January 12, 2023

DSML : CC Fundamentals

Anova & Correlations

Recap:

* Null Hypothesis : H_0
* Alternate Hypothesis : H_a
* p-value.
* Test statistic
* Significance level : α.
* z-test
* t-test (1-sample).
* t-test (2-sample)
* KS test
* χ^2 test.

TERMINOLOGY

EVERYWHERE!

Class begins @ 9:05 p.m.

Recap:

What all tests have we seen so far?

Male income. Female income. μ, σ

Numerical vs. categorical.

* Z- test
* T- test

Numerical vs. categorical.

* KS -test
* χ^2 - test

categorical vs. categorical.

* ANOVA — Numerical vs. $\underbrace{\text{categorical}}_{\text{multiple categories!}}$.

* Pearson R correlation test
  ↳ numerical vs. numerical.

Recap: Degrees of freedom

Case 1: If we know the sample mean of n numbers, d.o.f $= n-1$

Case 2: If we know the sample means of 2 sets of numbers, d.o.f $= ?$

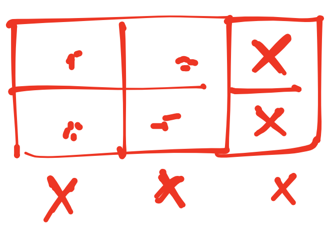
$n_1 + n_2 - 2$.

1 n_1 $\quad \bar{x}_1 \to n_1 - 1$

2 n_2 $\quad \bar{x}_2 \to n_2 - 1$

Case 3: In a contingency table with $\underline{m}$ rows and $\underline{n}$ columns,

d.o.f $= ?$



$(m-1) \times (n-1)$

ANOVA - Analysis of Variance, Intuition

Setup 1 :

(Heights) Indonesian College students → $\mu = 60$ in

(Heights) Indian College students → $\mu = 66$ in.

(Heights) American Basketball players → $\mu = 72$ in

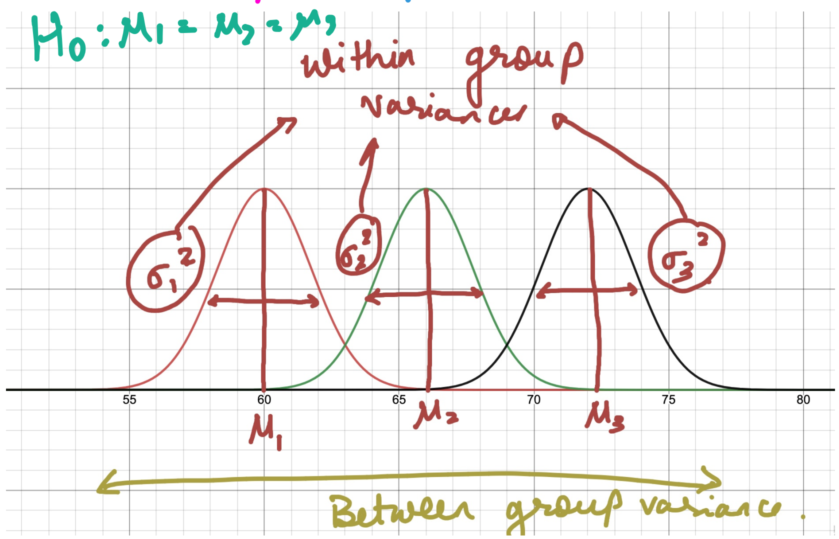
Setup 2 :

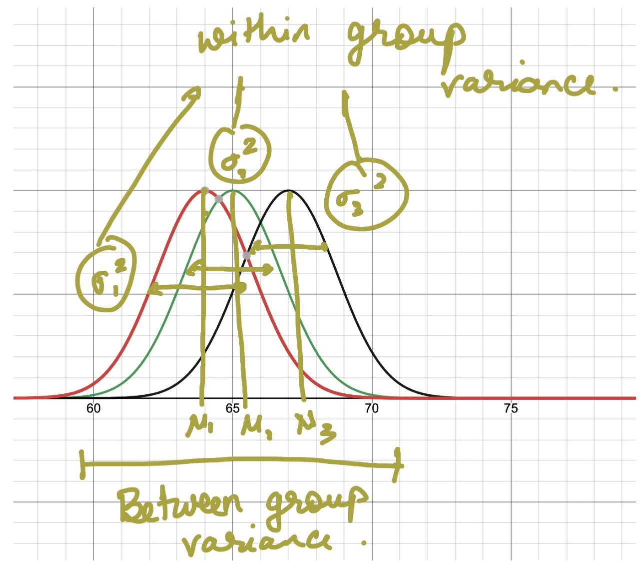
(Heights) College 1 (India) → 64 in

(Heights) College 2 (India) → 65 in

(Heights) College 3 (India). → 67 in

Setup 1. > Setup 2





F- distribution.

F-ratio : $\dfrac{\text{Variance between groups}}{\text{Variance within groups}}$.

If F-ratio is ↑ → reject

If F-ratio is ↓ → fail to reject.

→ Test statistic for ANOVA!

→ ppf.
↓
to get critical value.

α

ANOVA: Performing the test.

iPhone sale data.

	A	B	C
0	25	30	18
1	25	30	30
2	27	21	29
3	30	24	29
4	23	26	24
5	~~20~~	~~28~~	~~26~~
	25	26.5	26

$\boxed{25.83}$

$6 \times 3 = 18.$

15

$H_0 : \mu_1 = \mu_2 = \mu_3.$

Step 1 : Compute individual means.

Step 2 : Compute mean of these 3.

Step 3 : Calculate between group variances.

$$SSB = 6 \times (25 - 25.83)^2 + 6 \times (26.5 - 25.83)^2 + 6 \times (26 - 25.83)^2$$

$DF_1 = 2.$

$$MSB = \frac{SSB}{DF} = \frac{6.9}{2} = 3.49.$$

Step 4 : Within groups.

$$SSW = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = 223.5.$$

$DF = 15.$

$$MSW = \frac{SSW}{DF_2} = 14.9.$$

For ANOVA,
Dependent → continuous
Independent → categorical.

$$\text{F-ratio} = \frac{MSB}{MSW} = \frac{3.49}{14.9} \approx 0.23.$$

ANOVA : Assumptions.

1] Normality - each sample is from a normally distributed pop.

2] Independence : sample

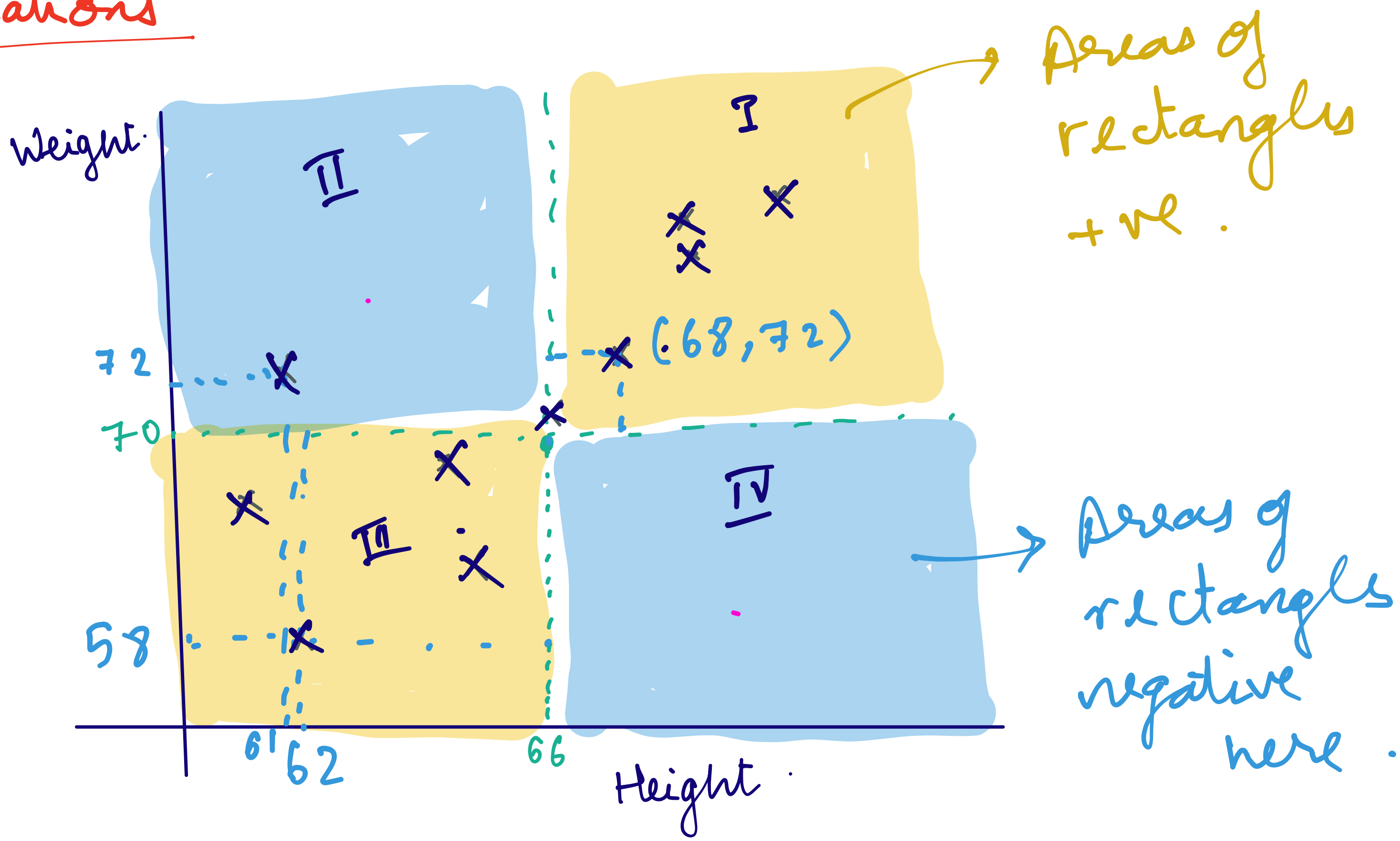
3] Equal variance of data in ~~Nearly~~ different groups.

In case assumptions fail,

→ Use Kruskal - Wallis test!!

Correlations

Height (inches)	Weight (kg)
68	72
62	58
64	67
61	72
70	79
66	61
61	68
65	64
71	80
72	79
$\bar{h} = 66$	$\bar{w} = 70$



Covariance: $(68-66)\times(72-70) = 2\times 2 = 4.$

$(62-66)\times(58-70) = (-4)\times(-12) = 48.$

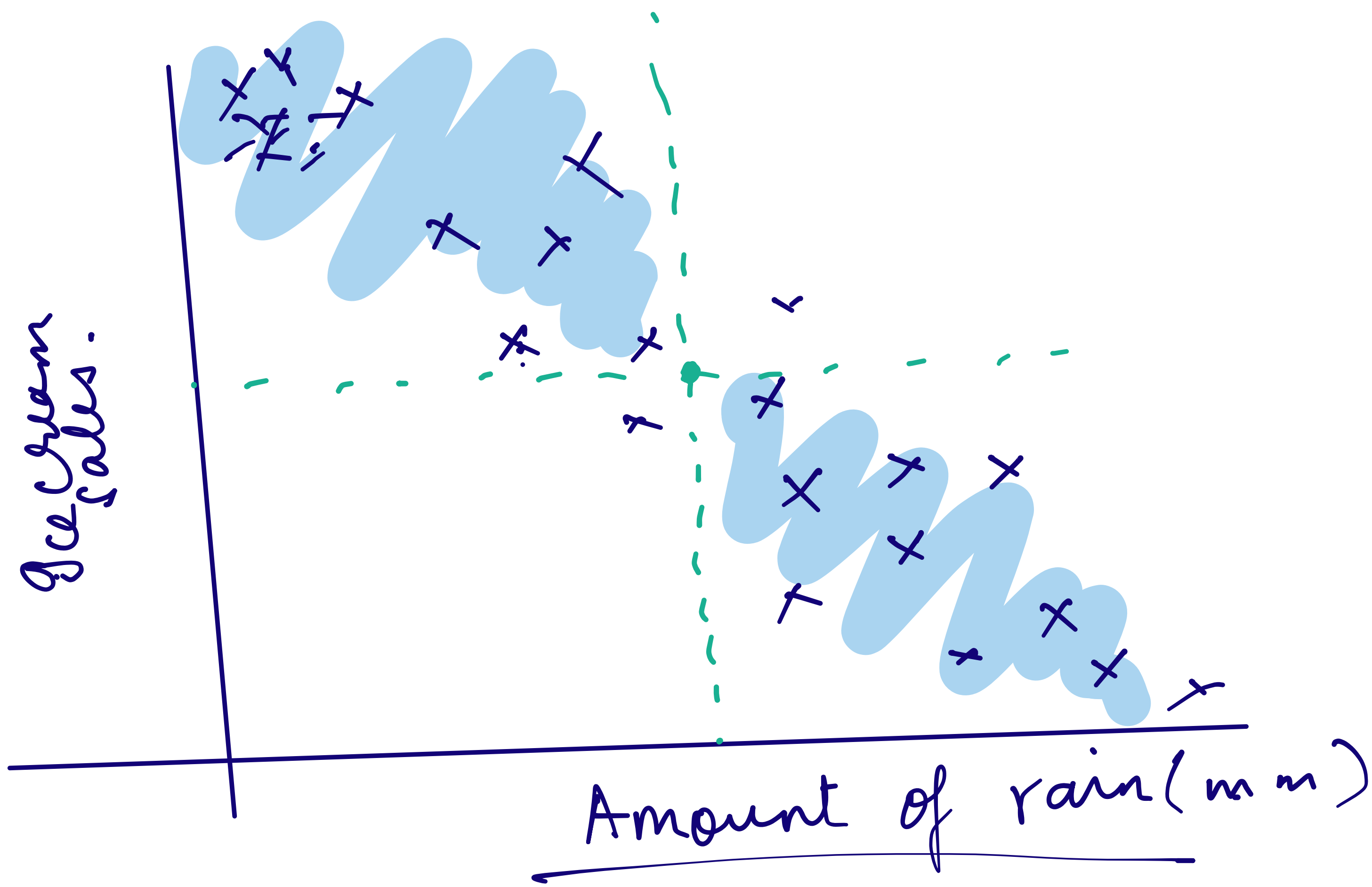
$(61-66)\times(72-70) = -5 \times 2 = -10$

$(\quad)\times(\quad)$

$\vdots$

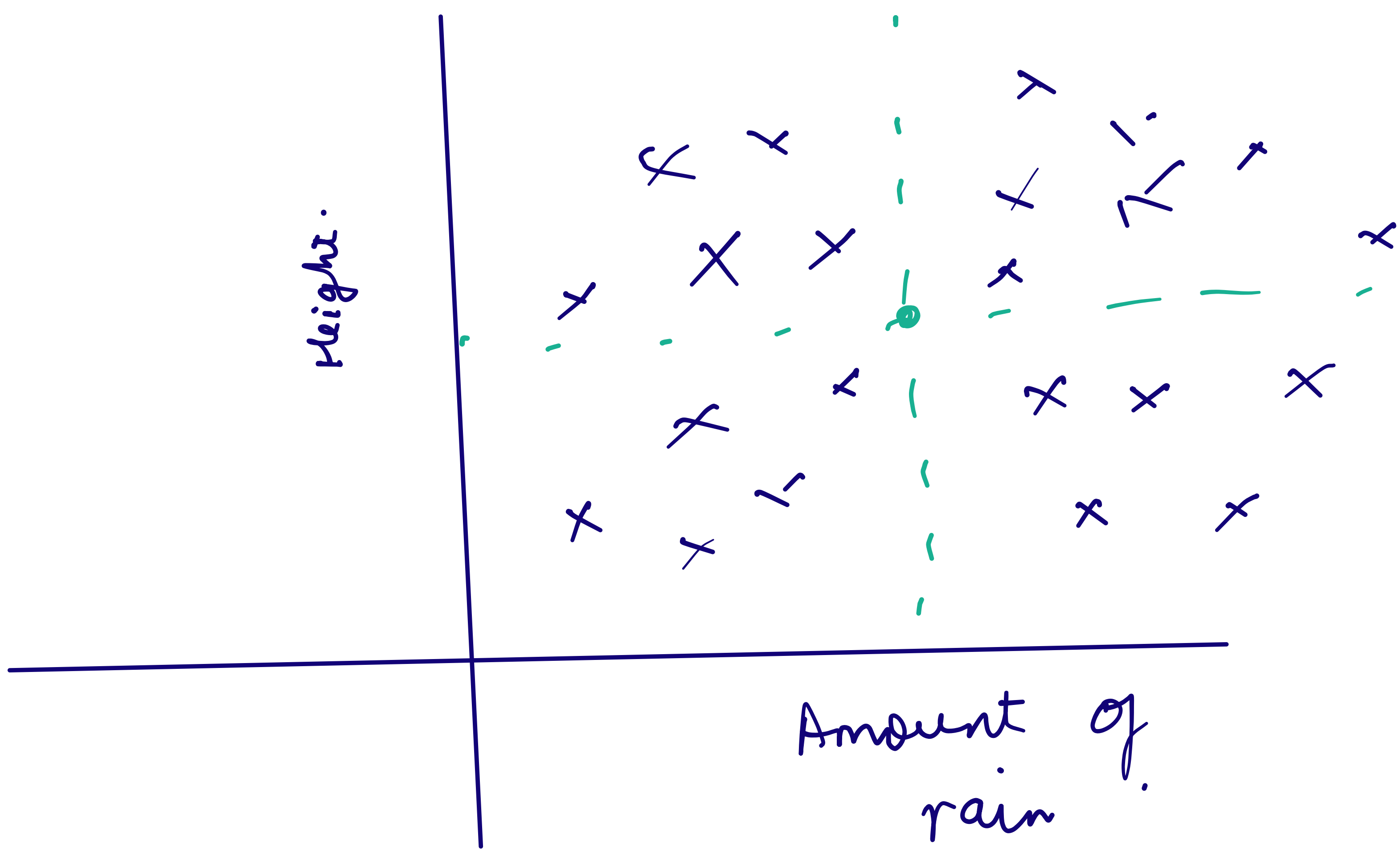
$$\mathrm{cov}(h,w) = \frac{1}{n}\sum_{i=1}^{n}(h_i-\bar{h})(w_i-\bar{w})$$

Ice-cream vs. Rain



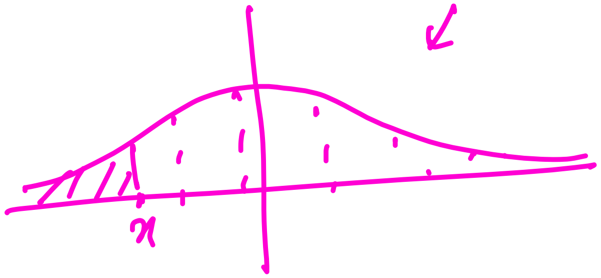
cov (rain, ice-cream) → -ve.

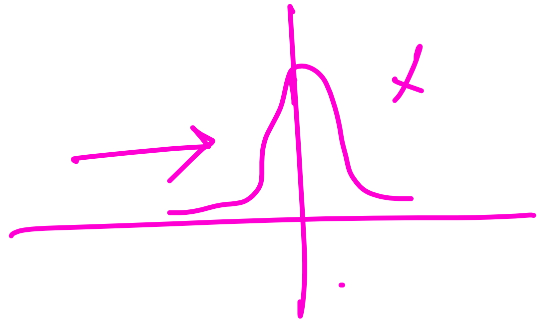
Height vs. Rain



Cov → +ve, -ve, very close to 0?

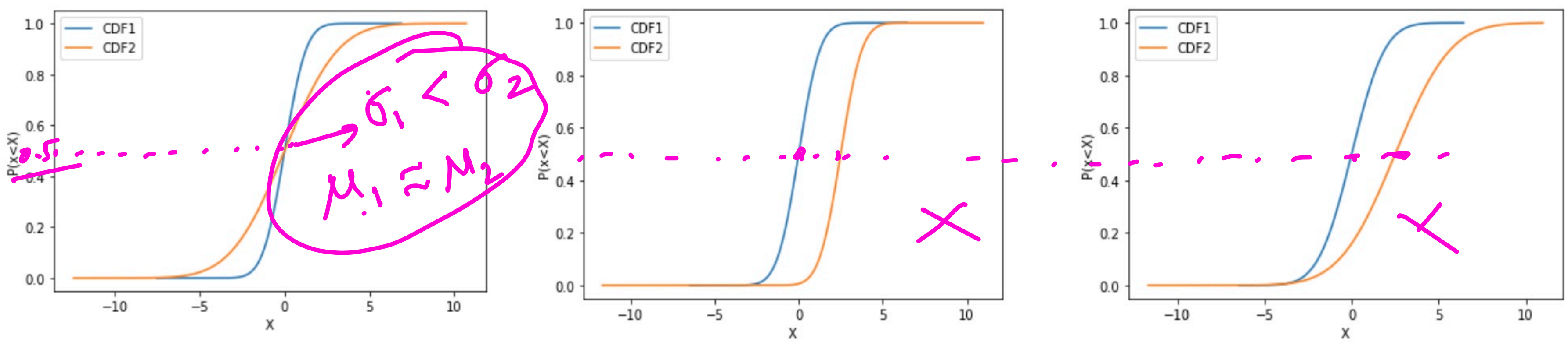
	Male	Female
Q_1		
Q_2		
Q_3		
Q_4.		





Compare Distributions

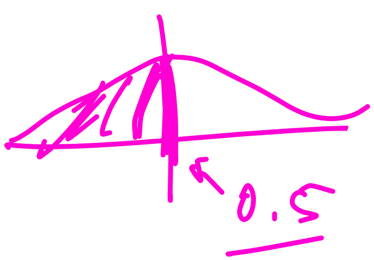
Following each figure contains **CDFs** plotted for two different **normal distributions**.

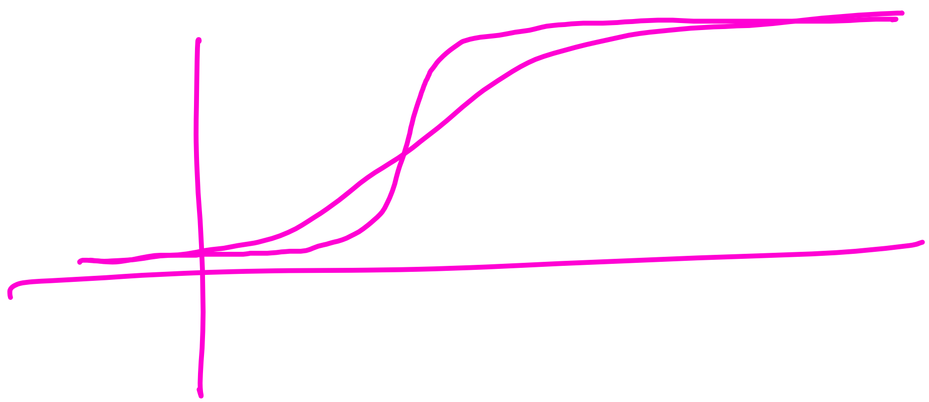


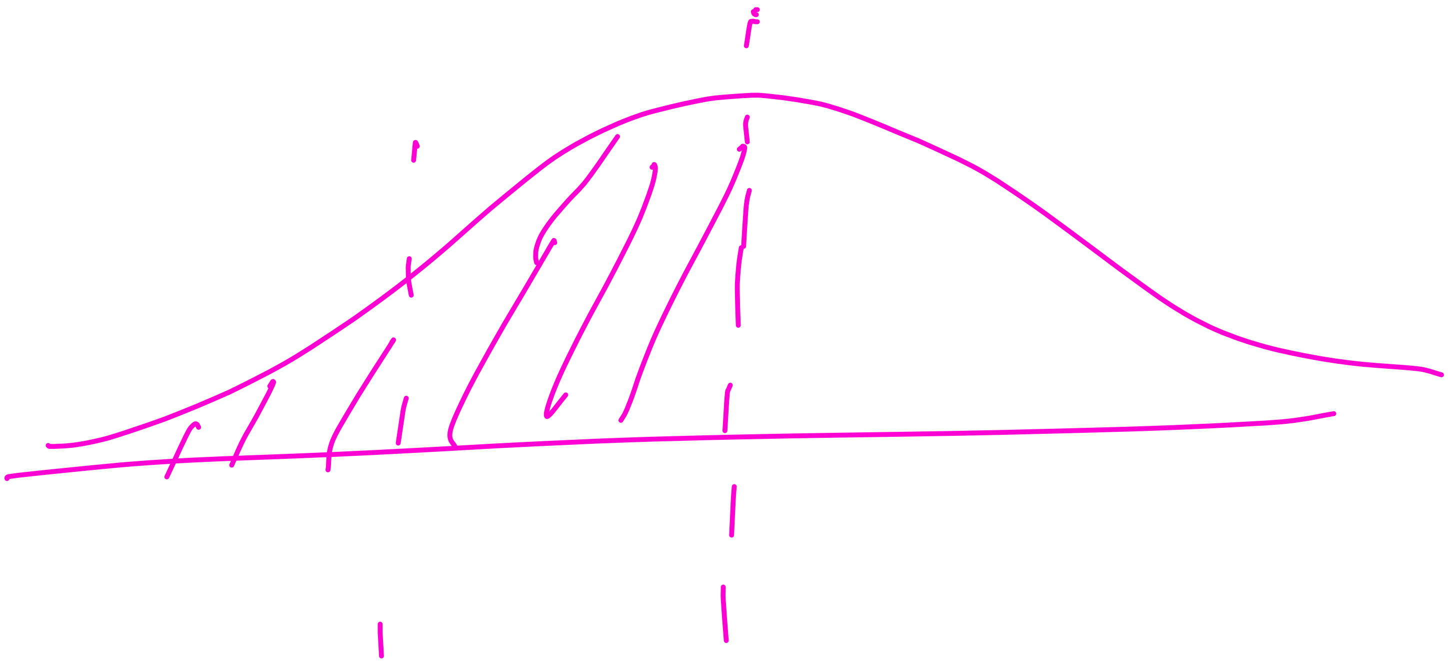
What is the **relation** that each case describes?

Note: Option follows following format:

Comparision for Figure 1; Figure 2; Figure 3







cdf (mean) = 0.5?