

Homework 1

Wage	Experience	CollGrad
11	2	1
4	3	0
14	7	0
12	3	1
24	15	0
6	4	0

1. Wages:

$$\text{Mean } (\bar{x}) = \sum x_i/n = (11+4+14+12+24+6)/6 = \mathbf{11.8333}$$

$$\text{Variance } (V_x) = \sum (x_i^2)/n - \sum (x_i/n)^2 = (121+15+196+144+576+36)/6 - (11.833)^2 = \mathbf{41.31}$$

$$\text{Standard Deviation } (S_x) = \sqrt{V_x} = \sqrt{41.47} = \mathbf{6.44}$$

Experiences:

$$\text{Mean } (\bar{x}) = \sum x_i/n = (2+3+7+3+15+4)/6 = \mathbf{5.6667}$$

$$\text{Variance } (V_x) = \sum (x_i^2)/n - \sum (x_i/n)^2 = (4+9+49+9+225+16)/6 - (5.6667)^2 = \mathbf{19.8885}$$

$$\text{Standard Deviation } (S_x) = \sqrt{V_x} = \sqrt{19.8885} = \mathbf{4.45965}$$

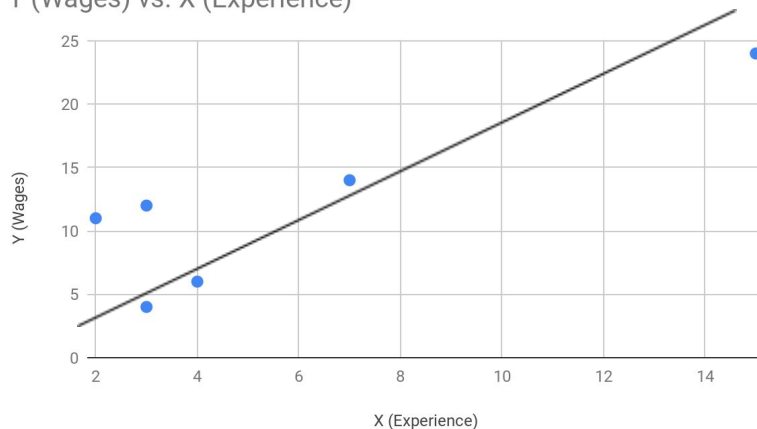
College Graduation (CollGrad):

$$\text{Mean } (\bar{x}) = \sum x_i/n = (1+0+0+1+0+0)/6 = \mathbf{0.3333}$$

$$\text{Variance } (V_x) = \sum (x_i^2)/n - \sum (x_i/n)^2 = (1^2+0^2+0^2+1^2+0^2+0^2)/6 - (0.3333)^2 = \mathbf{0.22221}$$

$$\text{Standard Deviation } (S_x) = \sqrt{V_x} = \sqrt{0.22221} = \mathbf{0.47139}$$

Y (Wages) vs. X (Experience)



The relationship between experience and wage seem to be positive

x_i	y_i	$(x_i - \bar{x}_i)$	$(y_i - \bar{y}_i)$	$(x_i - \bar{x}_i)(y_i - \bar{y}_i)$
11	2	-0.83	-3.67	3.05
4	3	-7.83	-2.67	20.91
14	7	2.17	1.33	2.89
12	3	0.17	-2.67	-0.45
24	15	12.17	9.33	113.55
6	4	-5.83	-1.67	9.74

Covariance(Wage, Experience):

$$[\sum (x_i - \bar{x}_i)(y_i - \bar{y}_i)] / n = (3.05 + 20.91 + 2.89 - 0.45 + 113.55 + 9.74) / 6 = \mathbf{24.95}$$

Yes it is consistent with our linear relationship

Correlation (Wages, Experience):

$$Cov(X, Y) / \sigma_x \sigma_y = 24.95 / (6.44 * 4.46) = \mathbf{0.87}$$

Yes Correlation and Covariance will have same signs since standard deviation of variable can not be negative

2. Data suggests that college graduates wages increase faster with less years of experience than those who didn't go to college.

Covariance (Wages, CollGrad):

$$[\sum (x_i - \bar{x}_i)(y_i - \bar{y}_i)] / n =$$

$$((11-11.8)(-1.033) + (4-11.8)(0-0.33) + (14-11.8)(0-0.33) + (12-11.8)(1-0.33) + (24-11.8)(0.33)) / 5 = \mathbf{-3.45}$$

Our two answers are contradictory because one is positive while the other is negative.

3. New Wages in Euro: (11.67, 4.25, 14.86, 12.74, 25.47, 6.37)

$$\text{Means: } (11.67 + 4.25 + 14.86 + 12.74 + 25.47 + 6.37) / 6 = \mathbf{12.56}$$

Variance:

$$((11.67-12.56)^2 + (4-12.56)^2 + (14.86-12.56)^2 + (12.74-12.56)^2 + (25.47-12.56)^2 + (6.37-12.56)^2) / 6 = \mathbf{47.395}$$

The relationship between the mean and variance of wages in this question versus the first question with the change of currency decreasing the mean by **0.72** but increased variance by **6.08**.

4. $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$

$$\frac{1}{n-1} \sum_{i=1}^n (ax_i - \frac{1}{n} \sum_{i=1}^n ax_i)(by_i - \frac{1}{n} \sum_{i=1}^n by_i) = \frac{1}{n-1} \sum_{i=1}^n (ax_i - \overline{ax_i})(by_i - \overline{by_i}) =$$

$$\frac{1}{n-1} \sum_{i=1}^n a(x_i - \overline{x_i}) b(y_i - \overline{y_i}) = a \cdot b \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x_i})(y_i - \overline{y_i}) = ab\text{Cov}(X, Y)$$

5. Selection Biased, not everyone has a landline anymore, those who usually do are an older generation of folks causing the data to become inaccurate and biased. Non-response biased because most people work a long work day that aligns with work hours of business thus no one will be home to answer the phone.

Selection Biased because sample isn't random if data obtained from you going to ComicCon and not every person goes to ComicCon, and most of your data could be from the same gender. Non-response biased because not everyone would stop to take the survey and missing individuals wouldn't become random.

Sampling method for this could be done using U.S. Census and having random selection done by computer so that person does not select it themselves, it could select 50-60 people and go to their home and have them take the survey.

6. tab white

=1 if white	Freq.	Percent	Cum.
0	98	12.14	12.14
1	709	87.86	100.00
Total	807	100.00	

In the data there are a total of 807 people observed, with 709 of them being white and 98 of them being non-white

sum income

Variable	Obs	Mean	Std. Dev.	Min	Max
income	807	19304.83	9142.958	500	30000

sum income, detail

annual income, \$				
Percentiles	Smallest			
1%	1500	500		
5%	3500	500		
10%	6500	500	Obs	807
25%	12500	500	Sum of Wgt.	807

50%	20000		Mean	19304.83
		Largest	Std. Dev.	9142.958
75%	30000	30000		
90%	30000	30000	Variance	8.36e+07
95%	30000	30000	Skewness	-.2386615
99%	30000	30000	Kurtosis	1.826593

The minimum income is \$500 and maximum income is \$30000. The median income is \$20000.

tab educ

years of schooling	Freq.	Percent	Cum.
6	43	5.33	5.33
8	54	6.69	12.02
10	126	15.61	27.63
12	259	32.09	59.73
13.5	128	15.86	75.59
15	32	3.97	79.55
16	86	10.66	90.21
18	79	9.79	100.00
Total	807	100.00	

Most common level of education is 12 years and makes sense as people graduating high school put 12 years of education.

tab educ if age == 17

years of schooling	Freq.	Percent	Cum.
10	17	100.00	100.00
Total	17	100.00	

There are 17 people with the education level of 17 years.

sum cigs

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	807	8.686493	13.72152	0	80

The mean is 8.686493 and the Standard Deviation is 13.72152

sum cigs if educ > 12

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	325	7.556923	13.99892	0	80

The mean is 7.556923 and the Standard Deviation is 13.99892

7. drop if white == 1

(709 observations deleted)

corr income cigs

(obs=98)

	income	cigs
income	1.0000	
cigs	-0.0028	1.0000

The correlation between income and cigarettes are negative, one reason why this may be true is that people use cigarettes as an alternative form of stress relief.

corr educ cigs

(obs=98)

	educ	cigs
educ	1.0000	
cigs	-0.0562	1.0000

The correlation between education and cigarettes are negative, one reason why this may be true is that people with less education do not understand that cigarettes are bad for you

sum cigs if educ < 12 & income < 19000

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	15	7.666667	9.795529	0	30

sum cigs if educ > 12 & income < 19000

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	15	15.86667	19.4711	0	50

sum cigs if educ < 12 & income > 19000

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	15	8.333333	12.19875	0	40

sum cigs if educ > 12 & income > 19000

Variable	Obs	Mean	Std. Dev.	Min	Max
cigs	29	5.517241	11.90324	0	40

I believe the person with high income and low education will have the highest rate of smoking