has slope  $\frac{SD_y}{SD_x}$  (so it assumes  $+1$  SD in  $x$  corresponds to  $+1$  SD in  $y$ ) and the regression line has slope  $\frac{r SD_y}{SD_x}$  (so  $+1$  SD in  $x$  corresponds to  $+r$  SD in  $y$ ).

9. (4 points) How do outliers effect the following statistics? predict  $y$  from  $x$ .

(a) Average

skewed towards the outlier

(b) Median

No effect

(c) Standard deviation

increased (histogram is "stretched out").

(d) Interquartile range

No effect

10. (5 points) Donna is trying to find the correlation between two variables in a complicated chemistry experiment. She computes  $r$  and gets that  $r$  is very close to zero. She concludes that these two variables are not associated at all. Do you agree with her conclusion? Explain.

No. ~~the~~  $r$  close to zero only means there is no linear correlation. They may be related in some other complicated way.

11. (10 points) (a) What five summary statistics are used to understand a scatter plot? (assume the variables are  $x$  and  $y$ )

Ave  $x$

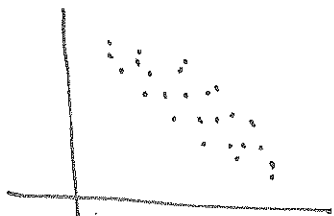
SD  $x$

Ave  $y$

SD  $y$

$r$  (correlation coef)

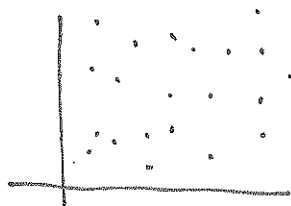
- (b) Sketch some data (scatter plot) which would have  $r \approx -0.95$ .



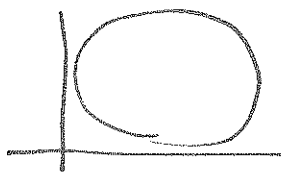
or



- (c) Sketch some data which would have  $r \approx 0.00$ .



or



- (d) Fill in the blanks. The correlation coefficient is always between -1 and +1.

- (e) What is the regression effect? (or you can explain the regression fallacy)

The regression effect ~~is~~ is the fact that the average  $y$  value of extreme points in  $x$  is not as extreme. For example, students who do poorly on one exam will (on average) do poorly on the next exam... but not as poorly. The regression fallacy is thinking this effect has some significance.

12. (10 points) Suppose a study has been done on the final grades of students enrolled in both math and physics courses. Each course is graded out of 100 possible points. It has been found that the average grade in Physics is 70 with a standard deviation of 10 and the average grade in Math is 75 with a standard deviation of 16. The correlation coefficient is  $r = 0.5$ . Suppose that Eric is in both of these classes and he got an 80 in Physics. Using regression, what would you expect his grade to be in Math?

His grade is  $+1$  SD above average in Physics ( $\frac{80-70}{10} = +1$ ) so his grade in Math should be  $+r$  SD above average. Thus his math grade should be 8 points above average ( $r \cdot 16 = .5 \cdot 16 = 8$ ). So his grade is predicted as  $75 + 8 = 83$

(extra space to work if you need it)