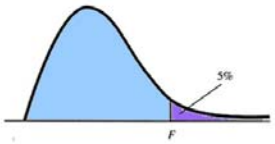
2 way ANOVA



- The principle use of a two - way ANOVA is situations where there are *two* possible factors which can affect the values of observations in a sample.

Example: Pollution

General Form: X = f(T,B)
Specific Form: $X_{jk} = \mu + \tau_j + \phi_k + \varepsilon_{jk}$
Verbal Form: Pollution = f(Time of Day, Freeway)

There is only one observation per treatment.

	Time of Day				
Freeway	12:00A - 6:00A	6:00A - 10:00A	10:00A - 3:00P	3:00P - 7:00P	7:00P - 12:00A
Chrysler	145	364	300	475	300
Davidson	125	205	215	301	201
Reuther	85	150	150	225	104
Lodge	150	346	294	450	274

X = nitric oxide pollution level (ppm)
T = time of day (5 treatments)
B = freeway (4 blocks)

- Example : Test H_2O from streams that run over different bases and at different stages of each rivers' development.
  - Sample the dissolved $CaCO_3$ content at (i) different stages of stream (ii) for streams / rivers that run over different bases.

	River Stage					
	Headwater	Middle	Mouth		$\bar{x}_i$	Deviation from $\bar{\bar{x}}$
River 1 (Shield)	0.30	0.40	0.45	i=1	0.38	-0.08
River 2 (Till)	0.35	0.32	0.32	i=2	0.33	-0.13
River 3 (Limestone)	0.60	0.71	0.72	i=3	0.67	0.21
	j=1	j=2	j=3			
$\bar{x}_j$	0.41	0.47	0.49			
Deviation from	-0.05	0.01	.03			

where r=number of rows
k= number of columns

- In general, it seems (i) bedrock seems to influence $CaCO_3$ content
- (ii) river stage seems to influence $CaCO_3$ content (increases downstream *but* are these relationships significant or due to chance?)

- There are *three* possible sources of variation in this data :
- (i) Variation between rows (streams) - MSR - estimate of population variance based on between row variation.
- (ii) Variation between columns (stages) - MSC - estimate of population variance based on between column variation.
- (iii) Remainder / residual variance - MSE - estimate of population variance based on overall variance.

- With two - way ANOVA, we can test 2 Null Hypotheses :
- (i) H_{0r} : There is no significant difference in $CaCO_3$ content between the different rivers. That is, the bedrock over which the river runs does not influence $CaCO_3$ content.
- $\mu_{(i = 1)} = \mu_{(i = 2)} = \mu_{(i = 3)}$

- (ii) H_{0k} : The is no significant difference in $CaCO_3$ content between the different rivers. That is, the distance over which the river flows over its bedrock does not influence $CaCO_3$ content.
- $\mu_{(j = 1)} = \mu_{(j = 2)} = \mu_{(j = 3)}$

- To Test (i) : $F = \frac{MSR}{MSE}$ $MSR = \frac{SSR}{r-1}$
  - df_1 = r-1=3-1=2
  - df_2 =(r-1)(k-1)= (2)*(2)=4
    - Where : T_i Sum of Observations in each row i (j = 1 ... k)
    - r = Number of Rows
    - k = Number of Columns N = Total Number of Observations

$$SSR = \frac{1}{k}\left(\sum_{i=1}^{r} T_i^2\right) - \frac{1}{n}T^2 \qquad T_i = \sum_{j=1}^{k} x_{ij}$$

- (Total Sum of Observations) $T = \sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}$
- T_1 = 0.30 + 0.40 + 0.45 = 1.15
- T_2 = 0.35 + 0.32 + 0.32 = 0.99
- T_3 = 0.60 + 0.71 + 0.72 = 2.03
- T = T_1 + T_2 + T_3 = 4.17

- To test (ii) : $F = \frac{MSC}{MSE}$
  - Where : df_1 = k-1 = 3-1=2
  - df_2 = (r-1)(k-1)= (2)*(2)= 4

$$MSC = \frac{SSC}{k-1}$$

$$SSC = \frac{1}{r}\left(\sum_{j=1}^{k} T_j^2\right) - \frac{1}{n}T^2 \qquad T_j = \sum_{i=1}^{r} x_{ij}$$

$$SSR = \frac{1}{3}(1.15^2 + 0.99^2 + 2.03^2) - \frac{4.17^2}{9} = 0.2091$$

$$MSR = \frac{0.2091}{3-1} = 0.1045$$

- Sum of Observations in each column j (i=1... r)
- r = Number of Rows
- T_1 = 0.30 + 0.35 + 0.60 = 1.25
- T_2 = 0.40 + 0.32 + 0.71 = 1.43
- T_3 = 0.45 + 0.32 + 0.72 = 1.49

$$SSC = 1.94324 - 1.9320 = 0.0104$$

$$SSC = \frac{1}{3}(1.25^2 + 1.43^2 + 1.49^2) = \frac{1}{9}(4.17^2)$$

$$MSC = \frac{0.0104}{3-1} = 0.0052$$

- For both (i) and (ii), we need to find MSE.

$$MSE = \frac{SSE}{(r-1)(k-1)}$$

If : SST = SSR + SSC + SSE
Then : SST − SSR − SSC = SSE
↑ Unknown ↘ Known

$$SST = (\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}^2) - \frac{1}{n}T^2$$

$$\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij} = 2.16$$

$$\frac{1}{n}T^2 = 1.932$$

$$SST = 2.1623 - 1.9320 = 0.2302$$

- SSE=SST-SSR-SSC
- SSE=0.2302 - 0.2091 - 0.01040 = 0.0107
- MSE = 0.0107/4
- MSE = 0.0027

Presentation of results

Source of variation	df	Sum of Squares	Mean square	F ratio
Between rows	r-1 (2)	SSR (0.2091)	SSR/r-1 (0.1046)	MSR/MSE (39.10)
Between columns	k-1 (2)	SSC (0.0104)	SSC/k-1 (0.0052)	MSC/MSE (1.94)
Error	(r-1)(k-1) (4)	SSE (0.0107)	SSE/(r-1)(k-1) (0.0027)	
Total	n-1	(0.2302)		

- F_{Rows} = 39.10
- F_{Cols} = 1.94
- Critical Values -
  - Rows (df_1 = 2, df_2 = 4, α = 0.05) = 6.94
  - Cols (df_1 = 2, df_2 = 4 , α = 0.05) = 6.94

Critical Values of the *F* Distribution

$\alpha = .05$

df_2 \ df_1	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18.5	19.0	19.2	19.2	19.3	19.3
3	10.1	9.6	9.3	9.1	9.0	8.9
4	7.7	6.9	6.6	6.4	6.3	6.2
5	6.6	5.8	5.4	5.2	5.1	5.0
6	6.0	5.1	4.8	4.5	4.4	4.3
7	5.6	4.7	4.4	4.1	4.0	3.9
8	5.3	4.5	4.1	3.8	3.7	3.6

- F_{ROWS} > Critical Value :
- Can reject H_{0r} - There is a significant difference in $CaCO_3$ content between rivers. Thus bedrock over which rivers flow influences $CaCO_3$ content of H_2O.

- F_{COLS} < Critical Value :
- Cannot reject H_{0k} - There is no significant difference in $CaCO_3$ content between river stages. Thus, the distance over which rivers flows does not influences $CaCO_3$ content of H_2O.

2 way with interactions

- the way to do a 2 way ANOVA with interaction effects is to break it into steps
- 1) perform a 1 way ANOVA on each variable separately to obtain the sums of squares for each

$$SSC = \sum_{i=1}^{r} n_i (\bar{x}_i - \bar{\bar{x}})^2$$

- 2) determine the sum of squares for the combination of the 2 variables A and B by using formula
  - where n_{ij}= is the frequency of the cell
  - $\bar{x}_{ij}$ is the mean of the cell
  - $\bar{\bar{x}}$ is the overall mean

$$SSAB = \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}(\bar{x}_{ij} - \bar{\bar{x}})^2$$

- 3) using the fact that SSI = SSAB -SSC –SSR determine the interaction term SSI
- 4) find the total sum of squares

$$SST = (\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}^2) - \frac{1}{n}T^2$$

- 5) then using the fact that
- TSS = SSC + SSR + SSI + SSE
- find the residual sum of squares

With Interaction

Model Form

$X_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$

Note α_j and β_k are called the *main effects.*

Hypotheses

H_0: $\alpha_j = 0$ (no row effect exists)

H_1: $\alpha_j \neq 0$ (row effect exists)

H_0: $\beta_k = 0$ (no column effect exists)

H_1: $\beta_k \neq 0$ (column effect exists)

H_0: $\alpha\beta_{jk} = 0$ (no interaction exists)

H_1: $\alpha\beta_{jk} \neq 0$ (interaction effect exists)

Definitions

X_{ijk} = i^{th} obs. in row j and col. k

μ = common mean

α_j = effect due to row j

β_k = effect due to column k

$\alpha\beta_{jk}$ = effect due to interaction

ε_{ijk} = random error

DVD Sales

General Form: X = f(A,B,AB)
Specific Form: $X_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$
Verbal Form: Sales = f(StoreSize, DisplayLocation, StoreSize*DisplayLocation)

Store Size	Display Location: Check Out	Mid-Shelf	Aisle End
Small	255	187	209
	233	175	212
Medium	414	350	401
	412	367	395
Large	643	512	619
	702	487	505

Sales were recorded for two stores in each treatment.

X = weekly sales in hundreds of units
A = display location (3 treatment levels)
B = store size (3 treatment levels)
AB = interaction effect

- You must have multiple observations in each combination of main effects
- If you don't, you have a ANOVA with no interaction. Why?
- If at all possible the design should be balanced

- **Balanced/Unbalanced Factorial Designs**. A **balanced** factorial design is one that has the same number of observations in every cell. Unbalanced designs do not have the same number.
- The calculations for unbalanced designs are more complex and the interpretation can be very unclear.
- It is best to avoid these unbalanced data, but in survey research such analyses are common.

correlation ratio or eta squared

$$H^2 = \frac{TSS - RSS}{TSS}$$

- where TSS is total sum of squares
- RSS is the residual sum of squares
- it provides the % of variance explained by the variables in the ANOVA

Example

		country		mean
		Germany	US	
Education	Low	5 4 (4.5)	7 5 (6.0)	5.25
	High	4 4 3 (3.67)	6 6 6 (6.0)	4.83
Mean		4.0	6.0	5.00

$$SST = (\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}^2) - \frac{1}{n}T^2$$

- TSS = 25+16+49+25+16+16+9+36+36+36 - .1(2500) = 264 - 250 = 14
- TSS = SSR + SSC + SSI + SSE
- 14 = 10 + .42 + .38 + ?
- SSE = 3.2

eta squared

- for country: 14 - (.42+.38+3.2)/14 = 10/14 = .714
- for education: 14 - (10+.38+3.2)/14=.03
- for interaction 14 - (10+.42+3.2)/14=.03

Source of variation	Sum of squares	df	Mean square	F
country	10.0	(r-1) 1	10.0	18.87
education	.42	(k-1) 1	.42	.79
interaction	.38	(r-1)(k-1) 1	.38	.74
residual	3.2	rk(n-1)	.53	
total	14.0	N-1		

- critical value is df=6,1 5.99 at α=0.05
- when n is number of replications in a cell formula only good for same number of replications per cell
- for nonequal number of observations per cell residual df seems to be N-k-r as in this example
- the best way to calculate df for residual is to subtract total df for SSC+SSR +SSI from N-1