Analysis of Variance (ANOVA).

* t-statistic test for 2 means can not be generalized for more than 2 different groups or populations.

∴ ANOVA !

Data organization & representation.

Data :- y_{ij} , $i = 1, 2, \ldots, t \longrightarrow$ t different groups (populations)

$j = 1, 2, \ldots, n_i$ (for i^{th} gp.).

for each of t groups, we have $n_1, n_2, \ldots, n_t$ sets of obs.

$$\text{Total} = \sum_{i=1}^{t} n_i \text{ observations.}$$

Null hypothesis :-

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

H_1 : atleast 1 of the above equality is not satisfied.

Assumption :- Each group is distributed $N(\mu_i, \sigma^2)$

* σ^2 is same across population groups.

Organization of Data :-

Factor levels	Observations	$\sum_j y_{ij}$ Totals	$\frac{Y_{i.}}{n_i}$ Means	Sum of sqs.
1	y_{11} y_{12} - - - - - - - - y_{1n_1}	$Y_{1.}$	$\bar{y}_{1.}$	SS_1
2	y_{21} y_{22} - - - - - - - y_{2n_2}	$Y_{2.}$	$\bar{y}_{2.}$	SS_2
3	y_{31} y_{32} - - - - - - - y_{3n_3}	$Y_{3.}$	$\bar{y}_{3.}$	SS_3
⋮	⋮	⋮	⋮	⋮
t	y_{t1} y_{t2} . . . . y_{tn_t}	$Y_{t.}$	$\bar{y}_{t.}$	SS_t
	Overall :-	$Y_{..}$	$\bar{y}_{..}$	SS_p

Sum of sqs.

$$SS_i = \sum_j (y_{ij} - \bar{y}_{i.})^2 \quad ; \quad i = 1, 2, \ldots t$$

$$\equiv \sum_j y_{ij}^2 - \frac{(Y_{i.})^2}{n_i}$$

$$\text{Pooled sum of sqs.} = SS_p = \sum_{i=1}^{t} SS_i$$

$$\text{Pooled d.o.f.} = \sum_{i=1}^{t} n_i - t \xrightarrow{\text{if } n_1 = n_2 = \cdots = n_t = n} t(n-1)$$

$$s_p^2 = \frac{SS_p}{\sum_{i=1}^{t} n_i - t}$$

(pooled variance)

* If the individual variances are available s_i^2

then $$s_p^2 = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i n_i - t}$$

the variance estimate of the factor level means

$$s^2_{means} = \frac{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2}{(t-1)}$$

* Under the null hypothesis & the 1st th^m on Sampling D^n (ref. to notes on Sampling D^n); the factor level means have D^n w/ mean μ & variance σ^2/n (provided the variance at each factor level is the same σ^2).

$$s^2_{means} = \frac{\sigma^2}{n} \Rightarrow n s^2_{means} = \sigma^2$$

w/ $(t-1)$ d.o.f.

* An alternate estimate of σ^2 is s_p^2 w/ $t(n-1)$ d.o.f.

From the def^n of F-D^n; F_{value} represents the ratio of 2 independent estimates of a common variance

$$\therefore F_{cal} = \frac{n s^2_{means}}{s_p^2} > F_\alpha(t-1, (n-1)t) \Rightarrow \text{then reject } H_0!$$

Alternate set of calculations of ANOVA (leads to same inference).

SSB (sum of sqs. bet'n gps) $= \sum_i \frac{(Y_{i\cdot})^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{\sum_i n_i}$ w/ d.o.f$_B$ = dfB = (t − 1)

SSW (sum of sqs. w/in gps.) $= \sum_{i,j} y_{ij}^2 - \sum_i \frac{Y_{i\cdot}^2}{n_i}$ w/ d.o.f$_W$ = dfW = $\sum_i n_i - t$

$$\sum_{i,j} y_{ij}^2 \rightarrow \sum_j \sum_i y_{ij}^2$$

Total sum of sqs. = TSS = SSB + SSW

1-way ANOVA table.

Source	d.o.f.	SS	$MS = \frac{SS}{df}$	F_{cal}
B/n groups	$t-1$	SSB	MSB	$\frac{MSB}{MSW}$
W/in gps.	$\sum_i n_i - t$	SSW	MSW	
Total	$\sum_i n_i - 1$	TSS		

If $F_{cal} > F_{table\,(D^n)}$ $\Rightarrow$ reject H_0 in favor of H_1!