Analysis of Variance

- ANOVA is used to compare the means of two or more populations.
- Procedure is based on the spread (variance) between sample averages of populations and spread within sample averages.
- Possibly the most widely used procedure across disciplines.

Example

Consider a cereal manufacturer who wants to evaluate the impact on sales of four package designs. Ten stores are randomly assigned to one of the designs and sales data are collected for a given period.

Package	Store					Number
Design	1	2	3	Total	Mean	of stores
1	12	18		30	15	2
2	14	12	13	39	13	3
3	19	17	21	57	19	3
4	24	30		54	27	2
All designs				180	18	10

This type of design is called a **Completely Randomized Design.**

It's called analysis of VARIANCE!

Recall that variance is the "almost average" of the squared differences of a set of data around its mean.

For this set of data then, we have:

$(12 - 18)^2 + (14 - 18)^2 + \ldots + (21 - 14)^2 = 304$ units of variation

Variation

What can account for this variation?

- type of package design (SSB) → treatment
- everything else (SSE) → error

The total variation can be expressed in this relationship:

$$SST = SSB + SSE$$

Why didn't we sell the same amount of each package type?

Why not the same at each store for package 1? (or 2?, or 3?, or 4?)

----- thousands of extraneous factors!

$SSE = (12 - 15)^2 + (18 - 15)^2$	package 1
$+ (14 - 13)^2 + (12 - 13)^2 + (13 - 13)^2$	package 2
$+ (19 - 19)^2 + (17 - 19)^2 + (21 - 19)^2$	package 3
$+ (24 - 27)^2 + (30 - 27)^2$	package 4

= 46 total units of variation

What number best represents the long-term average for package 1?

The package 1 average (just as package 2 average does for package 2 designs, and so forth.

How do these package averages vary from the overall average?

$(15 - 18)^2 + (13 - 18)^2 + (19 - 18)^2 + (27 - 18)^2 = 116$

But, we must weigh each one by the number of observations in that average -- that gives us ... $2(9) + 3(25) + 3(1) + 2(81) = 258$

(note: 304 = 258 + 46)

OK -- we've got some 'sum of squares'

But this procedure is called 'analysis of variance'

FACT: variance = sum of squares divided by appropriate df

Lets organize our results so far in an ANOVA table:

Sources of variation	Sum of squares	df	Variance
SSB	258	3	86
SSE	46	6	7.67
Total	304	9	

degrees of freedom

As the name implies, this is the number of things that are free to vary and still get the same result. For example, if I told you the average of five numbers is 7, you could pick any four numbers, and if I can pick the fifth I can ensure the average is 7.

Generally speaking, the df will be one less than the number of things being compared. For example,

SSB df = 4(package designs) - 1 = 3

SSE df = (2 - 1) + (3 - 1) + (3 - 1) + (2 - 1) = 6

Total df = 10 - 1 = 9

A ratio of variances

We next form a ratio of variances = 86 / 7.67 = 11.2

We need a reference distribution to evaluate this -- we compare it to the Fisher distribution (F distribution)

A Short Fisher Table

df for Denominator	Degree of Confidence	df for Numerator 1	2	3
2	95	18.51	19.00	19.16
	99	98.50	99.00	99.17
6	95	5.99	5.14	4.76
	99	13.75	10.92	9.78

Making a decision

- We are really carrying out a hypothesis test.

  Our Ho is that are package design means are equal.

  Ho: $m_1 = m_2 = m_3 = m_4$

  Our Ha is that at least one mean is different than the rest.

- We can make two decisions with our data:

1. No difference in means. This is one of those less than 1 in a 100 times we would get a value this large.
2. The null is false -- we reject the null and accept H a.

Summary of ANOVA concept

1. Decompose the total sum of squares.
2. Convert sum of squares into variances.
3. Compute variance ratio and compare to F table.

Assumptions!

1. Populations being compared are normally distributed -- moderate departures OK -- "robust" in this regard.
2. Variances of the populations are equal -- can be tested -- if this assumption is not met there is "trouble in River City."
3. Observations are statistically independent (use randomization).

Now that we know the means are different what do we do?

We know at least one mean is different from the rest -- which one or ones?

We use ' multiple comparison procedures'

- Fisher's LSD (pre-Woodstock!)
- Fisher's LSD with Bonferroni adjustment
- Tukey's procedure

Two - Way Analysis of Variance

The most common two-way analysis is called a Randomized Block Design. The purpose of this design is to control some the extraneous variation that is in the error term.

Blocking is used to form 'homogeneous' groups.

Example

Suppose five people whose favorite soft drink is Big K are asked to rate four brands of soft drink. We are using a person as a 'block' in this case because we suspect some people will rate all drinks higher than some other person (reflected as a difference in block means), but what we want to find out is whether there is a difference in brands of soft drink (treatment means).

Results of survey

Person	Favorite	X	Y	Z	Block mean
Jones	63	59	62	61	61.25
Smith	61	62	57	63	60.75
Klein	61	64	60	58	60.75
Carlucci	62	62	60	62	61.50
Weill	58	63	61	61	60.75
Treatment mean	61	62	60	61	

Sum of squares

Total sum of squares =

treatment sum of squares +

block sum of squares +

error sum of squares

ANOVA table

Source of Variation	Sum of squares	Degrees of freedom	Variation	F
Treatment	10	3	3.33	0.74
Blocks	2	4	0.50	0.11
Error	54	12	4.50	
Total	66	19		

Computations

Total sum of squares = $(63-61)^2 + (59-61)^2 \ldots + (61-61)^2 = 66$

Treatments sum of squares = $5[(61-61)^2 + .. + (61-61)^2] = 10$

Block sum of squares = $4[(61.25-61)^2 + \ldots (60.75-61)^2] = 2$

Error sum of squares = 66 - 10 - 2 = 54

Conclusions

- Compare results to appropriate F values

  Treatment $F_{.05,3,12} = 3.49$

  Block $F_{.05,4,12} = 3.26$

- Conclude -- no difference in individuals (blocks) and no difference in average ratings of four brands of soft drink.

Blocking

Objective: to remove the impact of the blocking factor from the sum of squares due to extraneous factors term.

~~**Examples**~~ (with possible blocking factors):

- 3 types of exterior paint
  - * painter's experience
  - * orientation of exterior wall
- 4 teaching approaches for CPA review course
  - * years since graduation
  - * type of degree
  - * day vs. night course

Rule of Thumb

When you design an experiment, list as many extraneous factors as you can.

If there is one factor that you think will dominate (have a major impact) on the dependent variable, then you should block on this factor.

Common blocking factors include: shift, time of day, people, machines, etc.

Multifactor or Factorial Designs

DEFINITIONS:

Experimental factors -- the factors that we wish to test so that we can make timely decisions.

Factorial design -- has at least two experimental factors and does not use blocking. Uses randomization to minimize the impact of the remaining extraneous factors.

Example

Suppose we are investigating the effect of teaching approaches and GMAT score on student performance in a statistics class.

We will have two factors:

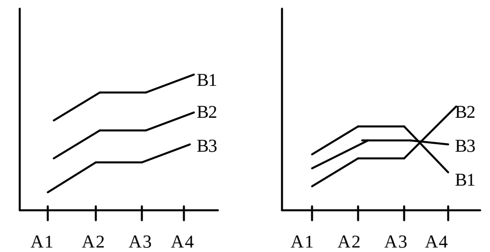
A -- at four levels	B -- at three levels
A1 = lecture	B1 = 400 - 450
A2 = discussion	B2 = 500 - 550
A3 = discovery	B3 = 600 - 650
A4 = classroom	

Visual representation of the design space

Factor A

Factor B	A1	A2	A3	A4
B1				
B2				
B3				

Graphical representation of possible results



Interaction – the best level of one factor depends on the level of the other factor.

Data analysis

Factor A

Factor B	A1	A2	A3	A4	
B1	58 62	60 60	64 66	59 61	61.25
B2	99 99	89 91	80 80	75 75	86
B3	74 76	80 80	85 85	94 96	83.75
	78	76.67	76.67	76.67	

Computing sum of squares

Total variation in all 24 observations =

SST = $(58\text{-}77)^2 + (62\text{-}77)^2 + \ldots + (96\text{-}77)^2 = 4174$ units

What can account for all this variation?

- factor A
- factor B
- extraneous factors
- interaction between A and B

Sum of squares for treatments

The four column averages best represent the effect due to the 4 teaching approaches --

SSTR - A = 6 [$(78\text{-}77)^2 + (76.67\text{-}77)^2 + (76.67\text{-}77)^2 + (76.67\text{-}77)^2$] = 7.96 units of variation

remember to account for the number of observations for each treatment method!

SSTR - B = 8 [$(61.25\text{-}77)^2 + (86\text{-}77)^2 + (83.75\text{-}77)^2$] = 2997 units

Sum of squares for extraneous factors

Extraneous factors -- best estimate for a cell (two observations with the same treatment combination) will be the cell average ...

SSE = (58-60)2 + (62-60)2 <-- extraneous variation: cell 1

+ (60-60)2 + (60-60)2 <-- extraneous variation: cell 2

.

.

.

+ (94-95)2 + (96-95)2 <-- extraneous variation: cell 12

= 12 units of variation

Sum of squares for interaction

The relationship is

Total = factor A + factor B + interaction + extraneous

Therefore, the interaction sum of squares will be what is left over = 1157.04 units of variation

Degrees of freedom

Total degrees of freedom = n-1 = 24-1 = 23

Factor A df = number of levels - 1 = 4-1 = 3

Factor B df = number of levels - 1 = 3-1 = 2

Extraneous factors df = two observations per cell --> 2-1 =1, but we have 12 cells --> df = 12

Interaction df = A df x B df = 3 x 2 = 6

ANOVA table

Sources	SS	df	Variance	Variance ratio
Factor A	7.96	3	2.65	2.65
Factor B	2997	2	1498.5	1498.5
Interaction	1157.04	6	192.8	192.8
Extraneous	12	12	1	
Total	4174	23		

F table (6,12,.01) = 4.82

Interpretation?

Profile