Ch 3. Summarizing Distributions

- Central Tendency: mean, median & mode
- Other Measures of Central Tendency
- Comparing Central Tendency
- Measures of Variability: Range, IQR, Variance
- Linear Transformation of variable
- Variance Sum Law I

Compute mean and variance from the **population** of size *N* :

Population Mean $\mu = E[X] = \frac{\sum X}{N}$

Population Variance $\sigma^2 = E[(X-\mu)^2] = \frac{\sum (X-\mu)^2}{N}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$

$\sigma^2 = E[X^2] - \mu^2$

N=7

Value	Deviation from mean	Squared deviation form mean
4	-3	9
5	-2	4
5	-2	4
7	0	0
8	1	1
9	2	4
11	4	16
Mean = μ	average deviation from mean	Average squared deviation from mean = σ^2
7	0	38/7=5.43

- Suppose we sample n values from a larger population of size N

Estimate mean and variance of the **population** from a **sample** of size *n* :

Sample Mean $$\bar{x} = \frac{\sum X}{n}$$

Sample Variance (unbiased) $$s^2 = \frac{\sum (X - \bar{x})^2}{n-1} \quad \text{or} \quad \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

Why n-1 and NOT n?

Intuition: Consider the **extreme** case: **n=1** and sampled value is **x**

- Variance estimate would be 0 **regardless** of the population if denominator = n
- 1 sample **cannot** give information for population variance (consistent with s^2=0/0)

Ch 3. Summarizing Distributions

- Central Tendency: mean, median & mode
- Other Measures of Central Tendency
- Comparing Central Tendency
- Measures of Variability: Range, IQR, Variance
- Linear Transformation of variable
- Variance Sum Law I

- Linear transformation of variable

  - Transform data from one measurement scale to another

Eg. Consider a taxi trip from point A to B. The taxi service initial charge is $3 and additional $0.50 per km for the trip. Let *y* be the cost of the taxi ride and *x* be the distance travelled. We have:

$$y = 0.5x + 3$$

The mean of y = 0.5 (mean of x) + 3

$$\mu_Y = \frac{1}{N}\sum Y = \frac{1}{N}\sum 0.5X + 3 = 0.5\mu_X + 3$$

The variance of $y = 0.5^2$ (variance of x)

$$\sigma_Y^2 = \frac{\sum(Y - \mu_Y)^2}{N} = \frac{\sum(0.5X - 0.5\mu_X)^2}{N} = 0.5^2\sigma_X^2$$

- Variance Sum Law I

  - Linear combination of 2 uncorrelated variables

Eg. Consider a delivery truck carrying x units of item A and y units of item B. The weight of each item A is 5kg while each item B is 10kg. The total weight carried is:

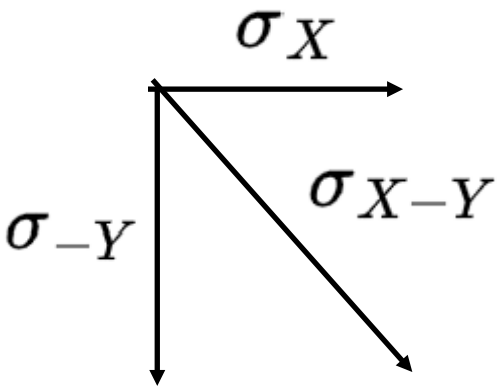
$$t = 5x + 10y$$

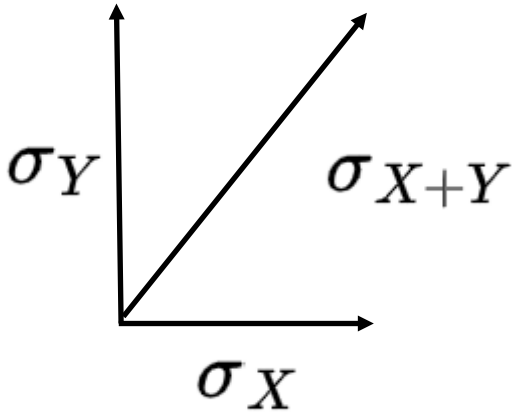
Weight difference d :

$$d = 5x - 10y$$

Mean = 5 (mean of x) – 10 (mean of y)

Variance = $5^2\,\sigma_x^2 + 10^2\,\sigma_y^2$





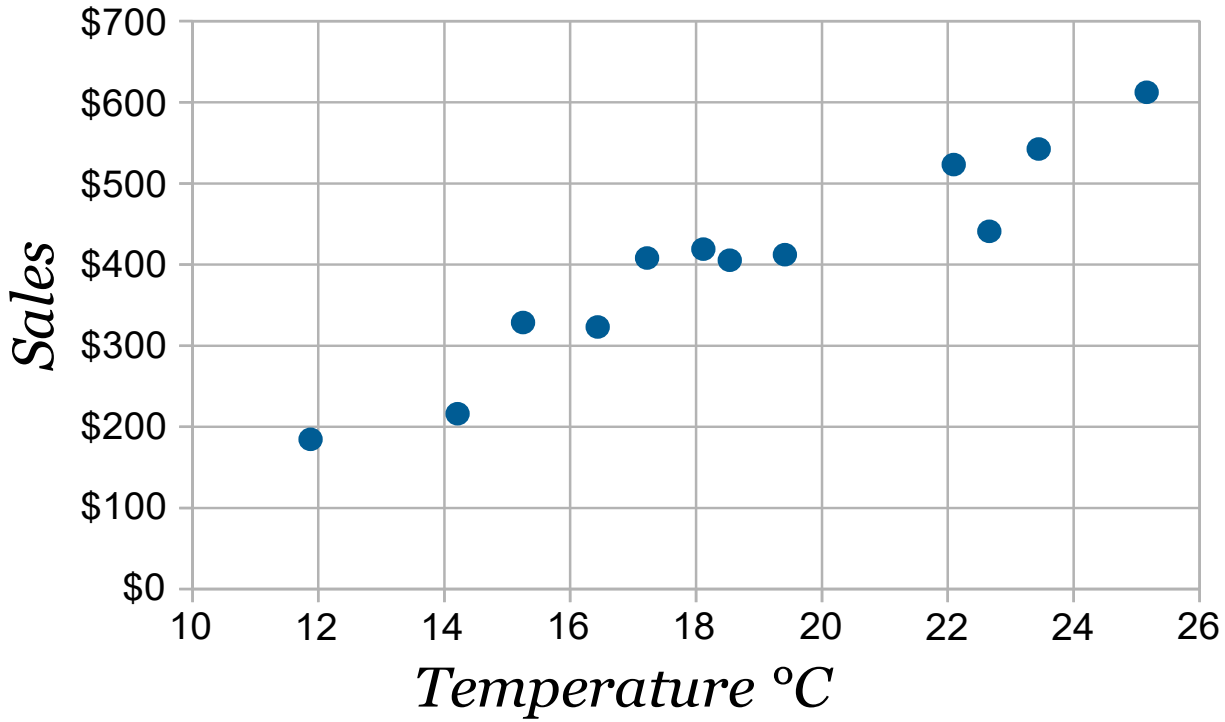
Ch 4. Bivariate Data

- Introduction to Bivariate Data
- Pearson Correlation and Covariance
- Properties of Person Correlation
- Variance Sum Law II

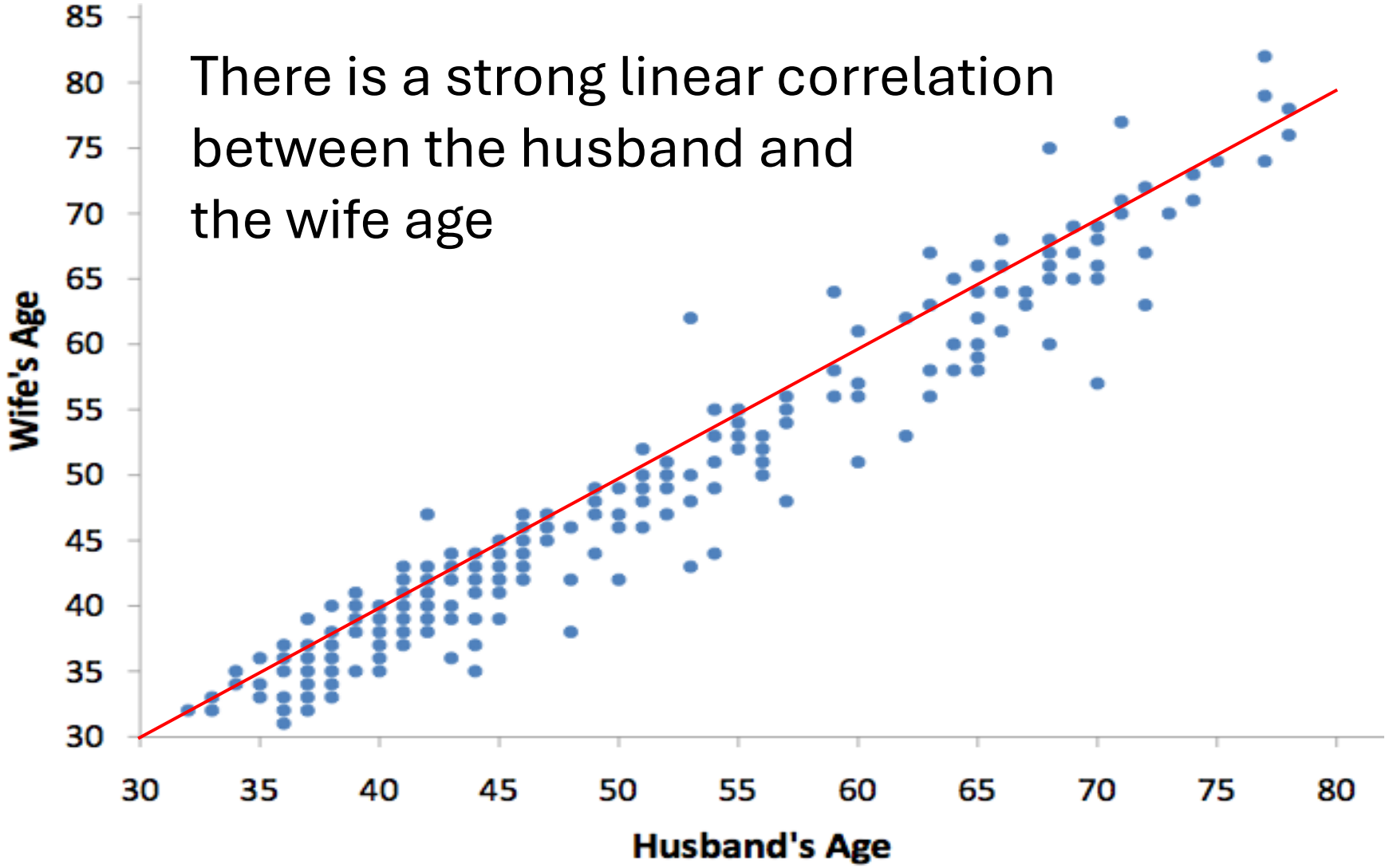
- Bivariate Data

  - A dataset with a pair of variables which may be correlated to one another.

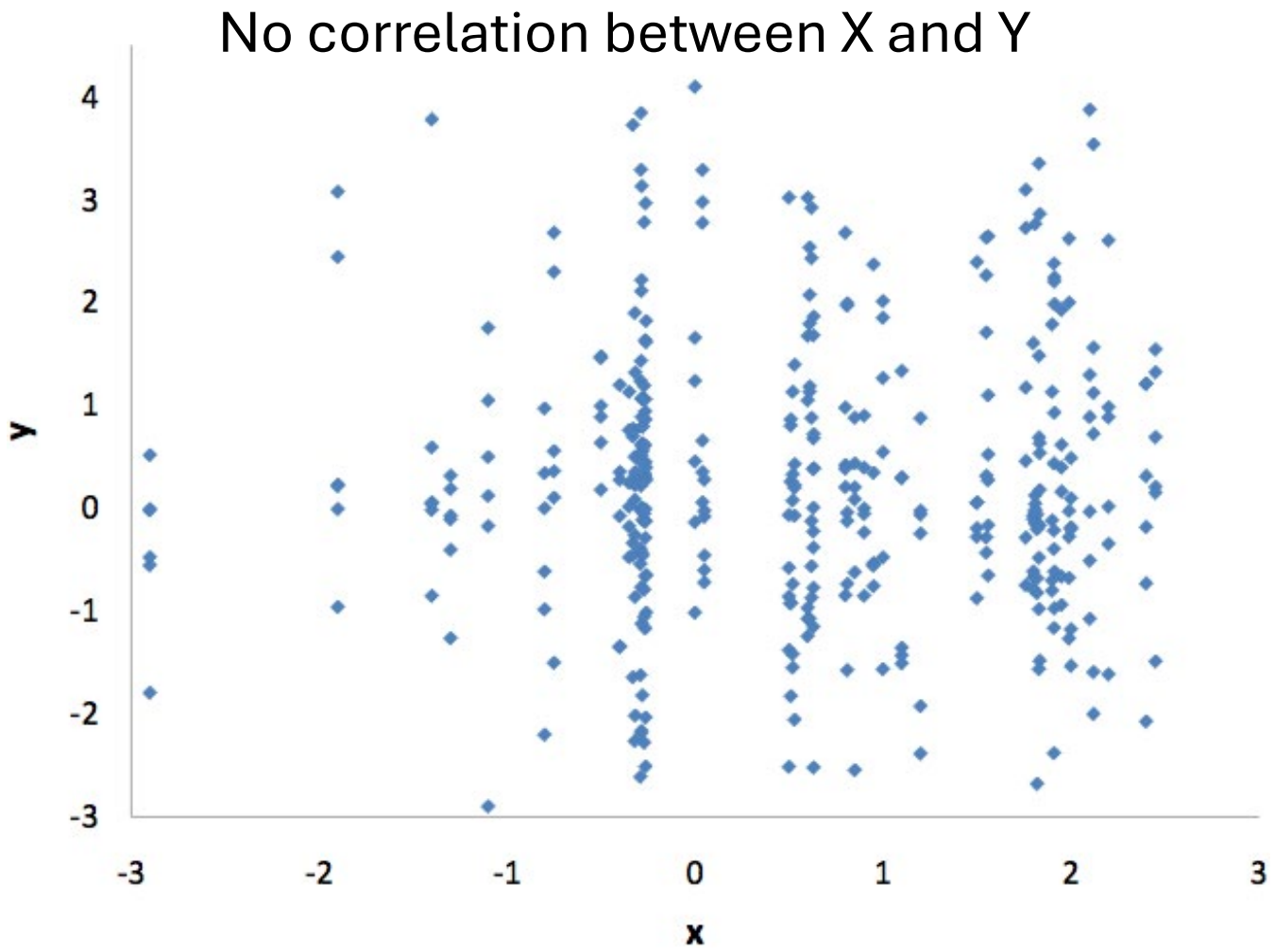
Eg: two variables – ice cream sales and temperature



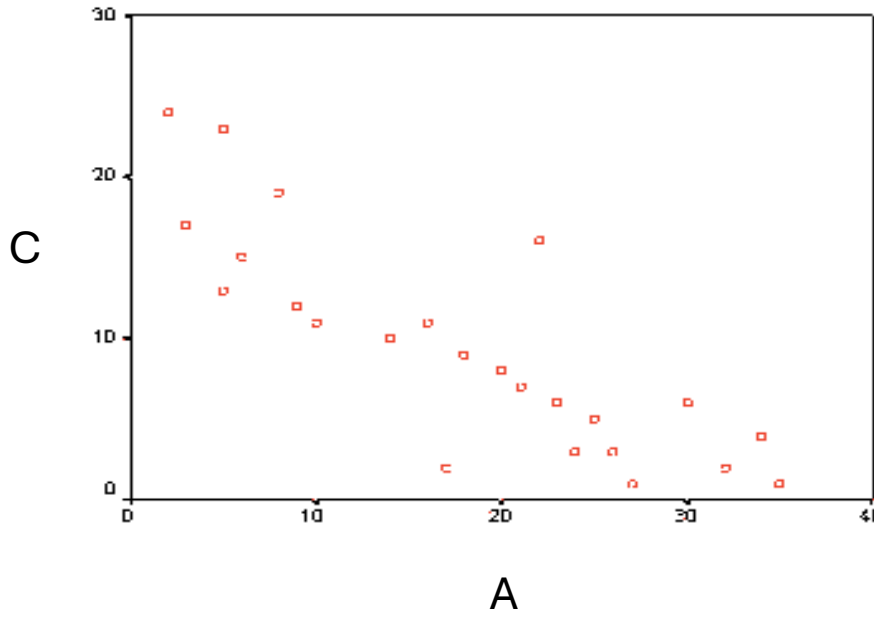
- Bivariate Data



Eg: Correlation of variables X and Y



Question: Describe the relationship between variables A and C. Think of things these variables could represent in real life.



Negative relationship between A and C.
There is a negative relationship between price and quantity of the products that we buy.

Pearson Correlation ρ

- An indicator on the strength of the linear relationship between two variables.

Definition: $\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$

Covariance of X and Y, denoted as cov(XY)

$$= \frac{E[XY]-\mu_X\mu_Y}{\sqrt{E[X^2]-(\mu_X)^2}\sqrt{E[Y^2]-(\mu_Y)^2}}$$

$$= \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

If $\mu_X = \mu_Y = 0$, then $\rho = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$

- Computation of Correlation based on a sample of size n

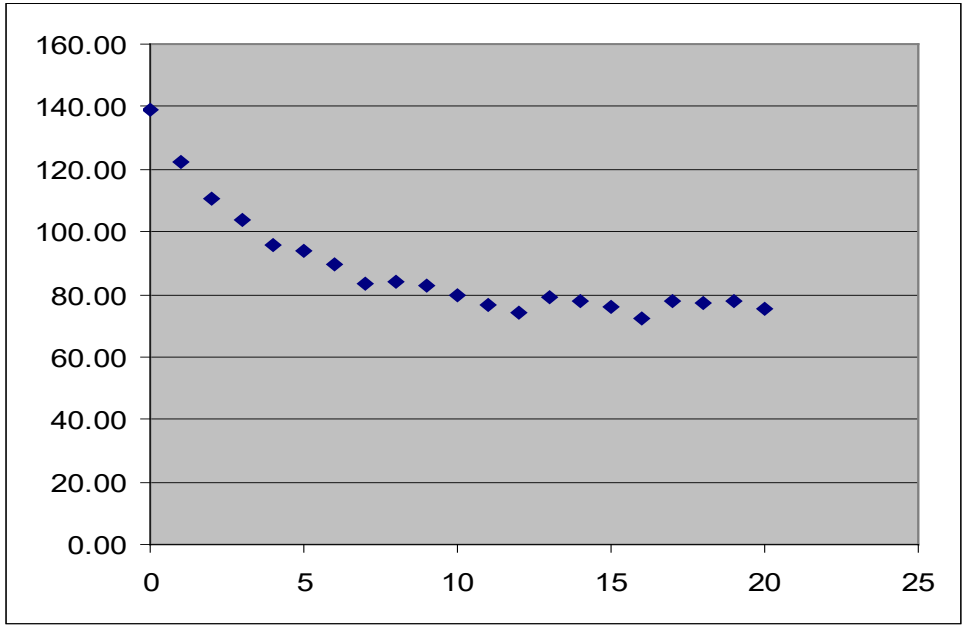
$$\text{cov(XY)} = \frac{1}{n-1}\sum (X-\bar{X})(Y-\bar{Y})$$

$$\text{Correlation } r = \frac{E[(X-\bar{X})(Y-\bar{Y})]}{s_X s_Y}$$

$$= \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

$$\text{If } \bar{X} = \bar{Y} = 0, \text{ then } r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$$

Given the data shown in the figure, is it appropriate to use Pearson Correlation to describe the relationship between Xand Y?



Eg: Given the data set, calculate σ_X, σ_Y, cov(X,Y) and the Pearson Correlation.

X	2	5	6	8	9
Y	8	5	2	4	1

Mean of X	6				
dev from mean X	-4	-1	0	2	3
Sq dev from mean X	16	1	0	4	9
Mean of Y	4				
dev from mean Y	4	1	-2	0	-3
Sq dev from mean Y	16	1	4	0	9
(X-μ_X)(Y-μ_Y)	-16	-1	0	0	-9

$$\sigma_X = \sqrt{30/5} = 2.45$$

$$\sigma_Y = \sqrt{30/5} = \sqrt{6} = 2.45$$

$$Cov(X,Y) = \frac{-26}{5} = -5.2$$

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = -\frac{5.2}{2.45 * 2.45} = 0.87$$

- Properties of Correlation

  - Value in the range of [–1, +1]

  - Symmetric: correlation of X with Y
    = correlation of Y with X

  - Unaffected by linear transformations:

    Correlation of Y with X
    = correlation of Y with A X + B
    where A and B are constants

Eg: If the correlation between weight (in pounds) and height (in feet) is 0.58, find:

(a) the correlation between weight (in pounds) and height (in yards)

(b) the correlation between weight (in kilograms) and height (in meters).

The correlation for both (a) and (b) is still 0.58 because linear transformations do not affect the value of Pearson's correlation, and both of the above instances are linear transformations.

Ch 4. Bivariate Data

- Introduction to Bivariate Data
- Pearson Correlation and Covariance
- Properties of Person Correlation
- Variance Sum Law II

- Variance Sum Law II

- Linear combination of 2 independent variables X and Y

Variance of $X \pm Y$: $\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$

- If the variables X and Y are correlated

Variance of $X \pm Y$: $\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y \pm 2\rho\sigma_x\sigma_y$

- For computation based on a sample

$$s^2_{x \pm y} = s^2_x + s^2_y \pm 2rs_xs_y$$

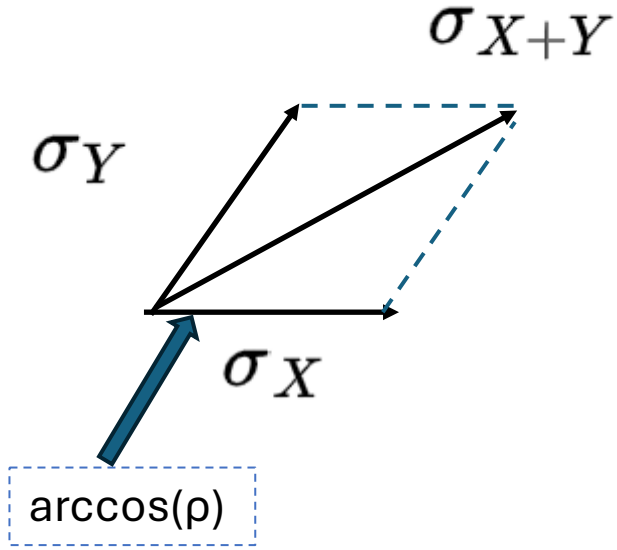
Eg: Students took 2 parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 49. The correlation between the test scores is 0.6.

(i) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be?

(ii) What would the variance of Part A - Part B be?

(i) Var (A + B) = 25 + 49 + 2*0.6*√25*√49
= 116

(ii) Var (A – B) = 25 + 49 - 2*0.6*√25*√49
= 32



Justification for sample variance formula: $s^2 = \frac{\sum(X-\bar{x})^2}{n-1}$

Suppose the denominator of s^2 is n, instead of $(n-1)$:

$$s^2 = \frac{\sum(X-\bar{x})^2}{n} = \frac{1}{n}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)$$

For unbiased estimate, we expect the mean of s^2 to be equal to σ^2 :

$$E[s^2] = E\left[\frac{1}{n}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\right]$$

$$= \frac{1}{n}\left(\sum \underbrace{E[X^2]}_{\sigma^2 + \mu^2} - \frac{E[(\sum X)^2]}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

$$E[s^2] = \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

Let $Y = \sum X$

$E[Y^2] = \sigma_Y^2 + \mu_Y^2$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{Var[\sum X] + (E[\sum X])^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{\sum Var[X] + (\sum E[X])^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{n\,\sigma^2 + (n\,\mu)^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \sigma^2 - n\,\mu^2\right) \quad = \frac{1}{n}(n - 1)\,\sigma^2$$

If the denominator is $(n - 1)$, then the mean of s^2 is equal to σ^2, i.e. $E[s^2] = \sigma^2$