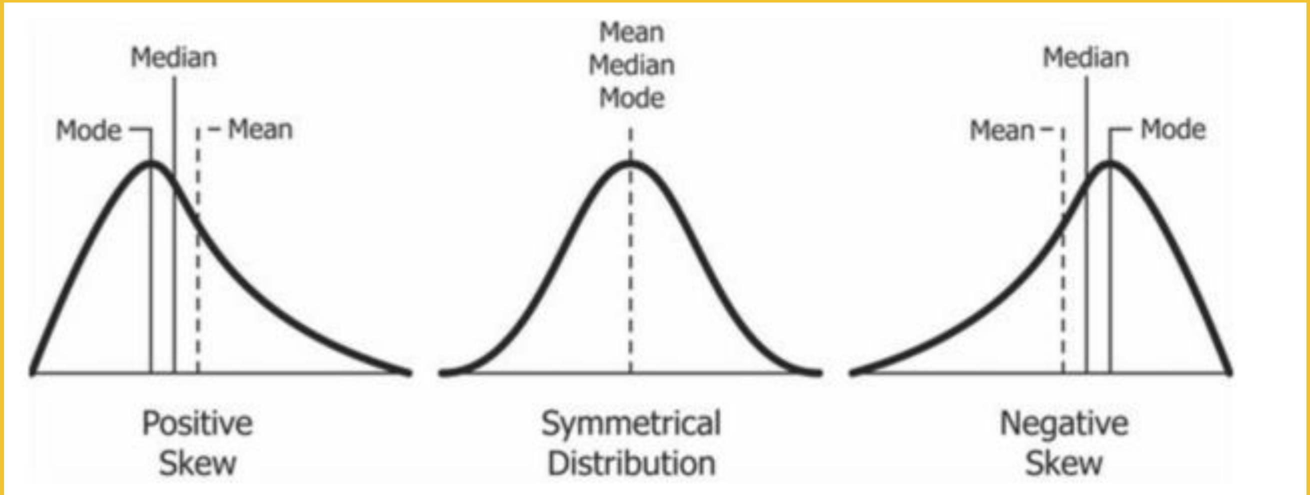


Econ 2250: Stats for Econ

Fall 2022

[Source for pic stats above.]

Announcements

- Homework 5 is due on Sunday

Resources:

- https://www.probabilitycourse.com/chapter3/3_2_2_expectation.php
- https://mixtape.scunning.com/02-probability_and_regression#variance

What we will do today?

- Deep dive on summary operator
- Deep dive on Expected value
- Revisit Variance
- Revisit Covariance
- Introduce Correlation

Summary Operator

$$\Sigma X = x_1 + x_2 + \ldots + x_n$$

Summary Operator

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n$$

Notice that the summary operator is a representation of a function that tells us put add every element of the input. If x is [1,2,3]

$$\Sigma x = x_1 + x_2 + x_3 = 1 + 2 + 3$$

But if x = [‘one’, ‘two’,’three’]

$$\Sigma x = x_1 + x_2 + x_3 = \text{‘one’} + \text{‘two’} + \text{‘three’}$$

Which is of course nonsense/undefined output, but the operator is an function format that defines what to do with the input. This will be an important idea with all operators.

Summary Operation (SIGMA?)

- We often use the greek symbol sigma to represent the summation operator
- It means to sum all of the elements that you pass it
- We often index with the letter ***i*** (meaning an observation) and often use the letter n to represent how many observations.
- Examples:

```
x = [3,12,4]
```

$$\sum_{i=1}^{n} x_i = \text{sum}(x) = x_1 + x_2 + x_3 = 3 + 12 + 4 = 19$$

Summary Operation

- First, some notes on indexing.
- This is a variable we are calling x

```
x = [3,12,4]
```

- It has three elements so we say n=3, where n is length. To get pedantic, this is an array of size (1x3).
- We can refer to the index of **x** that refers to the location in the array

$$x_1 = 3, \quad x_2 = 12, \quad x_3 = 4$$

- The notation for this is $\mathbf{x_i}$ where the ***ith*** refers to the location.

Summary Operation

- The summary operator has properties, all of which preserve the values of a row (it helps me, but might not help you, to think of this as what happens at each iteration of a loop).

```
x = [3,12,4]
```

$$\sum_{i=1}^{n} x_i^2 = \text{sum}(x^2) = x_1^2 + x_2^2 + x_3^2 = 9 + 144 + 16 = 169$$

Summary Operation (SIGMA?)

- We can pass more than one variable into the operator

```
x = [3,12,4]

y = [2,9,1]
```

$$\sum_{i=1}^{n} x_i y_i = \text{sum}(x*y) = x_1*y_1+x_2*y_2+x_3*y_3 =$$

$$3*2 + 12*9 + 4*1 = 118$$

Summary Operation (SIGMA?)

- And we can divide

```
x = [3,12,4]
y = [2,9,1]
sum(x*y)/sum(x**2)
```

$$\frac{\sum x_i y_i}{\sum x_i^2} = \frac{x_1 * y_1 + x_2 * y_2 + x_3 * y_3}{x_1^2 + x_2^2 + x_3^2}$$

$$= \frac{3 * 2 + 12 * 9 + 4 * 1}{3^2 + 12^2 + 4^2}$$

$$= \frac{6 + 108 + 4}{9 + 144 + 16}$$

$$= \frac{118}{169} = 0.6982249$$

SIGMA is a representation of a function

$$f(z) = \Sigma_i^n z = z_1 + z_2 + \ldots + z_n$$

Notice that **z** could be a vector (as we've seen with the variables **x** and **y** above), or another function (like we saw with $\mathbf{x^2}$ and **x*y** above) which could be written

$$f(z) = \Sigma_i^n z = sum(z)$$

$$g(z) = z^2$$

$$f(g(z)) = \Sigma_i^n g(z) = \Sigma_i^n z^2 = sum(z^2)$$

Or, a concrete example

$$\sum_{i=1}^{n} x_i^2 = \mathrm{sum}(x^2) = x_1^2 + x_2^2 + x_3^2 = 9 + 144 + 16 = 169$$

sum(sq(z)) $\sum_{i=1}^{n} x_i^2$

```
sq(z) = z²
def sq(x_in):
 return(x_in**2)

def _sum(old, new):
 return(old + new)

sum_xsq = 0
for i in range(len(x)):
 sum_xsq = _sum(sum_xsq, sq(x[i]))
 print(sum_xy)

Output:
9 (hint: 0 + sq(3))
153 (hint: 9 from above + sq(12))
169 (hint: 153 from above + sq(4))
```

sq(x[0]) + sq(x[1]) + sq(x[2])

$= x_1^2 + x_2^2 + x_3^2$

$= 3^2 + 12^2 + 4^2$

= 9 + 144 +16

= 169

x	sq(x)	cumsum
3	9	9
12	144	156
4	16	169

f(g(x))

Now we will look at a compound function (x + 1). Notice that the summary operator says to $sum(z) = z_1 + z_2 \dots z_{n,}$ so if the thing that we are summing is (x + 1) we just plug that function in

$sum((x + 1)) = (x_1 + 1) + (x_2 + 1) +...+ (x_n + 1)$

x = [3,7,2]

$$\sum_i^n (x_i + 1)$$

(x[0] + 1) + (x[1] + 1) + (x[3] + 1)

= (3 + 1) + (7 + 1) + (2 + 1) = 15

Summary

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n$$

Summary Operator Properties

1.) $\sum_{i=1}^{n} c = nc$

2.) $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$

3.) For any constant a and b: $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{j=1}^{n} y_i$

Gotchas! Be Careful

$$\sum_{i}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}$$

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

Summary Operator Property 1:

$$\sum_{i=1}^{n} c = nc$$

$$\Sigma_1^3 10 = 10 + 10 + 10 = 30 = 3 * 10$$

```
sum_x = 0
for i in range(3):
  sum_x = sum_x + 10
  print( sum_x)
```

```
10
20
30
```

x	cumsum
10	10
10	20
10	30

Summary Operator Property 2:

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$$

$$\Sigma_i^n x_i * c = (3 * 10) + (5 * 10) + (2 * 10)$$
$$= 10 * (3 + 5 + 2) = c * \Sigma_i^n x_i$$

```
x = [3,5,2]
c = 10
sum_x = 0
for i in range(3):
  sum_x = sum_x + x[i] * c
  print( sum_x)

sum_x == sum(x)*c
```

```
30
80
100
True
```

Summary Operator Property 3:

For any constant a and b: $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{j=1}^{n} y_i$

$$\Sigma_i^n (a * x_i + b * y_i) =$$

$$(a * x_1 + b * y_1) + (a * x_2 + b * y_2) + (a * x_3 + b * y_3) =$$

$$a * x_1 + b * y_1 + a * x_2 + b * y_2 + a * x_3 + b * y_3 =$$

$$a(x_1 + x_2 + x_3) + b(y_1 + y_2 + y_3) =$$

$$a\Sigma_i^n x_i + b\Sigma_i^n y_i$$

Expected Value Operator

$$E(x) = \Sigma\, x_i * Pr(x_i)$$

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k)$$
$$= \sum_{j=1}^{k} x_j f(x_j)$$

Notice that the expected value operator is a representation of a function that tells us put add every element of the input. If x is [1,2,3] with probability {⅓,⅓,⅓ }

$$E(x) = x_1 * P(x_1) + x_2 * P(x_2) + x_3 * P(x_3) = 1*⅓ + 2*⅓ + 3*⅓ = 2$$

But if x = [‘H’,’T’’] with prob (.5,.5)

$$E(x) = x_1 * P(x_1) + x_2 * P(x_2) = \text{‘H’}*0.5 + \text{‘T’}*0.5$$

Which is of course nonsense/undefined output, but the operator is an function format that defines what to do with the input. This will be an important idea with all operators.

Summary vs Expected Value Operators

x	p(x)	x * p(x)
3	1/3	1
12	1/3	4
4	1/3	1.33

$\Sigma x = 19$ $E(x) = 6⅓$

$x = [3,12,4]$

$P(x) = 1/3$

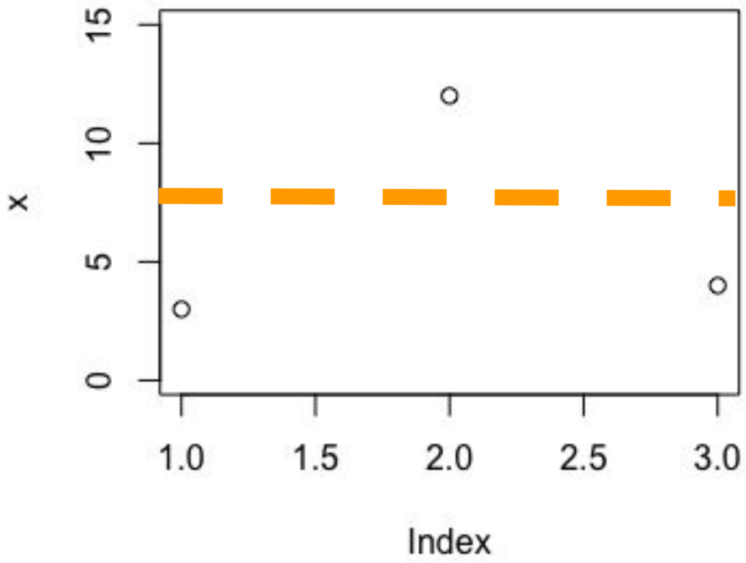
$\Sigma x = x_1 + x_2 + x_3 = 3 + 12 + 4 = 19$

$E(x) = x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3) = 3 * ⅓ + 12 * ⅓ + 4 * ⅓ = 6⅓$

Expected value is a measure of central tendency using probability, E(x) = x*p(x), it is what we expect given our information.

x = [3,12,4]

P(x) = 1/3

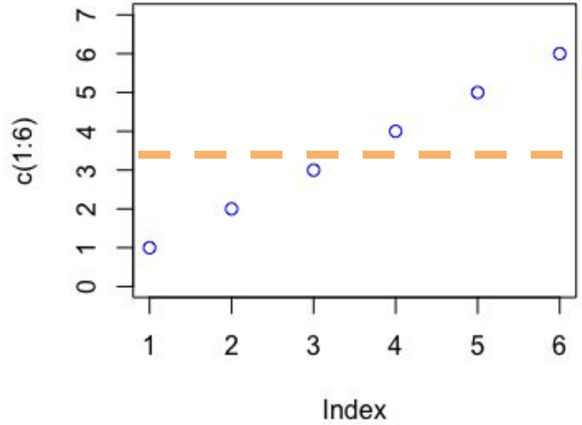


E(x) = 6⅓

Expected value

$x = [1,2,3,4,5,6]$

$p(x) = 1/6$



	x	p(x)	x * p(x)
	1	1/6	0.16666
	2	1/6	0.33333
	3	1/6	0.5
	4	1/6	0.66666
	5	1/6	0.83333
	6	1/6	1
Σ	21	1	3.5

$E(x) = x_1p(x_1) + x_2p(x_2) + x_3p(x_3) + x_4p(x_4) + x_5p(x_5) + x_6p(x_6) =$

$1*⅙ + 2*⅙ + 3*⅙ + 4*⅙ + 5*⅙ + 6*⅙ = 3.5$

When p(x) is uniform (same for all observations) E(x) is the average

$$
\begin{aligned}
\mathrm{E}[X] &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= 3.5 \\
&= \sum (x) * \frac{1}{6} \\
&= \frac{\sum (x)}{6} \\
&= \frac{\sum (x)}{n}
\end{aligned}
$$

$$
\begin{aligned}
E(X) &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) \\
&= \sum_{j=1}^{k} x_j f(x_j)
\end{aligned}
$$

Expected Value Operator Properties

$$E(c) = c$$

$$E(aX + b) = E(aX) + E(b) = aE(X) + b$$

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) \longrightarrow E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

$$E(W + H) = E(W) + E(H), \qquad E\Big(W - E(W)\Big) = 0$$

Expected Value Operator Property 1: $E(c) = c$

C = 10

E(c) = c * p(c) = c* 1 = c

While this is kind of obvious, it will come in handy in lots of proofs.

Expected Value Operator Property 2:

$$E(aX + b) = E(aX) + E(b) = aE(X) + b.$$

x = [3,6,2]

p(x) = ⅓

a = 5

b = 4

E(aX + b) = E(aX) + E(b)

$= ax_1{*}p(x_1) + ax_2{*}p(x_2) + ax_3{*}p(x_3) + b = a(x_1{*}p(x_1) + x_2{*}p(x_2) + x_3{*}p(x_3)) + b$

$= a(E(x)) + b = 5(3{*}\tfrac{1}{3} + 6{*}\tfrac{1}{3} + 2{*}\tfrac{1}{3}) + 4 = 22\tfrac{1}{3}$

An important extension of this linearity is that

$$E(W + H) = E(W) + E(H)$$

Expected Value Operator Property 3:

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) \longrightarrow E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

The above left is a another way to distribute out the linearity in property 2

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n)$$

And the right is the special case when a = 1.

Variance

$$V(X) = E((X - E(X))^2)$$

Variance is a measure of the spread of the data

We get the central tendency using the expected value E(X), and to get a measure of the spread of the data we take the expectation of the squared deviations

Expected value of X: $E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$

which is the average for equally weighted data $E(X) = \frac{\sum x}{n} = \mu_x$

To get a deviation we subtract off the mean

$$\text{deviation of x} = X - \mu_x = X - E(X)$$

And square this so that it does sum to zero

$$\text{squared deviation of x} = (X - \mu_x)^2 = (X - E(X))^2$$

Example: demean squared

x= [3,12,4]

E(x) = mu = xbar = 3*⅓ + 12*⅓ + 4*⅓= (3+12+4)/3=6.3

x	mu	demean	demean_sq
3	6.3	-3.3	11.1
12	6.3	5.7	32.1
4	6.3	-2.3	5.4

Expectation is our best guess of what something will equal, so take the expectation of the squared deviation

$$E[(X - \mu_x)^2] = \sum (x_i - \mu_x)^2 * P(x_i)$$

if $P(x_i)$ is $\frac{1}{n}$ for all $i = 1, 2, \ldots, n$

$$V(X) = \sum (x_i - \mu_x)^2 * \frac{1}{n} = \frac{1}{n} \sum (x_i - \mu_x)^2$$

From the example above

$$\sigma^2 = \frac{1}{n}\sum(x_i - \mu_x)^2$$

x	mu	demean	demean_sq
3	6.3	-3.3	11.1
12	6.3	5.7	32.1
4	6.3	-2.3	5.4

Sum of demean_sq = 11.1 + 32.1 + 5.4 = 48.6

48.6/3 = 16.2

But, notice our deviations (-3.3, 5.7, -2.3), 16.2 is an awful absolute value estimate. That is because we squared the errors, and x is in levels (not squared).

Standard Deviation = square root of σ^2 , sqrt(16.2) = 4.02

Variance Overview

$$V(X) \equiv \sigma^2 = E[(X - E(X))^2]$$

Population model:

$$V(X) = \sigma^2 = E[(X - E(X))^2]$$
$$= E[(X - \mu_x)^2] = \sum (x_i - \mu_x)^2 * P(x_i)$$

if $P(x_i)$ is $\frac{1}{n}$ for all $i = 1, 2, \ldots, n$

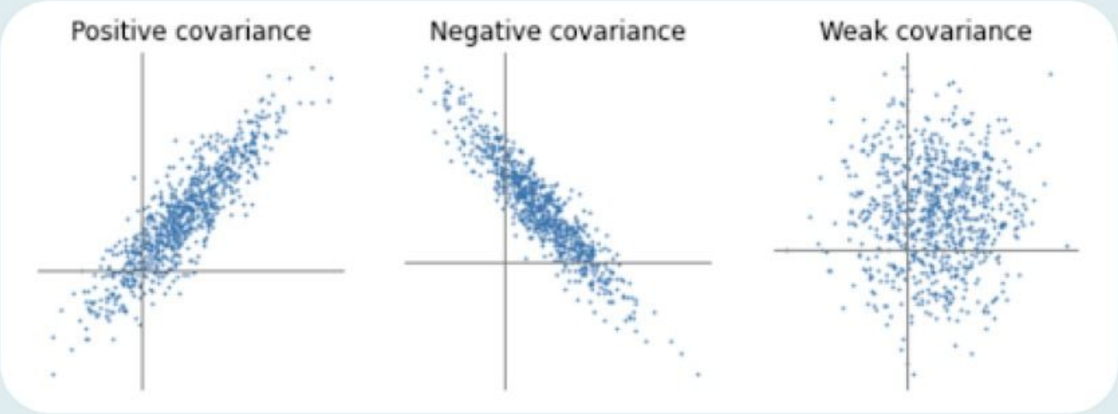
$$V(X) = \sum (x_i - \mu_x)^2 * \frac{1}{n} = \frac{1}{n} \sum (x_i - \mu_x)^2$$

bring squared values back the units of x

$$\sqrt{V(X)} = \sqrt{\frac{1}{n} \sum (x_i - \mu_x)^2}$$

Nice correlation app

https://shiny.rit.albany.edu/stat/rectangles/



Covariance

$Cov(X, Y)$ = E[(X-E(X)(Y-E(Y)]

image

Covariance

$$
\begin{aligned}
Cov(x, y) &= E[(X - E(X))(Y - E(Y))] \\
&= \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n - 1}
\end{aligned}
$$

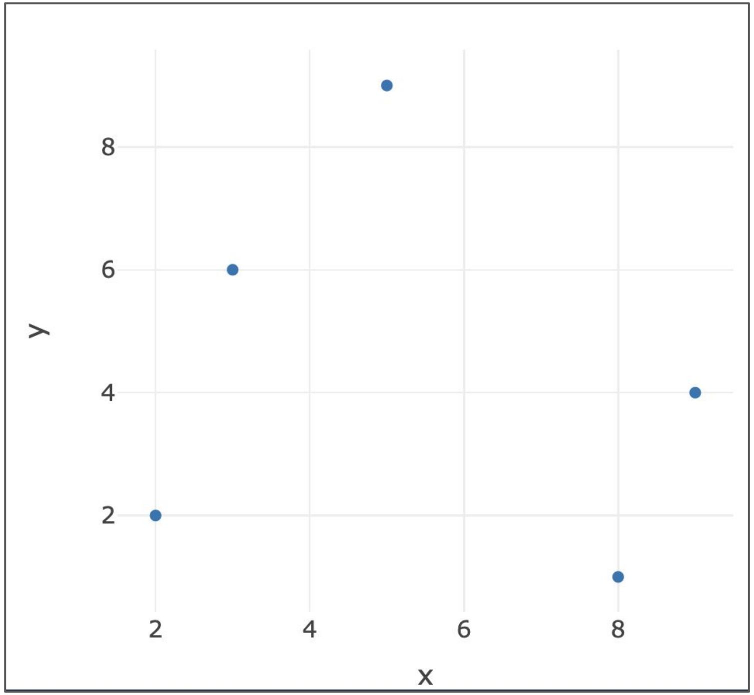
Example covariance

$$\frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$

x	y	demean_x	demean_y	demean_x*demean_y
3	6	-2.4	1.6	-3.84
5	9	-0.4	4.6	-1.84
2	2	-3.4	-2.4	8.16
8	1	2.6	-3.4	-8.84
9	4	3.6	-0.4	-1.44

-7.8	sum
-1.95	sum/(n-1)

mean_y	4.4
mean_x	5.4



Correlation

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \quad \text{if } \sigma_X \sigma_Y > 0$$

$$r_{xy} \quad \stackrel{\text{def}}{=} \quad \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 \sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}$$

Example

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x	y	demean_x	demean_x_sq	demean_y	demean_y_sq	demean_x*demean_y	
3	6	-2.4	5.76	1.6	2.56	-3.84	
5	9	-0.4	0.16	4.6	21.16	-1.84	
2	2	-3.4	11.56	-2.4	5.76	8.16	
8	1	2.6	6.76	-3.4	11.56	-8.84	
9	4	3.6	12.96	-0.4	0.16	-1.44	
			37.2		**41.2**	-7.8	sum
						-1.95	sum/(n-1)

mean_y	4.4
mean_x	5.4

numerator	**-1.95**	-1.95	**-0.22**	**correlation**
denom	sqrt(**37.2** + **41.2**)	8.85		

End of class form

https://forms.gle/kgT2w9wPZo3vJcjA8