

Problem Set 2 Solution

Part A.

Chapter 3.

1. 66 & 71.999

2. See attached

(a) age 1 : $8/5 = 1.6\%$

age 11 : $13/9 = 1.444\%$

Therefore, there are more children in age 1.

(b) age 21 : $10/7 = 1.4\%$

age 31 : $9/5 = 1.8\%$

Therefore, there are more 31-year-olds.

(c) more age 35-44

(d) around 50%

4. (a) $1.8(5) + 1(5) + .8(10) + .3(10) = 9+5+8+3 = 25$. Answer: 25%

(b) 99%

(c) 135-140mm : $1\% \times 5 = 5\%$

140-150mm : $.8\% \times 10 = 8\%$

? more women in 140-150mm interval.

(d) 135-140mm interval

(e) $5 \times 2.1\% = 10.5\%$

(f) 102-103mm

(g) 115-120mm

8.(a) True.

$7.3\%/5 = 1.46\%$ (10-15), $15.6\%/10 = 1.56\%$ (15-25), $15\%/10 = 1.5\%$ (25-35)

? the families that earn between \$10,000 and \$35,000 are spread fairly evenly.

(b) False.

$19.2\%/15 = 1.38\%$ (35-50), $19.6\%/25 = .78\%$ (50-75)

? the percent of families that earn between \$35,000 and \$50,000 is larger than the families earning between \$50,000 and 75,000.

(c) False. In the histogram, the height of a block represents percentage per horizontal unit. Here, the percentage is represented as area. Moreover, the x-axis is not divided by the determined unit.

9.

(a) True.

(b) True

(c) You can look at it in two ways. Maybe there are many students who studied just enough to pass the class. Therefore, many students barely passed and received a GPA of 2. Alternately, perhaps the instructor tended to be "soft" (unlike your TA) to the weakest students pushing them up to a "C."

10.

(a) see attached.

(b) It could be a misreport of either of respondent or surveyor. Consider the low level of education, lack of precise official birth certificate system, and lack of precise census tools in the past. People might not have known their precise birthdate because their birth certificates were not always issued and not everyone could read. Respondents might have answered their age with semi-decade or decade ending, which seems to be simpler and easier to answer. (e.g. "around 40") Because census is taken in years with "0" ending, there might be a tendency to answer with birth years which end in zero.

(c) The census has become more sophisticated over time, as has the issuance of birth certificates due to the adoption of social security and other factors. People are also more likely to be literate.

(d) Even in both times.

Chapter 4.

1.

(a) $(41+48+50+50+54+57)/6 = 50$.

$$[\{(50-41)^2 + (50-48)^2 + (50-50)^2 + (50-54)^2 + (57-50)^2 + (50-50)^2\}/6]^{1/2} = 5$$

? average = 50 & SD = 5

(b) $50 + .5(5) = 52.5$

$$50 - .5(5) = 47.5$$

? 48, 50, 50 are within .5 SDs of average.

$$50 + 1.5(5) = 57.5$$

$$50 - 1.4(5) = 42.5$$

? 48, 50, 50, 54, 57 are within 1.5 SDs of average.

2.

(a) (ii) has smaller SD. Since there is no difference due to the additional three 50s, and it is divided by 10 instead of 7, it generates smaller SD than (i).

(b) (i) has smaller SD. Two additional elements (1 and 99) will enlarge the distances from the average 50 which will way exceed the increased denominator.

3.

(a) 5

(b) Considering that average is 5, its SD should be around 3, since the average plus or minus 2SD should cover 95% of data.

5. Assuming that it has a normal distribution, the lower bound is 96 ($124 - 2 \times 14$) and the higher bound is 152 ($124 + 2 \times 14$). Thus, 80mm and 210mm are way to low and high respectively as compared to the average.

6. (a) (i) average 60

(ii) average 50

(iii) average 40

(b) (i) average < median

(ii) average = median

(iii) average > median

(c) 15

(d) False. (i) seems to be more dispersed, having a larger variance than (iii) in the graph.

7.

(a) Men average = 66, $SD_m = 9$

Women average = 55, $SD_f = 9$

	Average	SD
Men	145.2	19.8
Women	121	19.8

(b) $66 - 9 = 57$, $66 + 9 = 75$. Thus, 1 SD of the average, which includes 68% of men.

(c) Bigger than 9 kg.

It asks what happens if you pool the two variables together. The two variables have the same SD, but different means. Contrary to fact, imagine the two variables had the same mean. Now, split the sample, moving half to the left and half to the right by equal amounts, retaining the same SD of the split sample. The SD of the overall sample will increase. (Or, at least, that's how I'd develop the intuition.)

10.

(a) The best guess is 163.

(b) \$8

Chapter 8.

1.

	Average IQ	SD
Husbands	100	15
Wives	100	15

$$r = .6$$

ranges of x & y : 70-130 (15x2 = 30)

- (a) Averages are out of range.
- (b) Range is too small for x and y.
- (c) Range is too large for x and y.
- (d) Correct scatter diagram.

2.

(a) Negative. As a car gets older, it becomes a gas guzzler and gasoline economy decreases. In addition, new cars have to meet new fuel economy standards. Both factors will conspire to produce a negative correlation between age and fuel economy.

(c) People with higher incomes can afford newer cars which have better gasoline economy than older cars or used cars.

3. The correlation coefficient is 1 because there is a perfect linear relationship.

6. False. There is no direct relationship between two different correlation coefficients, because correlation coefficients are standardized figures.

7. As shown order...

8. .62-1

-.85 .97

.06 -.38

11. **answer : -1**

average right: 6.4 $SD_r = 2$

average wrong : 3.6 $SD_w = 2$

right = 10 - wrong

$$\text{Corr}(r,w) = \text{Cov}(r,w)/(SD_r \times SD_w)$$

$$\text{Cov}(r,w) = \sum (r_i - 6.4)(w_i - 3.6)/n = \sum (r_i w_i)/n - (6.4)(3.6)$$

$$= \sum r_i(10-r_i)/n - (6.4)(3.6) = \sum (10r_i - r_i^2)/n - (6.4)(3.6) = 10\sum r_i/n - \sum r_i^2/n - 23.04$$

$$= 10(6.4) - \sum r_i^2/n - 23.04$$

$$\sum r_i^2/n = \text{Var}(r) + \text{mean}(r)^2 = 4 + (6.4)^2 \quad ? \quad \text{Var}(r) = \sum r_i^2/n - \text{mean}(r)^2$$

$$? \quad \text{Cov}(r,w) = 10(6.4) - (4+6.4^2) - 23.04$$

$$= -4$$

$$? \quad \text{Corr}(r,w) = -4/(2 \times 2) = -1.$$

Intuitively, we can imagine that wrong answers and right answers will have a precise linear negative relationship.

Part B.

1. and 2. should result in the same answers as in Part C
3. There is one BIG outlier in (1), which is not so obvious in the logged version. The logged version may be better

Part C.

1. -0.566
2. -0.503
3. Graphs attached and see log below.

It is now more clear that the log correlation is better, as described in B3.

Problem C Log

```
. corr rating enrollment  
(obs=26)
```

```
-----+-----  
          | rating enroll~t  
-----+-----  
    rating |    1.0000  
enrollment | -0.5655    1.0000
```

```
. // -0.566  
. gen logenrol = log(enrollment)  
. corr rating logenrol  
(obs=26)
```

```
-----+-----  
          | rating logenrol  
-----+-----  
    rating |    1.0000  
logenrol  | -0.5032    1.0000
```

```
. \\ -0.503
```

```
. graph rating enrollment  
. graph rating logenrol  
. graph rating enrollment, xlog  
. summ rating enrollment logenrol
```

```
-----+-----  
Variable |      Obs      Mean   Std. Dev.   Min     Max  
-----+-----  
    rating |      26   5.626923   .7102437     3.8     6.6  
enrollment |      26  48.96154   95.02609      7    501  
    logenrol |      26   3.334779   .8608651   1.94591  6.216606
```

```
. log close
```

Part D.

1. See do file and log below. The correct correlation is 0.475. The relationship is not well described as the association is clearly not linear. Graphing the data makes this clear.

3. Depending on whether you did it as I did, as shown below, or used the weighting you would have gotten the following correct results. Either way, it is clear that central scanning makes a BIG difference in uncounted votes.

Weighted: 0.005, 0.057

Unweighted: 0.006, 0.061

Do File Pset2-D

```
use "E:\My Documents\course_number\fla_precinct_subset.dta", clear
gen resid= undervote + overvote
gen resrate= resid/ total_ball
corr blackrv resrate
sort county
save fla_precinct_subset, replace
use "E:\My Documents\course_number\fla_county_subset.dta", clear
sort county
save fla_county_subset, replace
merge county using fla_precinct_subset
save fla_merged, replace
table centraltab, c(mean resrate)
```

Problem D Log

```
. do Pset2-D
. use "E:\My Documents\course_number\fla_precinct_subset.dta", clear
. gen resid= undervote + overvote
. gen resrate= resid/ total_ball
(70 missing values generated)
. corr blackrv resrate
(obs=5816)
```

	blackrv	resrate
blackrv	1.0000	
resrate	0.4748	1.0000

```
. sort county
. save fla_precinct_subset, replace
file fla_precinct_subset.dta saved
. use "E:\My Documents\course_number\fla_county_subset.dta", clear
. sort county
. save fla_county_subset, replace
file fla_county_subset.dta saved
. merge county using fla_precinct_subset
. save fla_merged, replace
file fla_merged.dta saved
. table centraltab, c(mean resrate)
```

```
-----
CENTRAL |
OR       |
PRECINCT|
TAB     | mean(resrate)
-----+-----
          1 | .0063025
          2 | .0613287
          9 | .0420813
-----
```

.
end of do-file