STATISTICS CLASS 2

- Agenda for today
  - Sample v. Population
  - Moments
  - Covariance
  - Correlation

- Sample v. Population

  Population includes all data of a specified group.

  Sample is a subset of the population.

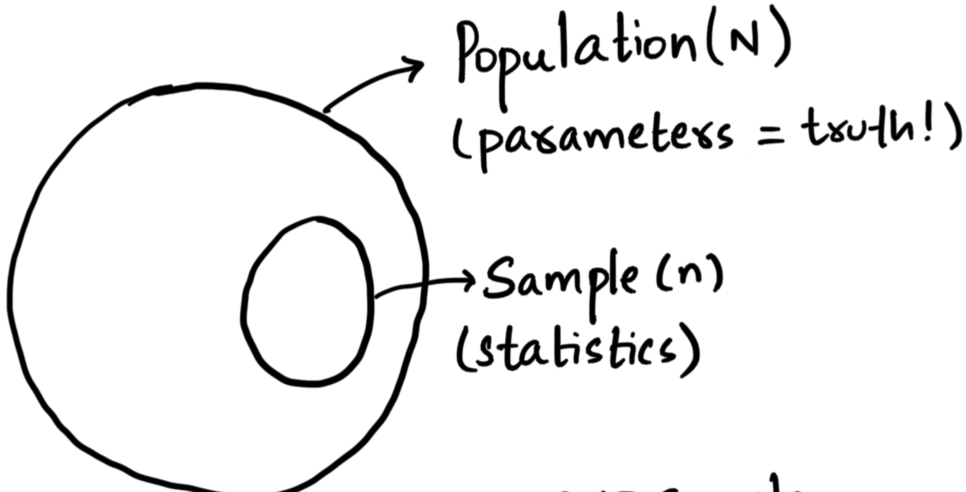
	Population	Sampling
Avg. Height of people in USA	~330 Mn	Select people from each state?
Avg. Height of people in UCLA	~ 50 k	MFE, MBA, Profs?
Avg. Weight of people in Japan	~125 Mn	Volunteering booths at diff

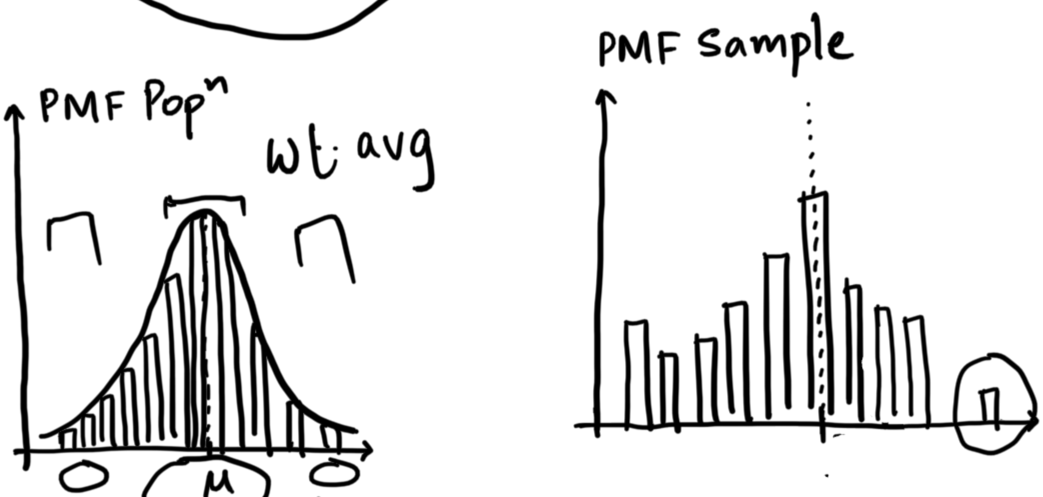
places in Japan?

Sampling bias → Undercoverage (ex:2)
→ Advertising (ex:3)
→ Non-response (ex:1)

Random sample → gold standard
(good proxy/less bias)

Ex1: Avg. Height of people in USA

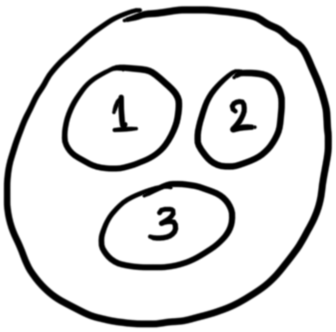




Population dist.

Distribution of sample data

Ex 2: Distribution of sample statistic

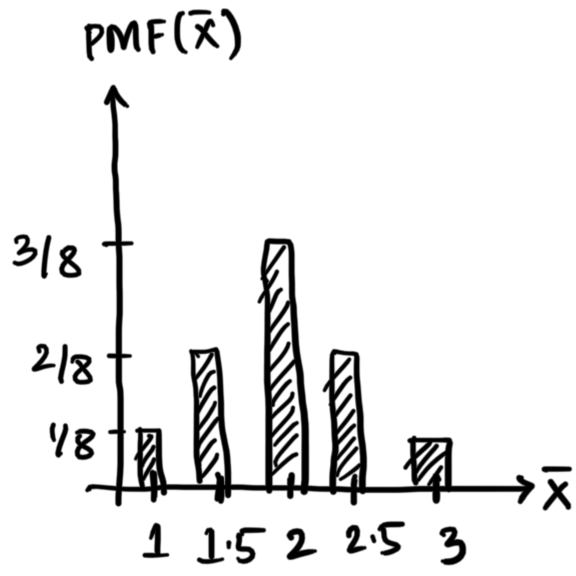


Population mean $= \mu = \dfrac{1+2+3}{3} = 2$

Sampling = 2 draws with replacement

Samples	Mean ($\bar{x}$)
(1,1)	1
(1,2)	1.5
(1,3)	2
(2,1)	1.5
(2,2)	2
(2,3)	2.5
(3,1)	2
(3,2)	2.5
(3,3)	3

Mean ($\bar{x}$) → proxies for μ



Depending on who samples from popⁿ, $\bar{x}$ changes

Similarly we can get PDF of other sample stats.
- → Mean (Normal) ✓
- → Variance (χ^2)
- ↳ Median / Quantiles

↳ skewness/kurtosis

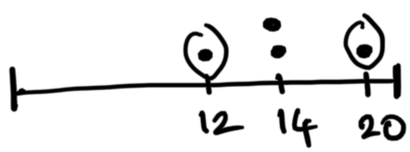
- Moments

Robust ways of summarizing data

$\{ \underbrace{\text{Mean, var, std dev}}, \text{Quantile}, \underbrace{\text{skew, kurt.}} \}$

Sample data 1

$X = \{ \underbrace{12, 14, 14, 20} \}$



Sample data 2

$Y = \{15, 15, 15, 15\}$

$$\bar{X} = \frac{12 + 14 + 14 + 20}{4} = 15$$

$$\bar{Y} = 15$$

Mean = Avg. distance from 0

↳ First moment

$$\underbrace{\frac{1}{n} \sum x_i^2} = \frac{12^2 + 14^2 + 14^2 + 20^2}{4} = 234$$

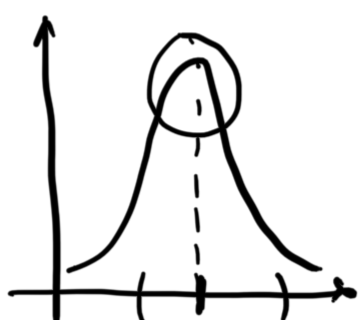
$$\frac{1}{n} \sum y_i^2 = 15^2 = 225$$

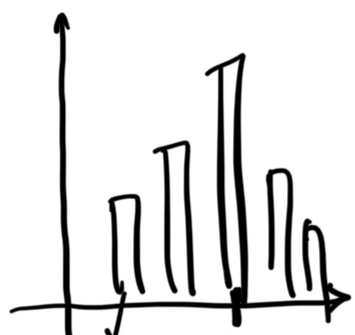
↳ squared dist. from 0 (increases with more spread

Alternatively we can compute dist. from mean

$$\underbrace{\frac{1}{n}\sum (x_i-\bar{x})^2}_{\text{centered } 2^{nd} \text{ moment}} = \frac{(-3)^2+2\cdot(-1)^2+(5)^2}{4} = 9 \qquad \frac{1}{n}\sum(y_i-\bar{y})^2 = \underbrace{0}_{\text{No spread in data}}$$

Remove effect of 1^{st} moment ($\bar{x}$) to capture any additional info. in the dataset.





Moment	Population (param)	Sample (statistic)
(1) Mean	$\mu = E[X] = \int x_i \cdot \underbrace{f_X(x)} \, dx$ $\mu = \sum x_i/N$	$\bar{X} = m = \sum_{i=1}^{n} x_i/n$

(2) Variance

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

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

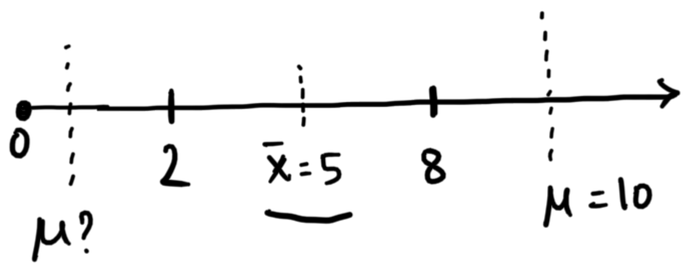
$$s^2 = \underbrace{\sum_{i=1} (x_i - m)^2 / n-1}$$

(3) skewness

(4) kurtosis

- Degrees of freedom

  - Intuition 1: Say $X = \{2, 8\}$



$$\frac{1}{n-1} \sum \underbrace{(x_i - \bar{x})^2} = (2-5)^2 + (8-5)^2 = 18$$

$$\frac{1}{n} \sum (x_i - \mu)^2 = (2-4)^2 + (8-4)^2 = 20 //$$

$$\sum (x_i - 6)^2 = (2-6)^2 + (8-6)^2 = 20 //$$

squared deviation at mean is smallest at $\bar{x}$

[s^2 computes variance centered at $\bar{x}$

[σ^2 computes variance centered at μ

If $\underbrace{s^2(\bar{x})}$ $= \frac{1}{n} \sum (x_i - \bar{x})^2$ } Function that takes $\bar{x}$ as input

then $\underbrace{s^2(\mu)} \geq \underbrace{s^2(\bar{x})}_{\text{lower bound}}$

Division by $n-1$ $\uparrow$ $s^2(\bar{x})$ to be closer to $s^2(\mu)$

- Intuition 2:

Say $X \sim N(\mu, \sigma^2)$ } True data generation process

If $\mu = 0, \sigma = 1$

[Population]: $I = \{0.5, 0.1, -0.2, 0.3, -1, 1.5\}$

[Sample]:

x	$x - \bar{x}$
1.5	1.3
0.1	-0.1
-1	-1.2

- If we know $\bar{x}$, one data point is redundant

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

free data points

$\bar{x} = 0.2$	0	- $\exists$ $n-1$ free data points after computing the mean!

[DF] : No: of data points available to compute sample statistics

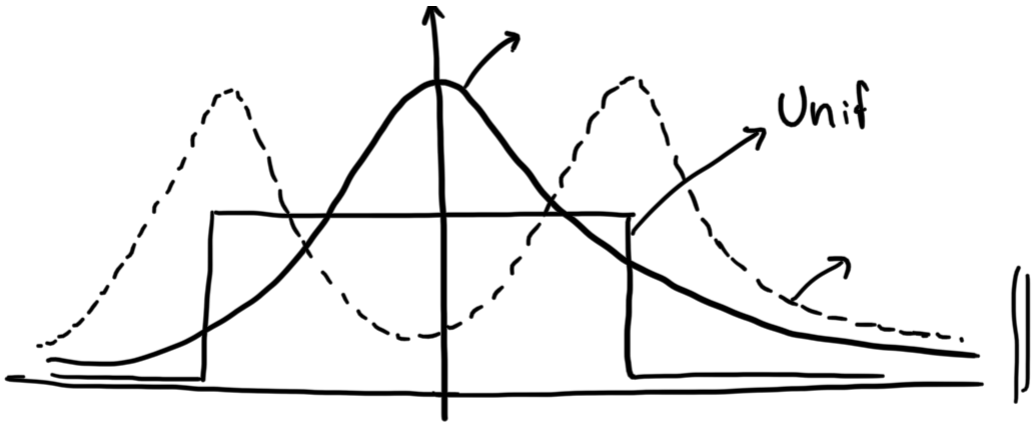
Sample stat	DF
Mean	n
Var	$n-1$
skew	$n-2$
kurtosis	$n-3$
$\vdots$	$\vdots$

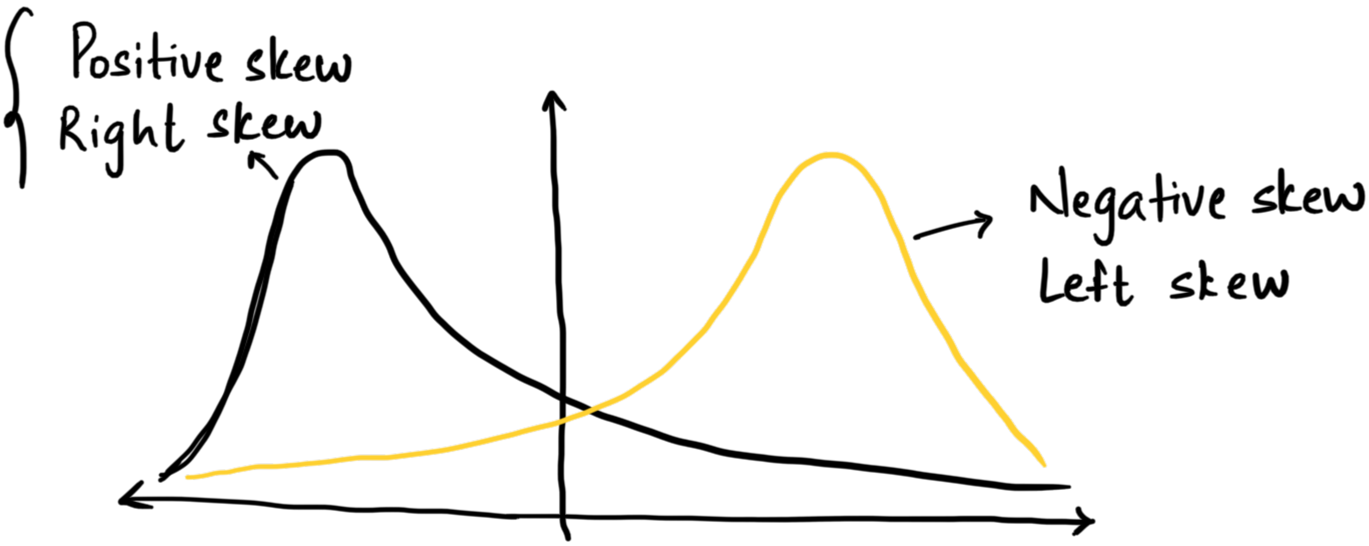
<u>Ex</u> : $X = \{44\}$ sample

(a) $\bar{x} = 44$

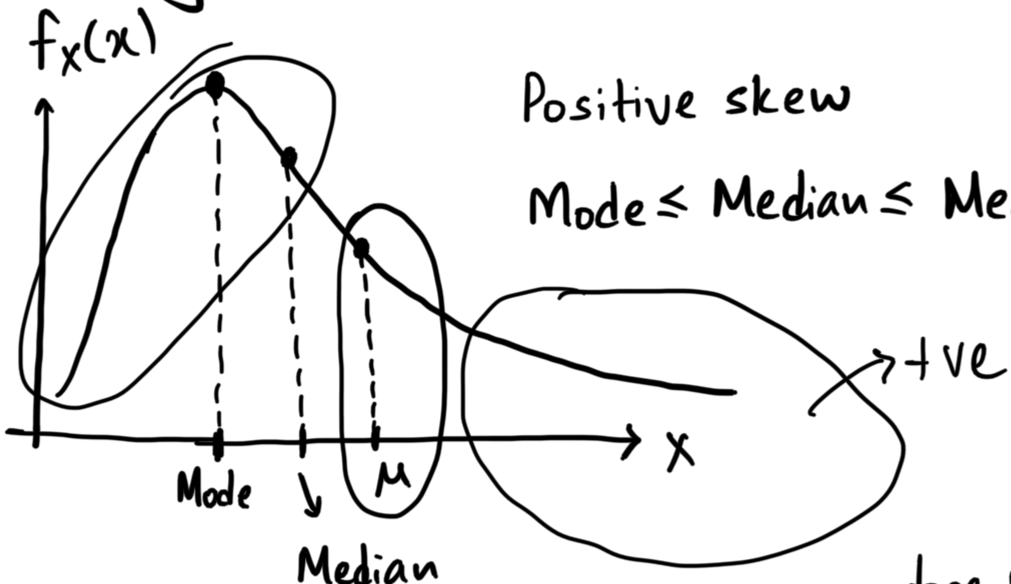
(b) $s^2 = \dfrac{(44-\overline{44})}{\underbrace{n-1}_{=0}} = \text{NaN}$

- Skewness





which way does the tail point?



Positive skew

Mode ≤ Median ≤ Mean

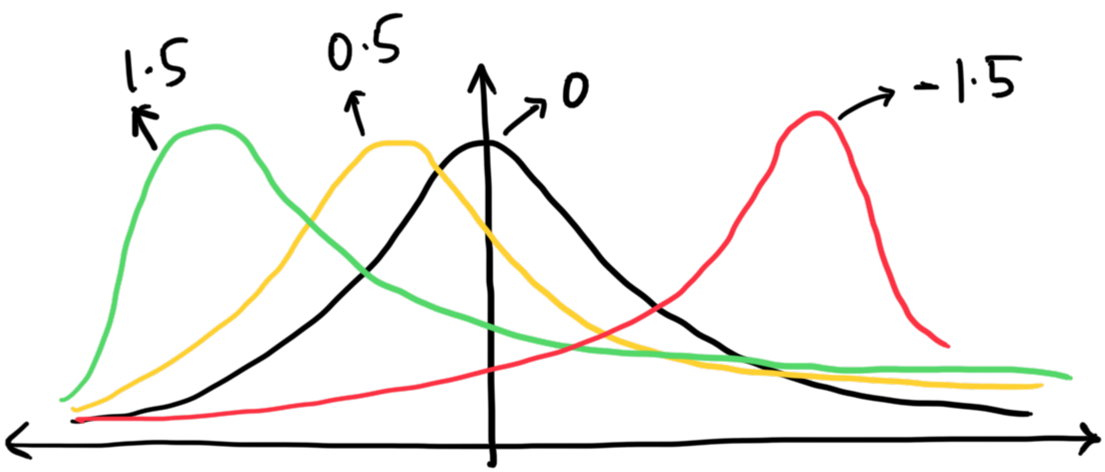
Mean − mode } does poorly on small

$$\text{Pearson's mode skewness} = \frac{\quad}{\text{std. dev}} \Bigg\}\ \text{samples}$$

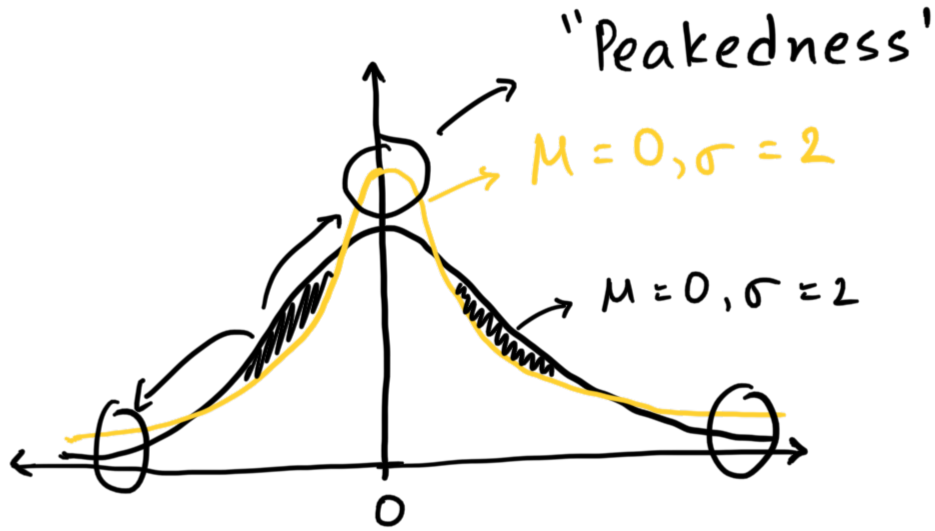
$$\text{Pearson's median skewness} = \frac{(\text{Mean} - \text{Median}) \times 3}{\text{std. dev}}$$

$$\text{Moment estim. skew} = \frac{1}{N} \underbrace{\left[\frac{\sum (x_i - \mu)^3}{\sigma^3} \right]}_{\text{Removes effect of mean and variance}} = \frac{1}{N} \sum \left(\frac{x - \mu}{\sigma} \right.$$

$\text{Skew}(N(\cdot,\cdot)) = 0$



- Kurtosis



Orange curve has fatter tails

Kurtosis measures weight on tails

Moment estim. kurtosis = $\frac{1}{N} \frac{\sum (x_i - \mu)^4}{\sigma^4}$

kurtosis $(N(0,1)) = 3$, kurt $\in [1, \infty)$

If kurtosis > 3 (leptokurtic)

kurtosis < 3 (platykurtic)

Excess kurtosis $= \underbrace{\text{kurt} - 3}$

sets Excess kurt. of

std. normal $= 0$

- Covariance

Captures joint variation of 2 RVs X, Y

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
100	10	0	0	0
102	9	2	-1	-2
98	11	-2	1	-2
110	14	10	4	40
90	6	-10	-4	40
$\bar{X} = 100$	$\bar{Y} = 10$			

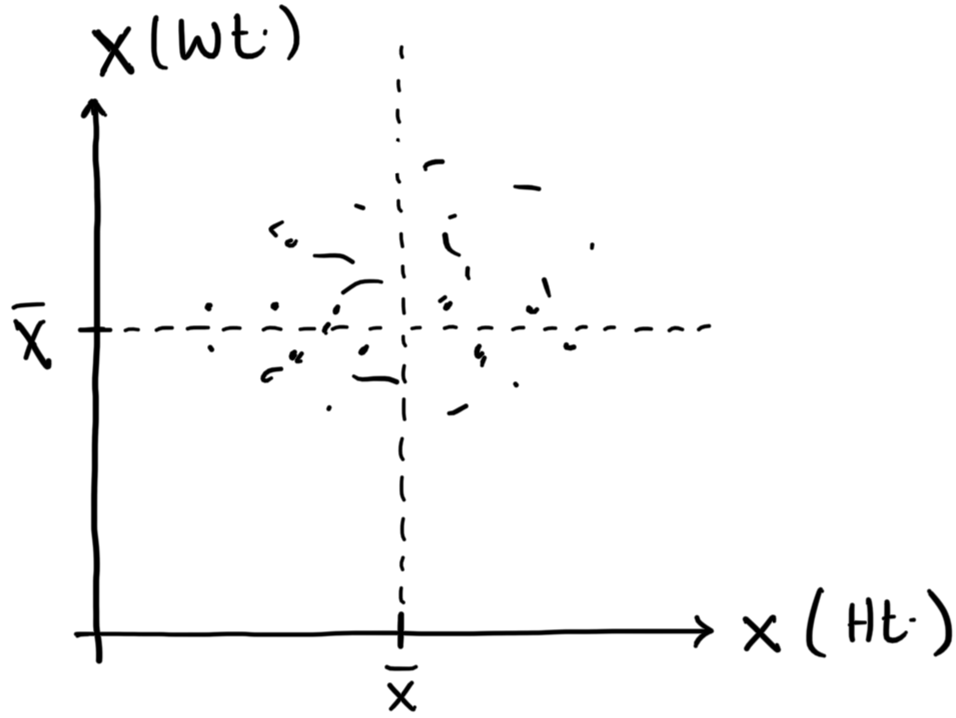
$$\text{cov}(X, Y) = \frac{1}{} \sum (x_i - \bar{X})(y_i - \bar{Y})$$

n

$cov(x,y) > 0$: x, y are above or below the mean together

$cov(x,y) < 0$: If x is above its mean, y is below (vice-versa)

Ex 1:



(a) $cov > 0$

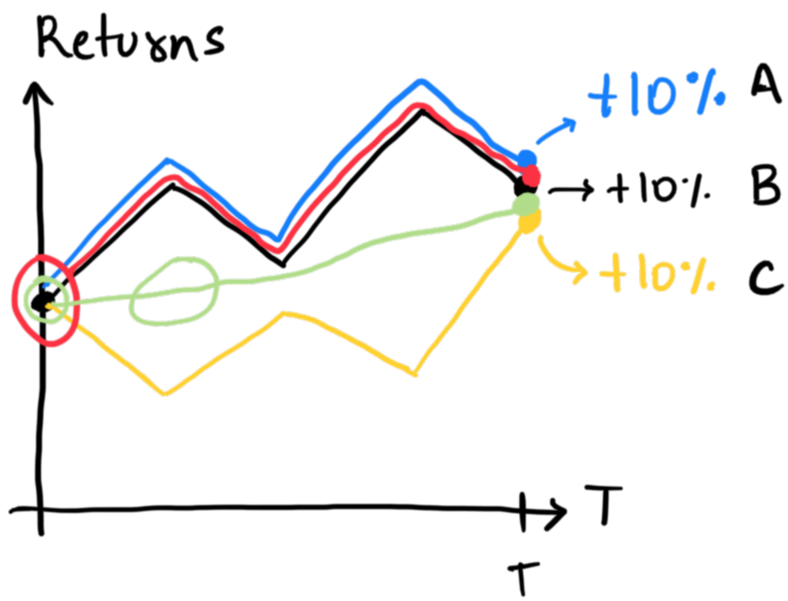
(b) $cov < 0$

(c) $cov(x,x)$

(d) $cov = 0$

(e) outlier

Ex 2:



$cov(A,B) > 0$

"Markowitz" 1952

$cov(A,C) < 0$

$cov(B,C) > 0$

Portfolio 1: 50% A, 50% B

Portfolio 2: 25% A, 25% B, 50% C

Properties of covariance

(a) $cov(x,y) = E[(x-\bar{x})(y-\bar{y})]$ } const. 0

$$= E[(X-\bar{X})\cdot Y] - \cancel{E[\bar{Y}(X-\bar{X})]}$$

$$= E[(X-\bar{X})\cdot Y]$$

(b) $$\mathrm{cov}(X,Y) = E[XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}]$$

$$= E[XY] - E[X]\cdot E[Y]$$

$$\Rightarrow E[XY] = E[X]\cdot E[Y] + \mathrm{cov}(X,Y)$$

(c) $$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

sensitivity to scale $\left\{ \mathrm{cov}(X,aY) = a\cdot \mathrm{cov}(X,Y) \right.$

$$\mathrm{cov}(X,c) = 0$$

$$\mathrm{cov}(X,Y+c) = \mathrm{cov}(X,Y)$$

$$\mathrm{cov}(X+Y,Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(Y,Z)$$

$$\mathrm{cov}(XY,Z) = E[(XY - E(XY))(Z - E[Z])]$$

$$= E\left[Z\left(XY - E[XY]\right)\right]$$

$$\neq \operatorname{cov}(X,Z) \cdot \operatorname{cov}(Y,Z)$$

(d) $$\operatorname{Var}(X+Y) = E\left[(X+Y-\bar{X}-\bar{Y})^2\right]$$

$$= E\left[(X+Y)^2 + (\bar{X}+\bar{Y})^2 - 2(X+Y)(\bar{X}+\bar{Y})\right]$$

$$= V[X] + V[Y] + 2 \cdot \operatorname{cov}(X,Y)$$

- Correlation and Causation

$$\rho_{xy} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_x \cdot \sigma_y} \quad \Big\} \text{ Scaled covariance}$$

Properties of correlation

(a) $\rho_{XY} \in [-1, 1]$

Scale invar. (b) $\text{Corr}(aX, Y) = \text{corr}(X, Y)$

(c) $\text{corr}(X+c, Y) = \text{corr}(X, Y)$

Causality is the fundamental problem of econometric inference.

$$X \sim Y$$

Education	Wage
Ice cream sales	Sunburns
Drug	Disease