☆ ANOVA - Analysis Of Variance

(1) Cat Vs. Numerical
(ex - Gender Vs. Income)
→ n > 30 — z-test
→ T-test

(2) cat. Vs. cat
(ex - Gender Vs. Product) → chi^2

(3) cat Vs. Cat Vs cat → ANOVA
(ex - Gender Vs. Product Vs. City)

When we have more than 2 columns, we use ANOVA

✡ ANOVA (Deep dive)

→ Basketball players

USA - 6.6 ± 2.5

India - 6.2 ± 2.5

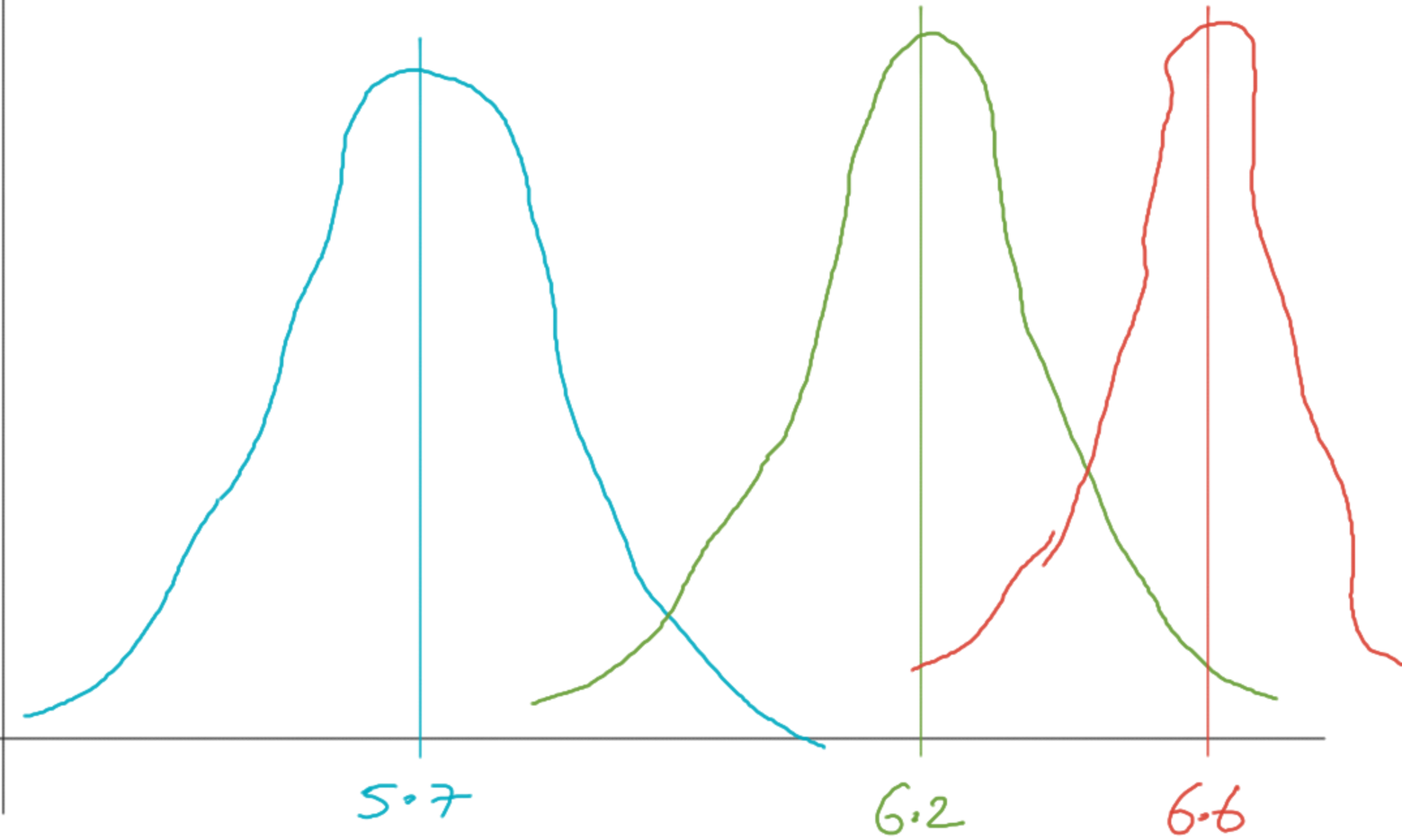
Indonesia - 5.7 ± 2.5

Question: Are mean height of Basket Ball players associated with groups (of country)?

H_0: Height in groups are not associated.

H_a: They are associated.

Case-1



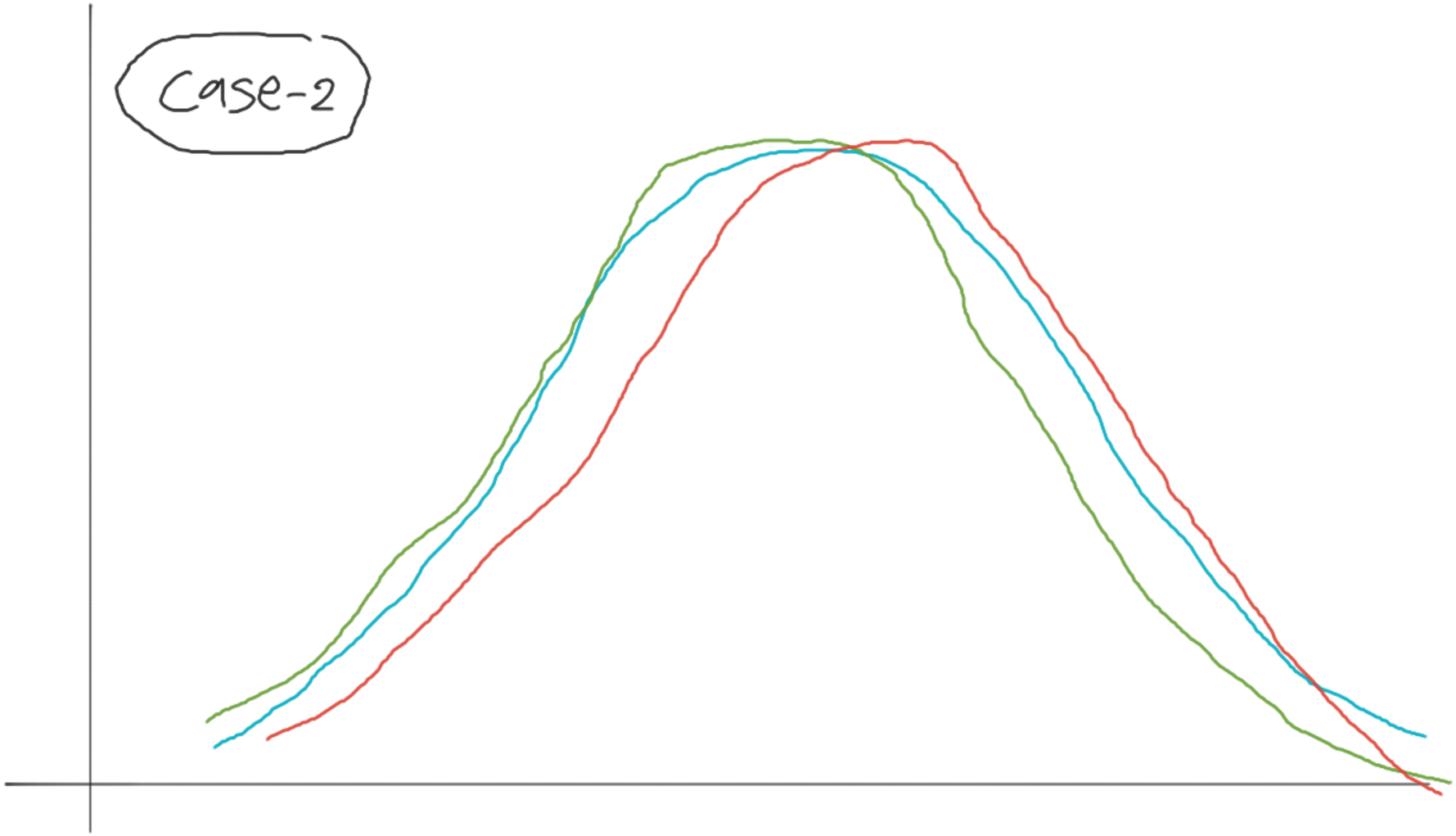
→ If we mix-up all of them and then randomly make 3 groups - g_1, g_2 & g_3

$mean(g_1)$

$\approx mean(g_2)$

$\approx mean(g_3)$

Standard deviation ↑



$$F = \frac{\text{Variance between the groups}}{\text{variance within the groups}}$$

Case-1: variance between the groups is high & variance within the groups is low. ∴ F-ratio is high ∴ p-value is lower ∴ Reject H_0

Case-2: is reverse. ∴ F-ratio is lower hence p-value is high ∴ Fail to reject H_0

✡ Assumptions of ANOVA

(1) Data must be Gaussian
- ↳ Plot the histogram
- ↳ Test-1 : Q-Q Plot
- ↳ Test-2 : Shapiro

(2) Rows must be independent of each other

(3) Variance within the groups must be equal.

If these assymptions are not followed then we use Kruskal Test

Logic behind Q-Q plot :