巨量資料管理學院碩士在職專班

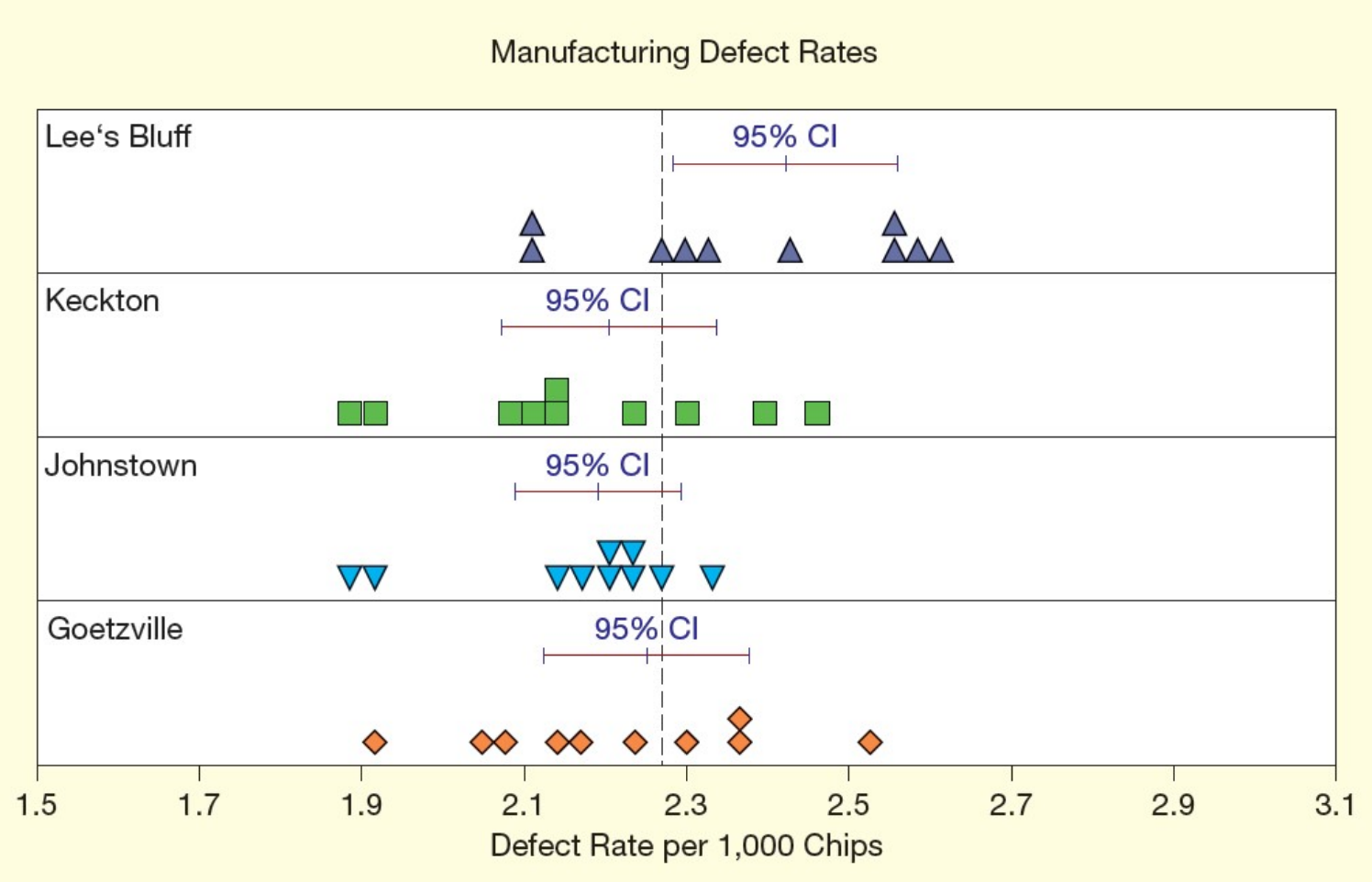
統計分析

2022/11/18

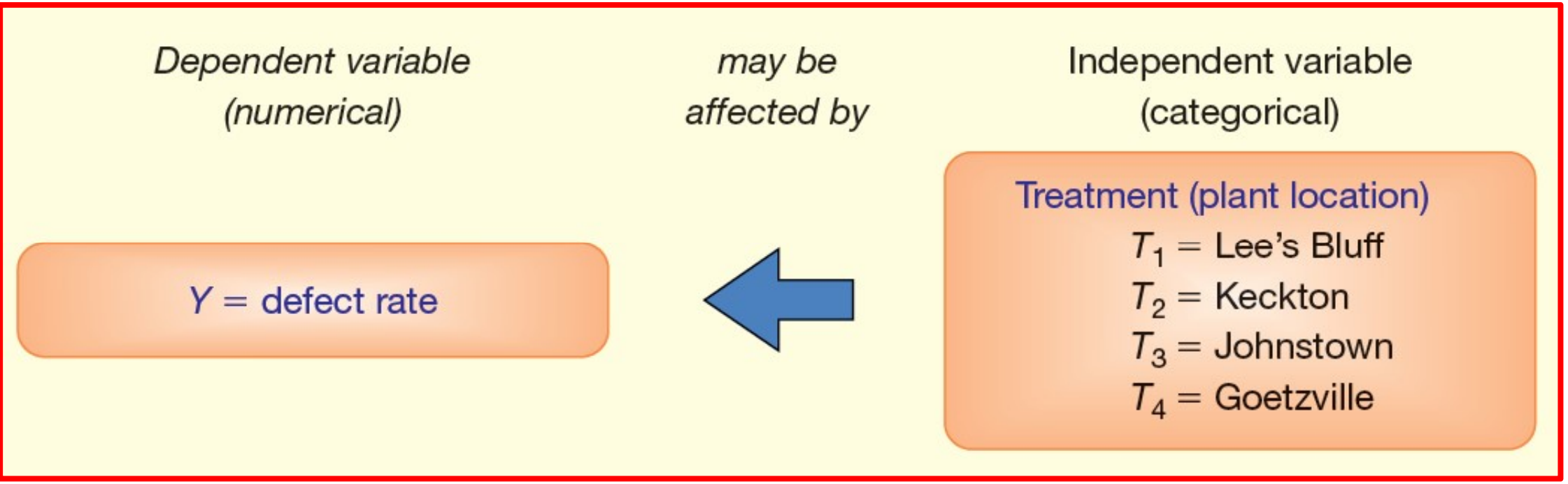
陳光宏

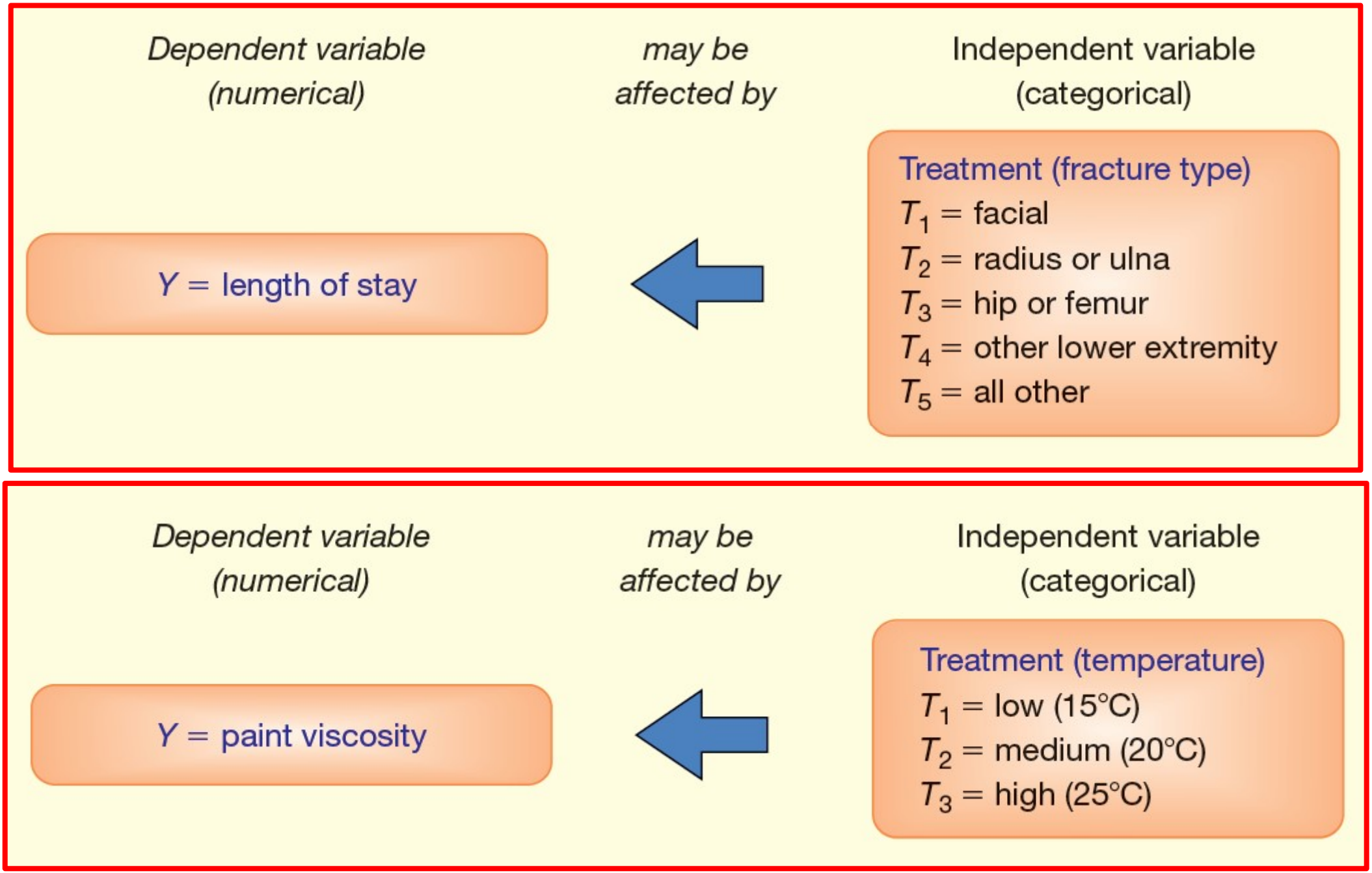
變異數分析
(Analysis of variance)

多個獨立樣本的檢定
Analysis of variance (ANOVA)



涉及兩個變項


Dependent variable
(numerical)
may be
affected by
Independent variable
(categorical)
Y = defect rate
Treatment (plant location)
T1 = Lee's Bluff
T2 = Keckton
T3 = Johnstown
T4 = Goetzville


Dependent variable (numerical)
may be affected by
Independent variable (categorical)
Y = length of stay
Treatment (fracture type)
T1 = facial
T2 = radius or ulna
T3 = hip or femur
T4 = other lower extremity
T5 = all other
Dependent variable (numerical)
may be affected by
Independent variable (categorical)
Y = paint viscosity
Treatment (temperature)
T1 = low (15°C)
T2 = medium (20°C)
T3 = high (25°C)

多個獨立樣本的檢定
Analysis of variance (ANOVA)

- 比較三組或三組以上的平均值
- 能否沿用兩個獨立樣本檢定的概念？
- 要解決什麼問題?
  - 假說怎麼建立？
  - 檢定統計量怎麼計算？

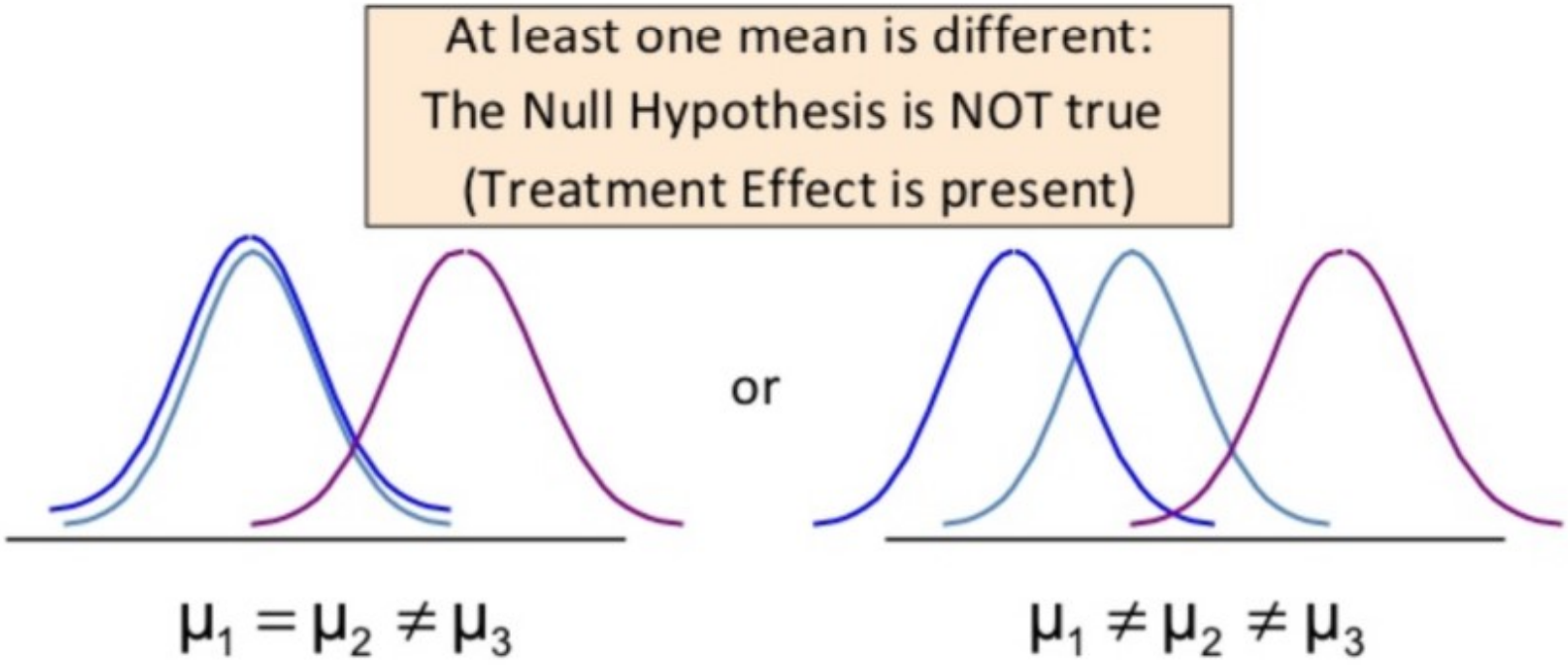
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

假說建立

- 虛無假說：三個平均值均相等
- 對立假說：至少有兩個平均值不相等

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

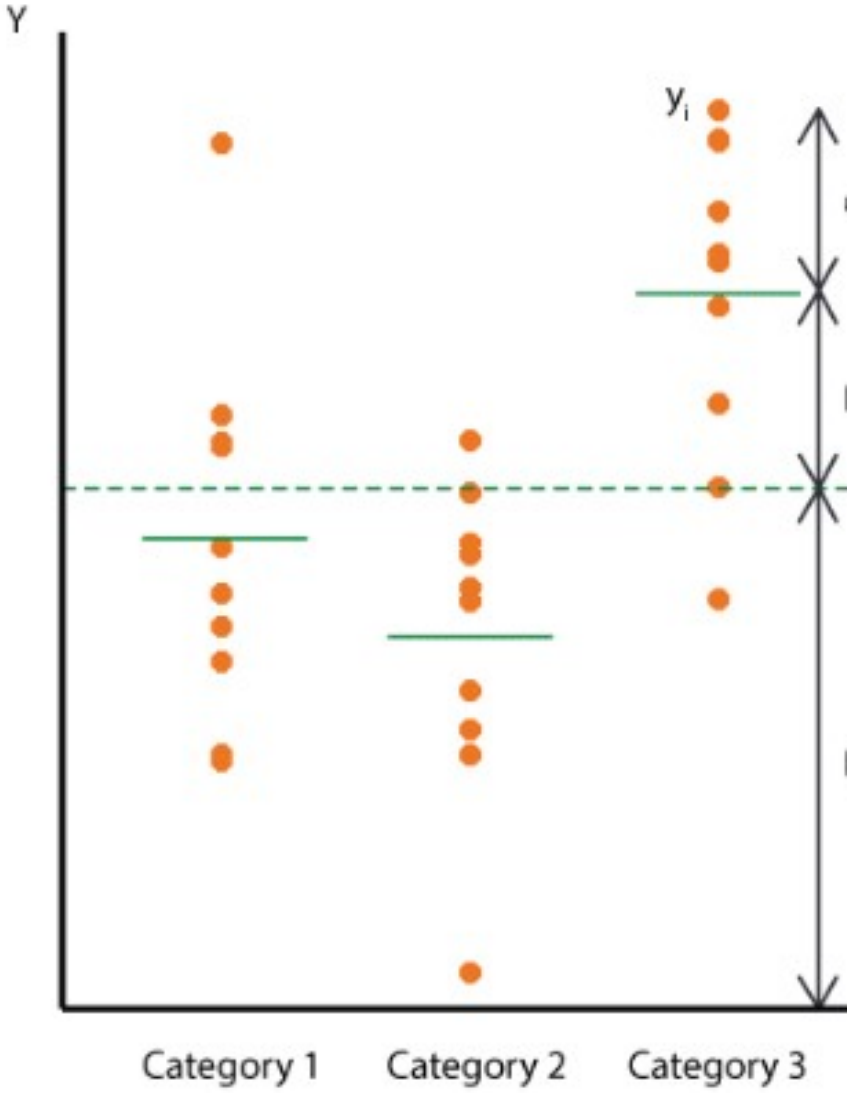
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$



概念

- 變異數分析
  - 組間變異 (Between-group variance)
  - 組內變異 (Within-group variance)
- 組間變異是否大於組內變異

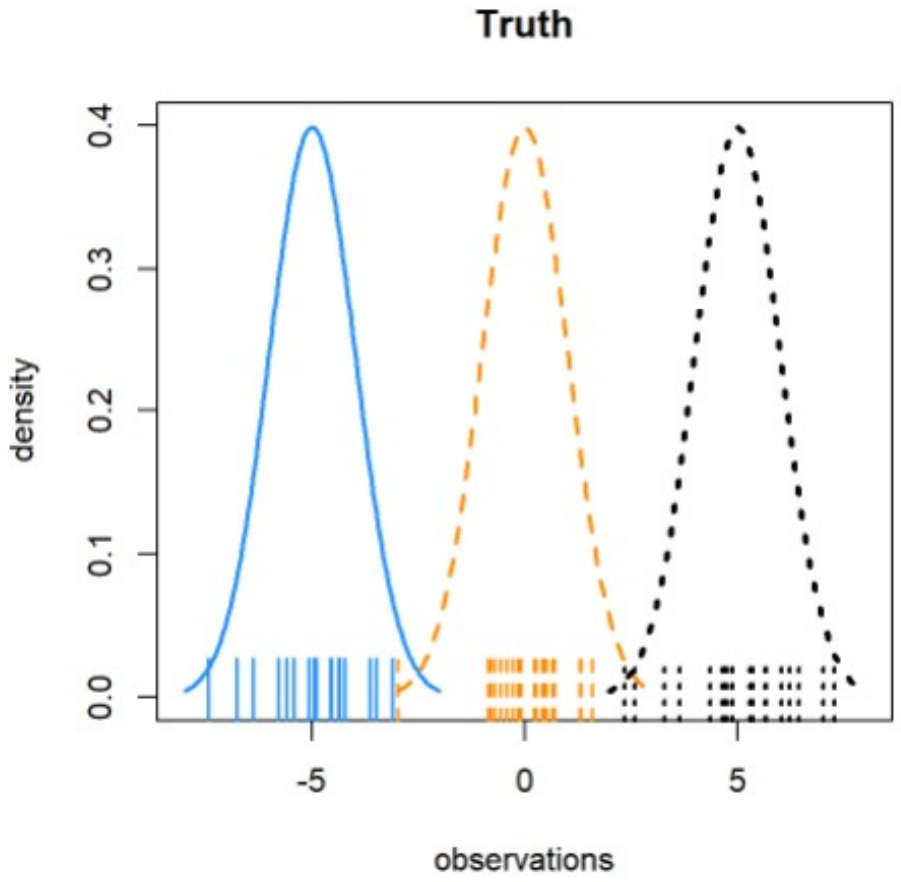
- 補充：變異 = 離平均的距離

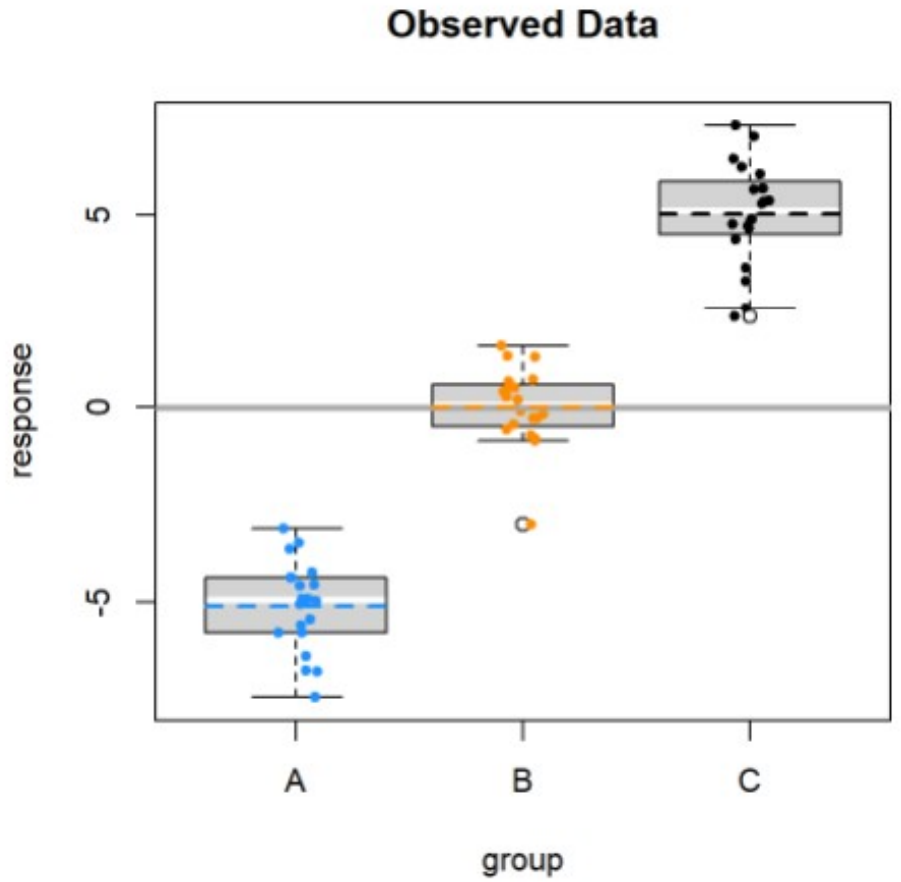


假想的例子

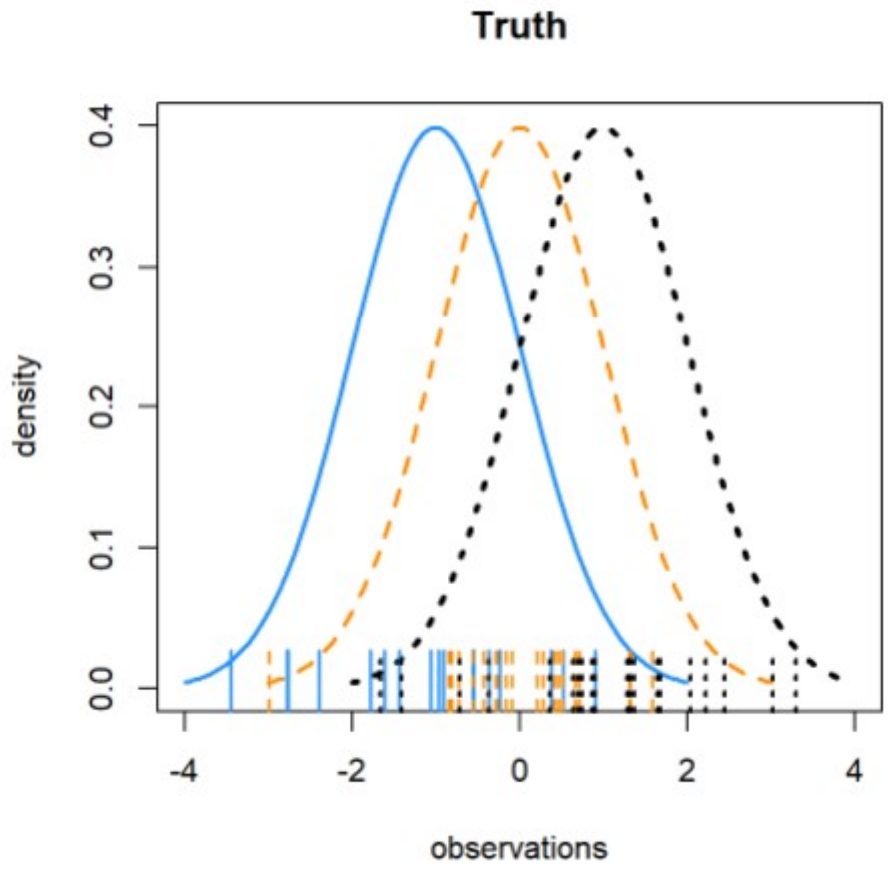
	Data 1			Data 2			Data 3		
	Group 1	Group 2	Group3	Group 1	Group 2	Group3	Group 1	Group 2	Grou
	2	300	20	1	1	1	2.1	5.1	6.1
	3	350	30	2	2	2	2.2	5.2	6.2
	4	400	40	2	5	6	2.2	5.5	6.6
	5	450	50	2	8	10	2.2	5.8	7.0
	6	500	60	3	9	11	2.3	5.9	7.1
均	4	400	40	2	5	6	2.2	5.5	6.6
異數	2	5000	200	0.4	10	16.4	0.004	0.1	0.16

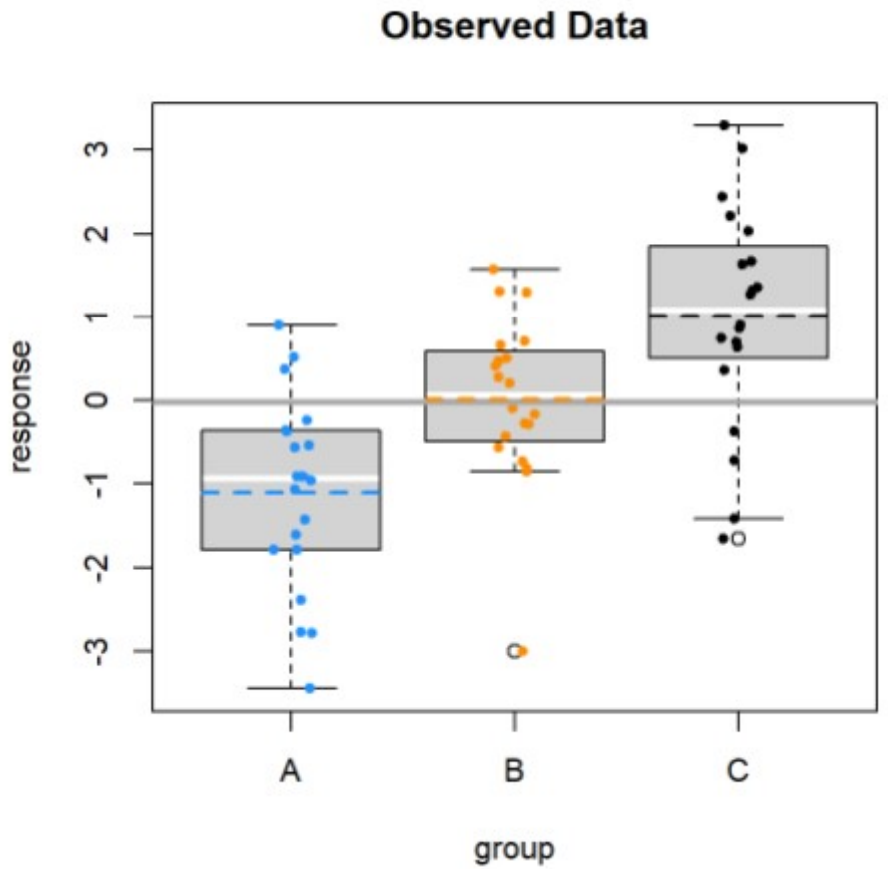
概念 - 組間 vs 組內變異


Truth
density
0.0
0.1
0.2
0.3
0.4
-5
0
5
observations


Observed Data
response
-5
0
5
A
B
C
group

概念 - 組間 vs 組內變異

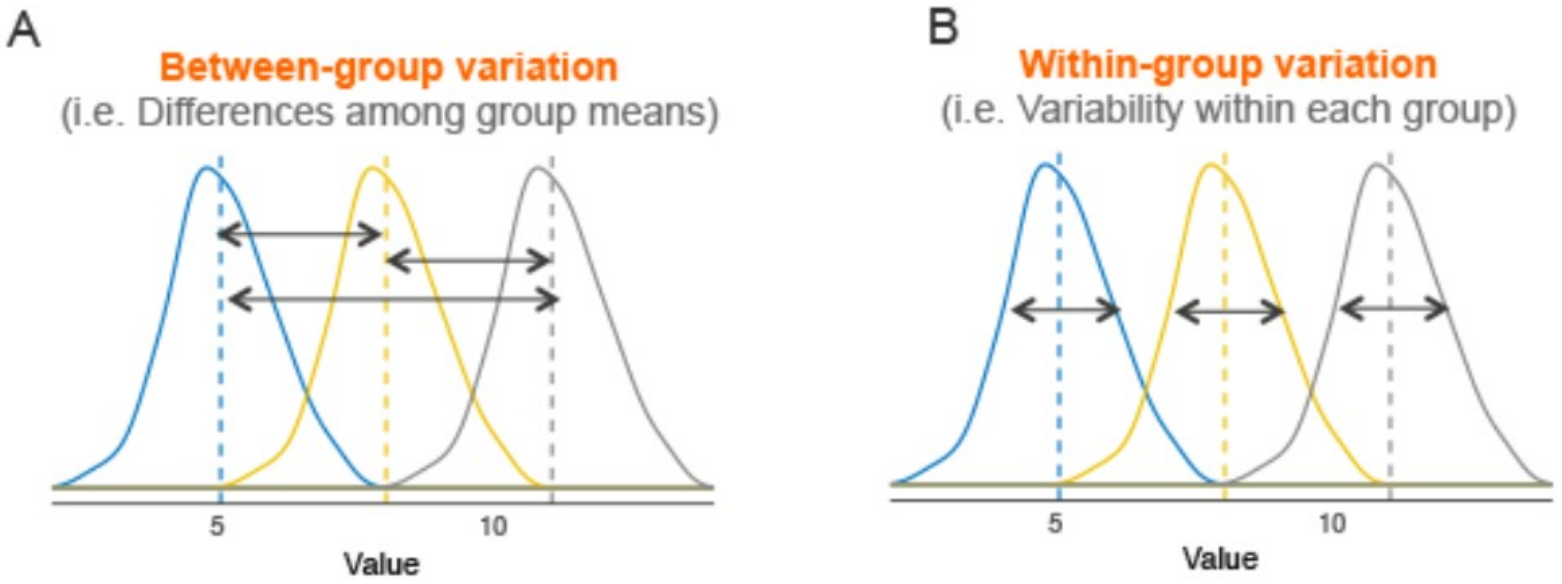

Truth
density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
observations


Observed Data
response
-3
-2
-1
0
1
2
3
A
B
C
group

拆解變異數

Variation in Y	=	Explained Variation	+	Unexplained Variation
(around its mean)		(due to factors)		(random error)

變異數 (Variance) = SD^2 = $\frac{\text{平方和 } Sum\ of\ squares\ (SS)}{\text{自由度 } Degree\ of\ freedom\ (df)}$



拆解變異數

Partitioned Sum of Squares				
Sum of Squares Total (SST)	=	Sum of Squares between Treatments (SSB) ↑ Explained by Treatments	+	Sum of Squares within Treatments (SSE) ↑ Unexplained Random Error

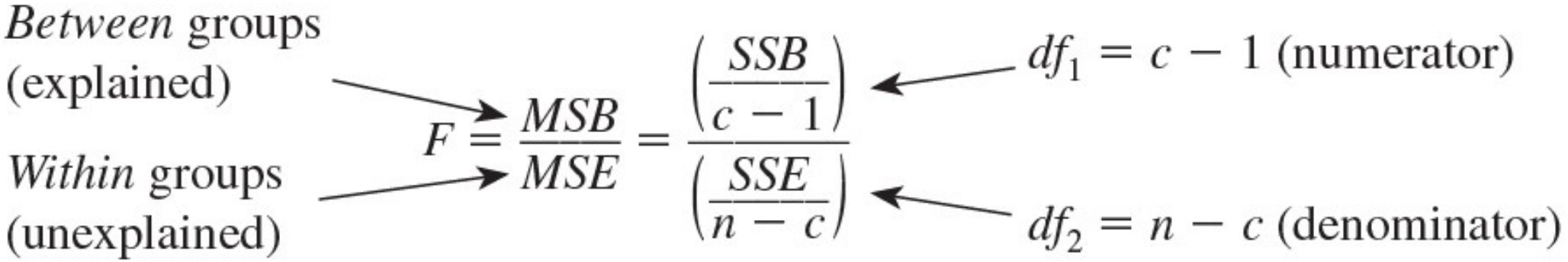
組間變異　　組內變異

如何拆解

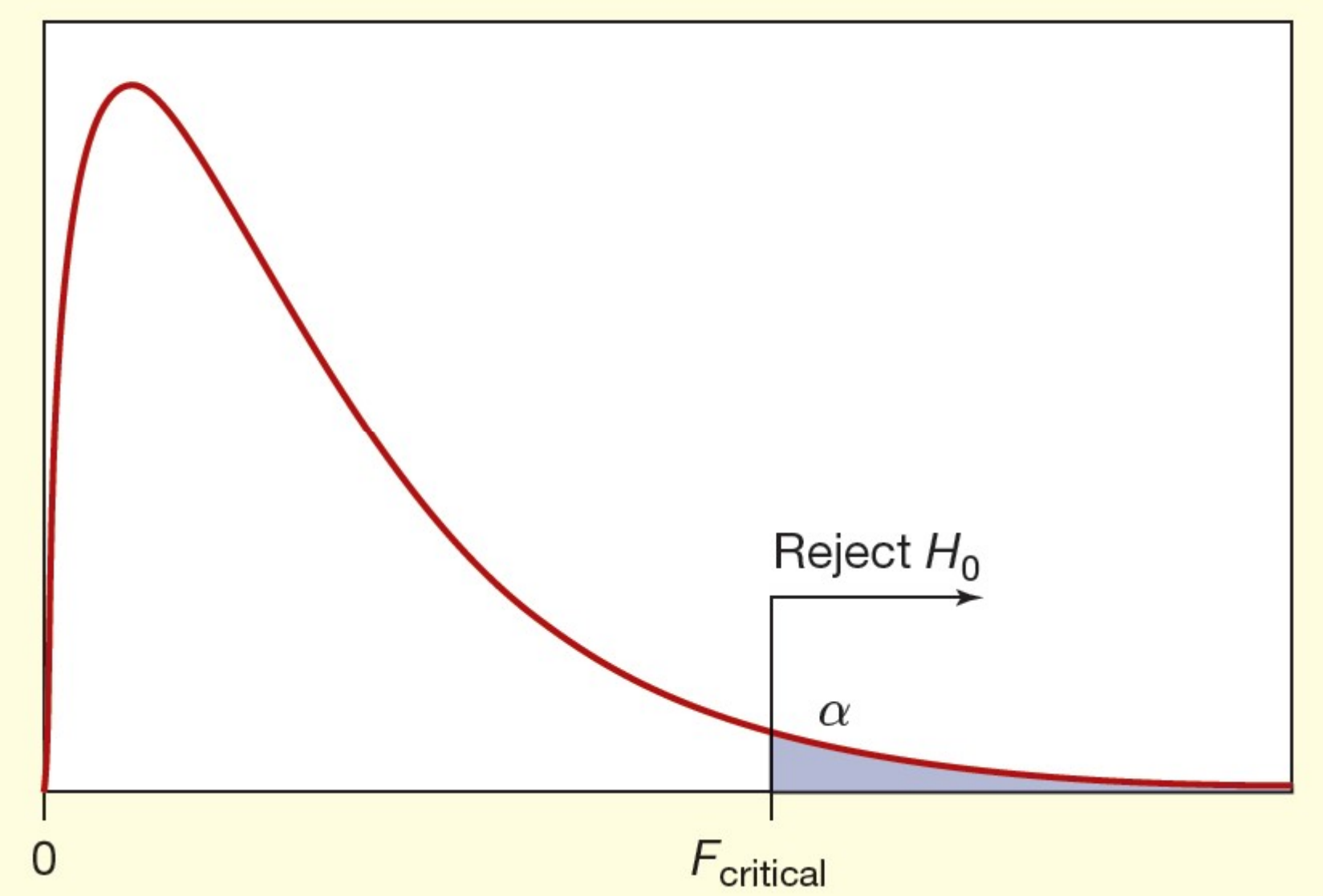
- ANOVA table

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Statistic*
Treatment (between groups)	$SSB = \sum_{j=1}^{c} n_j(\bar{y}_j - \bar{y})^2$	$c - 1$	$MSB = \frac{SSB}{c-1}$	$F = \frac{MSB}{MSE}$
Error (within groups)	$SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$	$n - c$	$MSE = \frac{SSE}{n-c}$	
Total	$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$	$n - 1$		

F檢定統計量 (F分布)


Between groups
(explained)
Within groups
(unexplained)
$F = \frac{MSB}{MSE} = \frac{\left(\frac{SSB}{c-1}\right)}{\left(\frac{SSE}{n-c}\right)}$
$df_1 = c - 1$ (numerator)
$df_2 = n - c$ (denominator)

F分布



範例

- 某公司有四個工作站，負責將商品裝箱，運送給零售商，每個工作站每天可以完成200箱以上。
- 請參考data_w11.xlsx檔案
- 請比較四個工作站平均裝箱數是否有差異？

範例解答

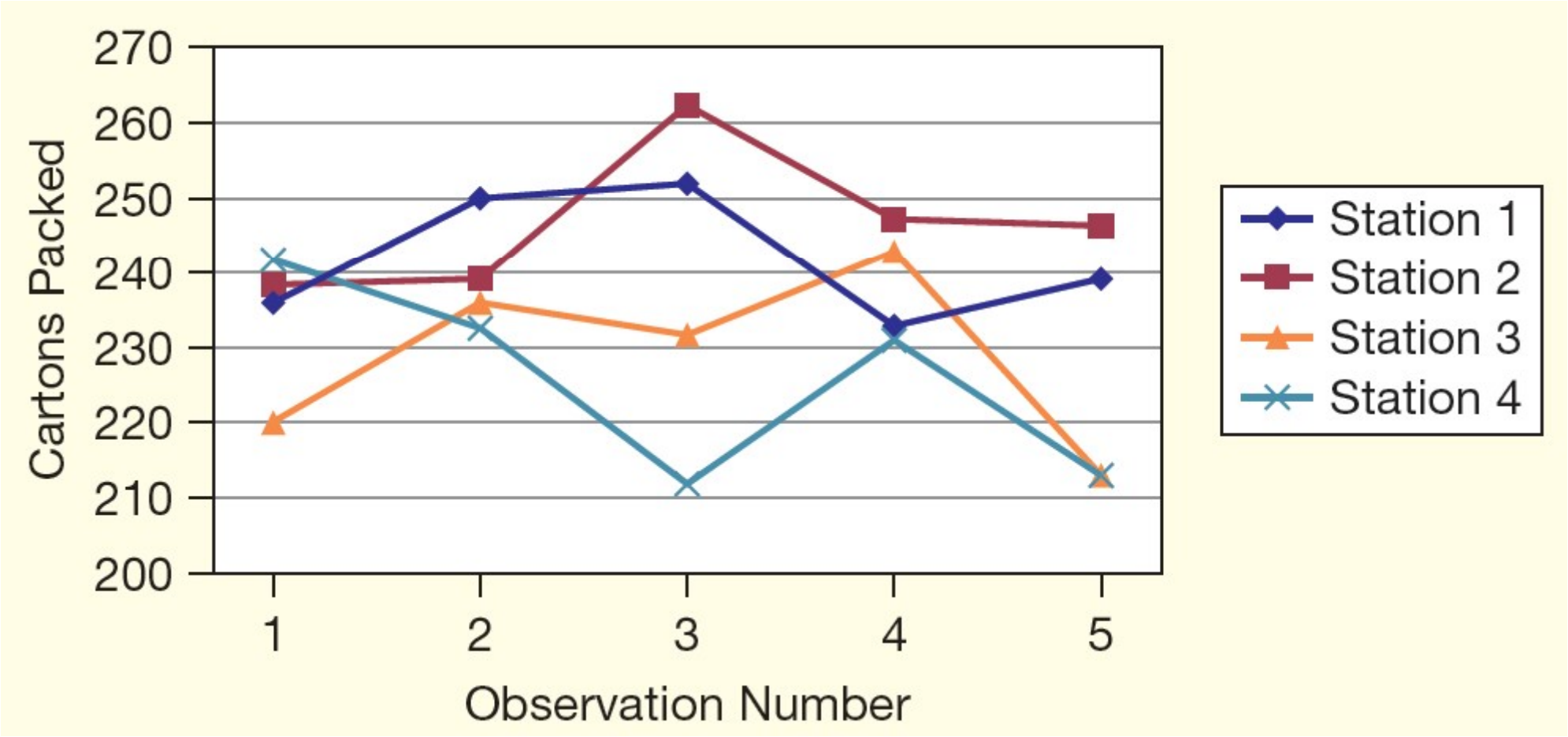
- 假說建立

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (the means are the same)

H_1: Not all the means are equal (at least one mean is different)

- 設定決策規則
  - P值法
  - 型I誤差 = 0.05

資料描述


270
260
250
240
230
220
210
200
Cartons Packed
1
2
3
4
5
Observation Number
Station 1
Station 2
Station 3
Station 4

nova: Single Factor

UMMARY

roups	*Count*	*Sum*	*Average*	*Variance*
tation 1	5	1210	242.0	72.5
tation 2	5	1232	246.4	92.3
tation 3	5	1144	228.8	147.7
tation 4	5	1130	226.0	166.0

NOVA

ource of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F cr*
etween Groups	1479.2	3	493.0667	4.121769	0.024124	3.23887
Vithin Groups	1914.0	16	119.6250			
otal	3393.2	19				

結論

- P值小於0.05，達到統計顯著，拒絕虛無假說。
- 至少有兩個工作站的平均裝箱數量有差異。

哪幾組有差異？

多重比較 (事後檢定)

$$H_0: \mu_j = \mu_k$$
$$H_1: \mu_j \neq \mu_k$$

- Bonferroni correction
  - $\alpha^* = \frac{0.05}{6}$

- Tukey事後檢定

$$T_{\text{calc}} = \frac{|\bar{y}_j - \bar{y}_k|}{\sqrt{MSE\left[\frac{1}{n_j} + \frac{1}{n_k}\right]}}$$

變異數分析的統計假設

- 獨立樣本 (Independence)
- 常態分布 (Normal distribution)
- 變異數相等 (Equal variance; homoscedasticity)