ONE WAY ANOVA :

One X (categorical) two (more) independent levels - Y numeric

Eg. X 60 individuals - diets A, B, C, D

Y - weight loss

	sample mean	sample S.D
A	9.18	2.29
B	8.91	2.78
C	12.11	1.99
D	10.54	2.23

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$

H_1: $\mu_p \neq \mu_q$, for $p \neq q$

* at least one differs, not which one or one or more differ

Assumes - normality

Other possibilities:

① non-parametric alternative: KRUSKAL WALLIS (doesn't require normality)

② Bootstrap / Resample

$a = (\text{individual}_3 - \bar{Y})$

Sum of squares total

$SS_{total} = \sum_{all} (\text{individual} - \bar{Y})^2$ — total variability in Y

$s^2_{total} = \frac{SS_{total}}{n-1}$ | sum of squares

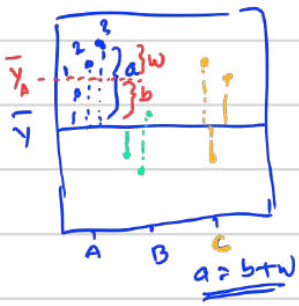
Why a's are different?

① Different diet - Explained by diet - called between (b)

② People are different - Unexplained by X (diet) (w)
- Also called within

So $SS_{total} = SS_{Explained} + SS_{unexplained}$

$= SS_b + SS_w$



$a = b + w$

SS_b or $SS_{Explained} = \sum_{all} (\text{Group } \bar{Y} - \text{Overall } \bar{Y})^2$

Signal → $s^2_b = \text{Mean square bet' groups} (MS_B) = \frac{SS_b}{k-1}$ k: number of groups

SS_w or $SS_{unexplained}$ — also called SS error or SS residual

$= \sum_{all} (\text{individual} - \text{Group } \bar{Y})^2$

Noise → $s^2_w = MS_w = \frac{SS_w}{n-k}$

On to test statistic for ANOVA:

	sample mean	Sample S.D
A	9.18	2.29
B	8.91	2.78
C	12.11	1.79
D	10.54	2.23

i = Groups (A, B, ... K)
$\bar{Y}_i$ = mean of group i
s_i = SD of group i
n_i = sample size of i
$\bar{Y}$ = Overall mean

j = individuals within group
Y_{ij} = individual observation in group i obs. j
Y_{53} = person 3 in 1st group (i.e. A)

Explained $s_B^2\ (MS_B) = \frac{SS_B}{df_B} = \sum_{\text{group}} \frac{n_i (\bar{Y}_i - \bar{\bar{Y}})^2}{K-1} = \frac{97.3}{3} = 32.4$

Unexplained $s_w^2\ (MS_w) = \frac{SS_w}{df_w} = \sum_{\text{all obs}} \frac{(Y_{ij} - \bar{Y}_i)^2}{n-k}$

$= \frac{\sum_{\text{groups}} (n_i - 1) s_i^2}{n-k}$ | like pool variance in t-test

$= \frac{297}{56} = 5.3$

If H_1 true, we expect $MS_B > MS_w$

$F_{\text{statistic}} = \frac{MS_B}{MS_w} > 1$

$F_{stat} = \frac{32.4}{5.3} = 6.1$

If H_0 true $\frac{MS_B}{MS_w} \approx 1$

[F dist^n with 2 df. → $df_n = k-1$, $df_d = n-k$

From: F dist^n → p-value = Prob. ($F_{STAT} \geq 6.1$ if should ≈ 1)

= 0.0011

→ Reject H_0

H_1 just says that dist has a difference "but not which one" → need multiple comparison.

ANOVA multiple comparison: Multiple Testing correction

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$ H_1: At least one pair differs

$F_{STAT} = \frac{s_B^2}{s_w^2} = \frac{32.4}{5.3} = 6.1 \rightarrow p = 0.0011 \Rightarrow$ Reject H_0

All pair wise comparison $\binom{4}{2} = 6$ possible → Independent 2 sample t-test

$(\bar{Y}_1 - \bar{Y}_2) \pm t_{df}^{crit} * SE_{(\bar{Y}_1 - \bar{Y}_2)}$

or $t_{STAT} = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{(\bar{Y}_1 - \bar{Y}_2)}}$

When we do more test, → probability of Type 1 error increases
One test – α 0.5% another test another 0.5% error

Eg. AB, AC, AD, BC, BD, CD – assume each comparison independent

Each test: we $\alpha = 0.05$ → prob. of making Type I error
$P(\text{Type I}) = 0.05$,
$\Rightarrow P(\text{not Type I}) = 0.95$

Overall $P(\text{At least 1 Type I}) = 1 - P(\text{making no type I error})$
(also called Family wise Error Rate) $= 1 - (0.95)^6$
$= 0.265$

Bonferroni approach:
use adjusted $\alpha^* = \frac{0.05}{\#\text{ comparison}} = \frac{0.05}{6} = 0.00833$
for C.I = 95% ⇒ 99.167%

Now, Overall $P(\text{atleast 1 Type error}) = 1 - (0.99167)^6$
$= 0.049$

Pearsons' Chi-square test of Independence:

independence test X & Y – both categorical, 2 / more levels.
Non parametric, but relies on chi-square dist^n like other parametric

X – vaccinated for MMR
Y – diagnosed for Autism

Any relationship? χ^2 – just tells if/not there is an assoc^n, but not how STRONG?

X \ Y	Yes	No	
Yes	621	440,034	440,655
No	117	96,51	96,648
	738	536,565	539,303

Call this observed table "O"

otherway to do → Two proportion test

H_0 : X, Y no association / Independent ⇒ $P_1 = P_2$ ⇒ $P_1 - P_2 = 0$
H_1 : X, Y associated ⇒ $P_1 \neq P_2$ ⇒ $P_1 - P_2 \neq 0$

$\hat{P}_1 = P(\text{AUT} | \text{Vacc}) = \frac{621}{440,655} = 0.00141$

$\hat{P}_2 = P(\text{AUT} | \text{no Vacc}) = \frac{117}{96,648} = 0.00121$

Expected table "E"

fill the values by taking the values in blue from "O" table

X \ Y	Y	N	
Y	605.25	☆	440,655
N	○	★	96,648
	738	536,565	537,303

These are not free ∴ 1 d.f (only 605.25 is free)

Now we know one value, rest can be filled

○ = 738 − 605.25

☆ = 440,655 − 605.25

★ = 96,648 − ○

Assumption of χ^2 test:
- Groups, Obs^s independent
- All cells ≥ 1
- All cells in "E" ≥ 5

If not met Then
- one can
  - → FISHER EXACT TEST
  - → BOOTSTRAPPING

$P(\text{Vacc and AUT}) = P(\text{Vacc})\,P(\text{Aut})$ | Assuming independee

$= \frac{440,655}{537,303} \times \frac{738}{537,303}$

⇒ Expected no. of people = total no. of people × P(Vacc. & Aut)

Vacc. and has AUT

$= 537,303 \times \frac{440,655}{537,303} \times \frac{738}{537,303}$

$= 605.25$

Actually this is simply $\frac{\text{row total} \times \text{column total}}{\text{overall total}}$

Test-static: compare "O" to "E" table

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = \frac{(621-605.25)^2}{605.25} + \dots + \frac{(96531-96515.25)^2}{96515.25}$$

$= 2.28$

↳ $\sim \chi^2$ dist^n (if H_0: TRUE) d.f = (# rows − 1)(# cols − 1)

So, here p-value = $P(\chi^2_{STAT} \geq 2.28 \text{ if } H_0 \text{ TRUE}) = 0.1309$

Accept H_0

NOTE:
- For large sample sizes, small differences may show up to be statistically significant
- For small sample sizes, big " " " NOT " " " " "

CONTINUITY CORRECTION: - Above example, had numbers of persons - discrete case

- In a discrete case $P(X > 10)$ is same as $P(X \geq 11)$
- In a continuous case $10 \neq 10.5 \neq 10.3$, suggestion is we a mid point

$P(X > 10) = P(X \geq 10.5)$

✱ To know how strong is assoc^n < RISK RATIO, RISK DIFFERENCE, ODDS RATIO