Displays for Statistics 5303

Lecture 32

November 20, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

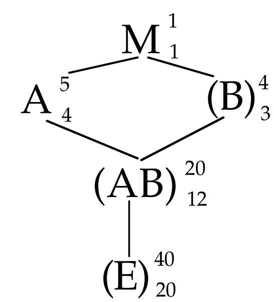
Hasse Diagrams

You have seen how important it is to
- find expected mean squares
- select denominators for tests

For balanced data there is an important tool -- the **Hasse Diagram**

This represents ANOVA model in semi-graphical form giving information on:
- Which factors are fixed and random
- The number of effects for each term
- The degrees of freedom for each term

Two way factorial with A fixed (a=5) and B random (b=4). There are 5 *nodes*, one for each term.
(...) means random term.
Subscript = DF
Superscript = number of effects.

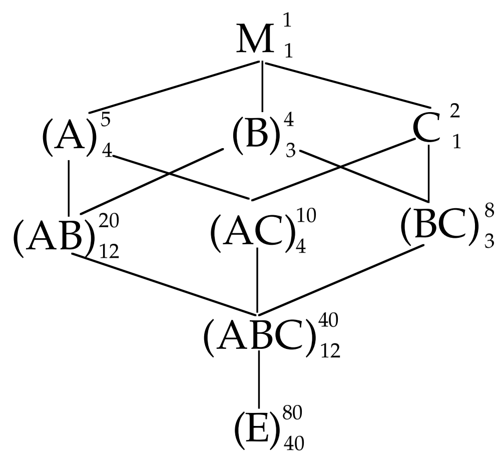


5 nodes

M_1^1 is *above* everything. A_4^5 and $(B)_3^4$ are above $(AB)_{12}^{20}$ which is above $(E)_{20}^{40}$

Dfn Node U is *above* V in a lower row if they are connected by a path of lines.

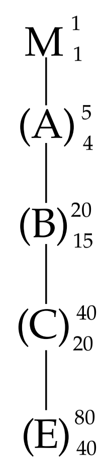
3-way factorial with A and B random (a=5, b=4), C fixed (c=2).



Fully nested with A and B random (a = 5, b = 4) and **C fixed** (c = 2).

C is in (...) because it is nested in a random factor.

This makes the actual levels of C present random, even though they are fixed once the levels of B are selected.



The rules for making and using Hasse diagrams are not as complicated as they may seem, but they are confusing.

I'm going to reverse the order in the book and start with the rules for making them.

A Hasse diagram consists of rows of "nodes", each consisting of a sequence of letters possibly contained in parentheses.

Each node corresponds to a potential model term corresponding in a natural way to the letters in the node.

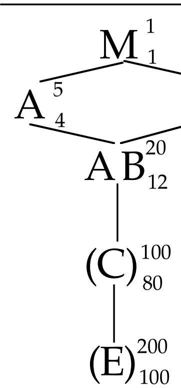
Some or all the nodes in row i are connected with lines to one or more notes in row i+1.

Conventionally, the top row is number 0.

One source of confusion is the concept of "implied" factors in a node. They contrast with *explicit* factors - factors whose letters are in the node.

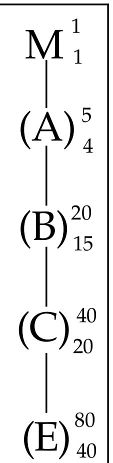
Implied factors occur only in a node representing a nested term. The implied factors are any "parent" factors.

a = 5, b = 4, c = 5, n = 2



A and B are implied in (C)

a = 5, b = 4, c = 5, n = 2



A and B are implied in (C)

A is implied in (B)

You could use a notation like (C_{AB}), (C.AB) or (C[A][B]) that that makes implied factors visible but that's not done.

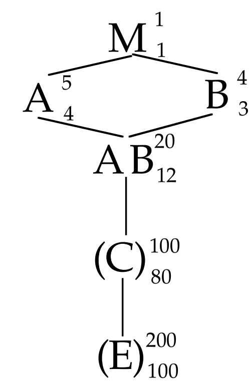
When you add a node, enclose it in (...) if the corresponding term is random.

Row	What you do
0	Put M, corresponding to μ
1	Put a node connected to M (by a line) for every factor except factors nested in another factor
2	(a) Put a node for any factor nested in a row 1 node and connect it to its parent node. (b) Add nodes for terms with 2 explicit or implied factors and connect them to including nodes above them
i	(a) Put a node for any factor nested in a row i-1 node and connect it to its parent node. (b) Add nodes for any term with i explicit or implied factors and connect them to including nodes in row i-1
Bottom	Add E connected to all nodes just above it

There remains adding the subscripts and superscripts.

Super-scripts	The **number of effects** in the term = product of the number of levels of explicit or implied factors (for example ac on AC)
Sub-scripts	The **degrees of freedom** for the term. M has 1 d.f. For any other term, subtract from the superscript the sum of all degrees of freedom for terms connected by one or more lines above it

For the degrees of freedom, of course you can use the usual formulas for main effects and interactions, such as a-b, b-1, (a-1)(b-1), ..., but the rule given covers all cases.



a = 5, b = 4, c = 5, n = 2
ab = 20, abc = 100,
abcn = 200

$df_A = 5 - 1 = 4$

$df_B = 4 - 1 = 3$

$df_{AB} = 20 - 4 - 3 - 1 = 12$

$df_C = 100 - 12 - 4 - 3 - 1 = 80$

$df_e = 200-80-12-4-3-1$

Problem 12.1 (a)

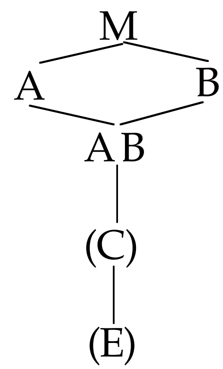
y = atmospheric sulfate concentration measured by drawing air through a filter and then chemically anlyzing the filter.

- There are a = 4 brands of filter.
- There are b = 2 analytical methods
- Randomly select 16 filters from each brand, 64 in all, c = 8 for each type of analysis and draw air through them.
- Cut each filter in n = 2 pieces and analyze each piece by the method for that filter.

A and B are fixed and crossed with a = 4 and b = 2.

C is random, nested in AB with c = 8

E is nested in C, n = 2



ab = 4×2 = 8
abc = 4×2×8 = 64
abcn = 4×2×8×2 = 128

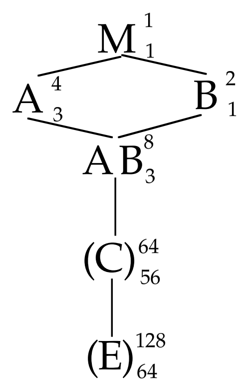
df_A = 4 - 1 = 3

df_B = 2 - 1 = 1

df_{AB} = 8 - 1 - 3 - 1 = 3

df_C = 64 - 3 - 1 - 3 - 1 = 56

df_{error} = 128 - 64 = 64



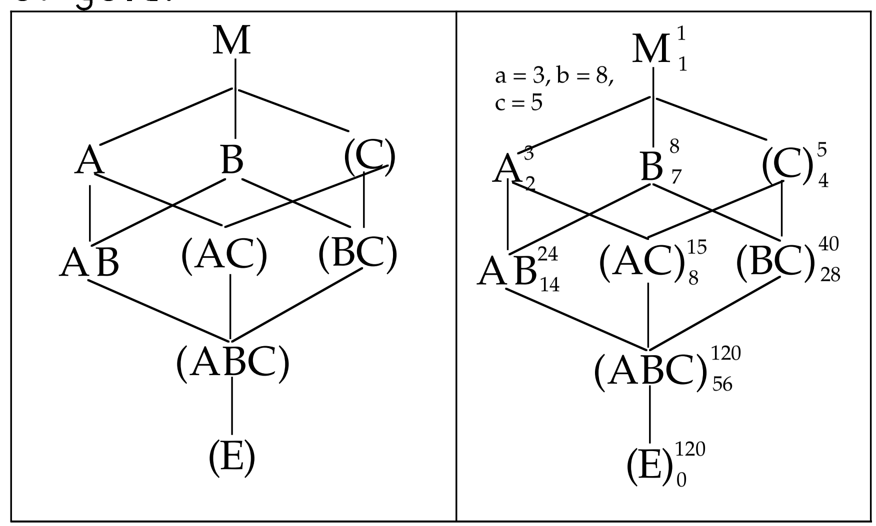
Problem 12.2

y = hardness of gold fillings

Fixed factor A = condensation method for producing gold alloy, a = 3

Fixed factor B = gold alloy used, b = 8

Random factor C = dentist who makes fillings; c = 5 dentists selected randomly, each to make a filling using gold from each of the 3×8 = 24 types of gold.



The d.f. for E is 0 because n = 1.

Using Hasse diagrams for finding F-statistics

Basically you use the Hasse diagram to find the denominator and then add things to the numerator MS so that the terms not tested are the same in both.

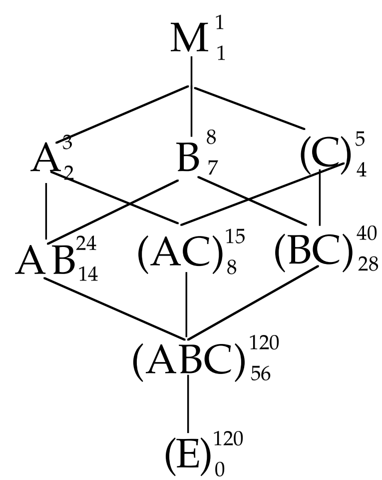
Key concepts are

- an **eligible** term for a denominator
- a **leading eligible** term.

The definition of what is eligible differs between restricted and unrestricted.

- Only random terms below U, the term tested, are ever eligible
- *Unrestricted*: *All* random terms below U are **eligible**
  Restricted: All random terms below U are **eligible** except those containing a fixed factor not in U
- An eligible term V below U is **leading** if there is no eligible random term below V and above U

- The denominator for U is a sum of all leading eligible terms
- If there is more than 1 leading eligible term, the F test will be approximate



Denominator for A
U&R: MS_{AC}

Denominator for B
U&R: MS_{BC}

Denominator for C
U $MS_{AC} + MS_{BC}$
Need MS_{ABC} on top
R MS_{ABC}

The denominator for AB is MS_{ABC} in *both* restricted and unrestricted case.

The *unrestricted* denominator for both AC and BC is MS_{ABC}.

The *restricted* denominator for AC and MS_{BC} is E because ABC has a fixed factor not in either of them.

Use in Expected mean squares

This uses the same definition of *eligibility* as for selecting F denominators

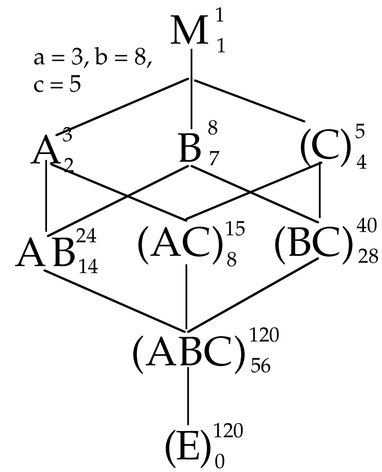
- *Unrestricted*: All random terms below U are **eligible**
  Restricted: All random terms below U are **eligible** except those containing a fixed factor not in U

The concept of **leading eligible** terms does *not* apply

Representative elements for term

- *Fixed*: Q = $\sum$(all effects)^2/DF
  e.g. $\sum_i\sum_j \alpha\beta_{ij}^2/(a-1)(b-1))$
- *Random*: V = variance component (σ_x^2 for pure random, $r_x\sigma_x^2$ for mixed)
- The contribution of a term is N/(number of effects) (e.g.. N/(bc))
- EMS_U = sum of contributions of all eligible random terms below U

U = unrestricted, **R** = unrestricted



R $EMS_A = 40Q_A + 8\sigma_{\alpha\gamma}^2 + \sigma^2$

U $EMS_A = 40Q_A + 8\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma^2$

R $EMS_B = 15Q_B + 3\sigma_{\beta\gamma}^2 + \sigma^2$

U $EMS_B = 15Q_B + 3\sigma_{\beta\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma^2$

R $EMS_C = 24\sigma_{\gamma}^2 + 3\sigma_{\beta\gamma}^2 + \sigma^2$

U $EMS_C = 24\sigma_{\gamma}^2 + 8\sigma_{\alpha\gamma}^2 + 3\sigma_{\beta\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma^2$

UR $EMS_{AB} = 5Q_{AB} + \sigma_{\alpha\beta\gamma}^2 + \sigma^2$

R $EMS_{AC} = 8\sigma_{\alpha\gamma}^2 + \sigma^2$

U $EMS_{AC} = 8\sigma_{\alpha\gamma}^2 + \sigma_{\alpha\beta\gamma}^2 + \sigma^2$

RU $EMS_{ABC} = \sigma_{\alpha\beta\gamma}^2 + \sigma^2$