Bad Statistics

R. A. Bailey

Queen Mary
University of London

r.a.bailey@qmul.ac.uk

Visualization and Presentation of Statistics,
Open University, 18 May 2011

Once upon a time . . .

. . . John Gower was head of the Statistics Department at Rothamsted Experimental Station
(and I was a member of the department).

Once upon a time . . .

. . . John Gower was head of the Statistics Department at Rothamsted Experimental Station
(and I was a member of the department).

Senior management told him that the only acceptable way to display the proportions in the categories A and not-A was a pie-chart.

Once upon a time . . .

. . . John Gower was head of the Statistics Department at Rothamsted Experimental Station
(and I was a member of the department).

Senior management told him that the only acceptable way to display the proportions in the categories A and not-A was a pie-chart.

John disagreed rather strongly.

From a conference poster in December 2010

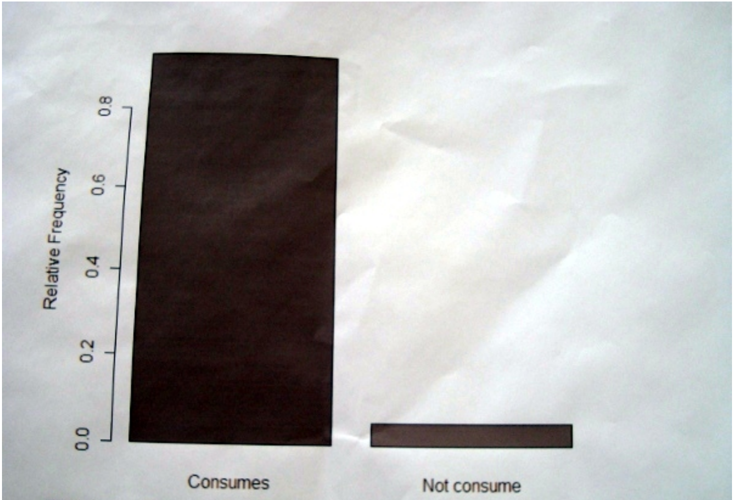
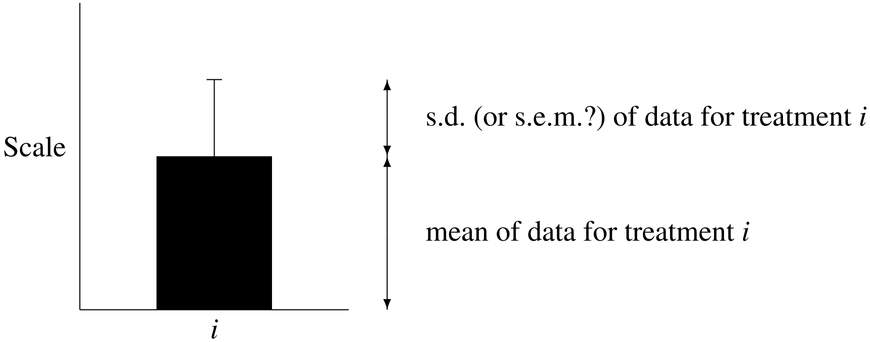


Figura 1: Proportion of residents who eat chocolate or not.

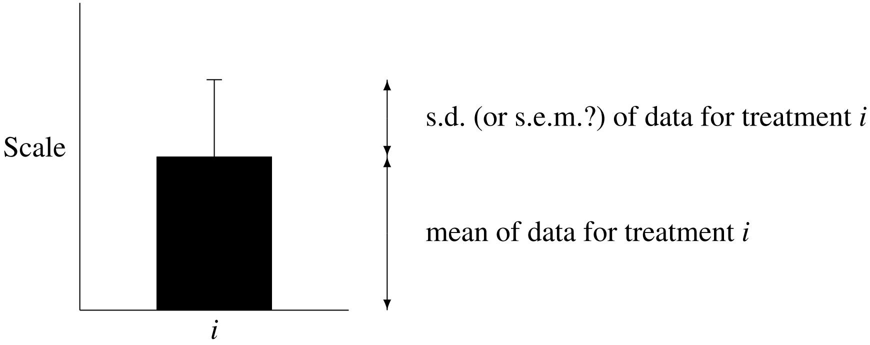
How should we display the results of an experiment?

Medical scientists, biologists, engineers, … seem fond of the convention that the data for each treatment should be summarized on a “bar-and-antenna” diagram.



How should we display the results of an experiment?

Medical scientists, biologists, engineers, ... seem fond of the convention that the data for each treatment should be summarized on a "bar-and-antenna" diagram.



Is this always appropriate?

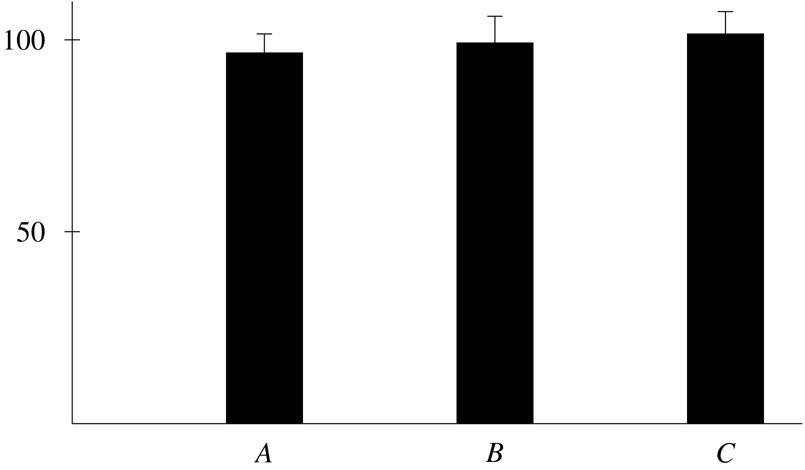
What if the experiment is in randomized complete blocks?

An experiment was conducted to compare two protective dyes (B and C) for metal, both with each other and with ‘no dye’ (A).

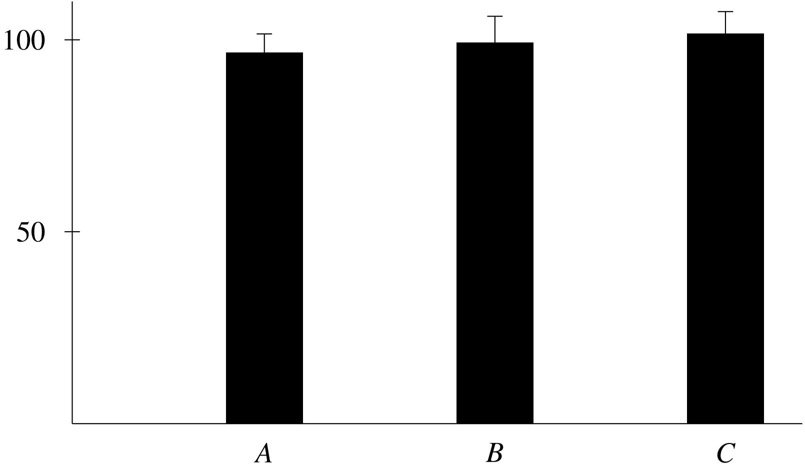
Ten braided metal cords were broken into three pieces. The three pieces of each cord were randomly allocated to the three treatments. Thus the cords were blocks.

After the dyes had been applied, the cords were left to weather for a fixed time, then their strengths were measured.

Simple display of results for dyes on metal cords



Simple display of results for dyes on metal cords



But these standard deviations include the variability between cords!

A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73

A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73
subtracting cord deviations	mean	96.67	99.29	101.62
	s.d.	3.03	3.96	2.55

A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73
subtracting cord deviations	mean	96.67	99.29	101.62
(but based on wrong df!)	s.d.	3.03	3.96	2.55

Pooled estimate of σ is 3.956 so

A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73
subtracting cord deviations	mean	96.67	99.29	101.62
(but based on wrong df!)	s.d.	3.03	3.96	2.55

Pooled estimate of σ is 3.956 so

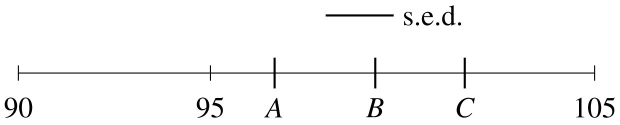
$$\text{standard error of difference} = \sqrt{\frac{2}{10}} \times 3.956 = 1.77.$$

A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73
subtracting cord deviations	mean	96.67	99.29	101.62
(but based on wrong df!)	s.d.	3.03	3.96	2.55

Pooled estimate of σ is 3.956 so

$$\text{standard error of difference} = \sqrt{\frac{2}{10}} \times 3.956 = 1.77.$$

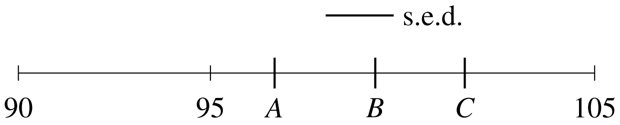


A more helpful display

	treatment	A	B	C
raw data	mean	96.67	99.29	101.62
	s.d.	4.87	6.84	5.73
subtracting cord deviations	mean	96.67	99.29	101.62
(but based on wrong df!)	s.d.	3.03	3.96	2.55

Pooled estimate of σ is 3.956 so

$$\text{standard error of difference} = \sqrt{\frac{2}{10}} \times 3.956 = 1.77.$$



Isn't this a more useful visual summary?

Experiments in randomized complete blocks: summary

The “antenna” part of the “bar-and-antenna” diagram is completely misleading.

What if blocks are incomplete?

Then the estimate of each treatment mean is no longer the same as the mean of the data for that treatment,

What if blocks are incomplete?

Then the estimate of each treatment mean is no longer the same as the mean of the data for that treatment,
so the "bar" part of the "bar-and-antenna" diagram
is even more misleading.

What if the treatments are structured?

Treatments may be factorial, or quantitative, or have other structures.

What if the treatments are structured?

Treatments may be factorial, or quantitative, or have other structures.

In this case, we probably want to consider several possible models for the response.

What if the treatments are structured?

Treatments may be factorial, or quantitative, or have other structures.

In this case, we probably want to consider several possible models for the response.

We can show the family of potential fitted models on a Hasse diagram.

What if the treatments are structured?

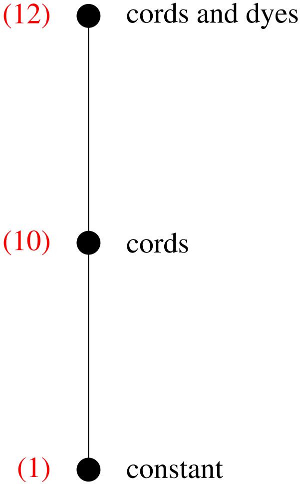
Treatments may be factorial, or quantitative, or have other structures.

In this case, we probably want to consider several possible models for the response.

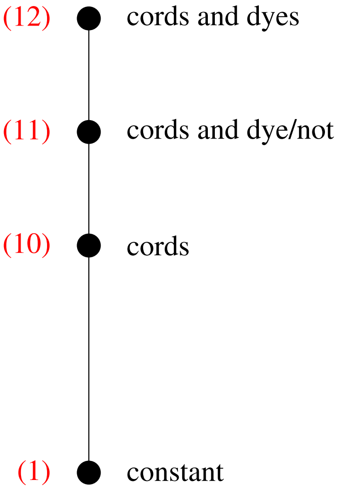
We can show the family of potential fitted models on a Hasse diagram.

We can summarize the ANOVA table graphically by scaling the lines in the Hasse diagram.

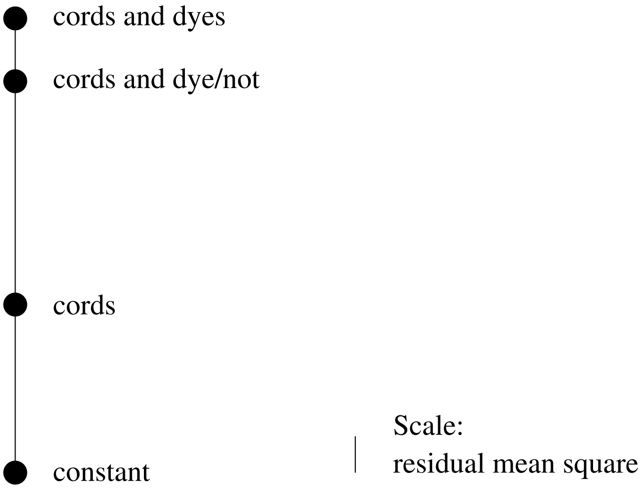
Metal cords: Hasse diagram of models



Metal cords: Hasse diagram of models



What the data showed: lengths are mean squares



An experiment on biodiversity

A, B, C, D, E, F — six types of freshwater "shrimp".
Put 12 shrimps in a jar containing stream water and alder leaf litter.
Measure how much leaf litter is eaten after 28 days.

An experiment on biodiversity

A, B, C, D, E, F — six types of freshwater “shrimp”.
Put 12 shrimps in a jar containing stream water and alder leaf litter.
Measure how much leaf litter is eaten after 28 days.

Experimental unit = jar.

An experiment on biodiversity

A, B, C, D, E, F — six types of freshwater "shrimp".
Put 12 shrimps in a jar containing stream water and alder leaf litter.
Measure how much leaf litter is eaten after 28 days.

Experimental unit = jar.

Treatment			Richness Level
A, ..., F	monoculture	12 of type A	1
AB, ..., EF	duoculture	6 of A, 6 of B	2
ABC, ..., DEF	triculture	4 of A, 4 of B, 4 of C	3

An experiment on biodiversity

A, B, C, D, E, F — six types of freshwater “shrimp”.
Put 12 shrimps in a jar containing stream water and alder leaf litter.
Measure how much leaf litter is eaten after 28 days.

Experimental unit = jar.

	Treatment			Richness Level
6	A, ..., F	monoculture	12 of type A	1
15	AB, ..., EF	duoculture	6 of A, 6 of B	2
20	ABC, ..., DEF	triculture	4 of A, 4 of B, 4 of C	3
41				

An experiment on biodiversity

A, B, C, D, E, F — six types of freshwater "shrimp".
Put 12 shrimps in a jar containing stream water and alder leaf litter.
Measure how much leaf litter is eaten after 28 days.

Experimental unit = jar.

	Treatment			Richness Level
6	A, …, F	monoculture	12 of type A	1
15	AB, …, EF	duoculture	6 of A, 6 of B	2
20	ABC, …, DEF	triculture	4 of A, 4 of B, 4 of C	3
41				

The experiment was carried out in 4 blocks of 41 jars.

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a “bar-and-antenna” for each of the 41 treatments.

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a “bar-and-antenna” for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a "bar-and-antenna" for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;
- explained that 41 treatment means were almost useless without some reference to the treatment structure and the models we were comparing;

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a "bar-and-antenna" for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;
- explained that 41 treatment means were almost useless without some reference to the treatment structure and the models we were comparing;
- explained that the "error bars" were misleading, because they ignored the fact that the experiment had been conducted in blocks;

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a “bar-and-antenna” for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;
- explained that 41 treatment means were almost useless without some reference to the treatment structure and the models we were comparing;
- explained that the “error bars” were misleading, because they ignored the fact that the experiment had been conducted in blocks;
- suggested that it would be better for us to make the complete data set available in a web appendix.

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a "bar-and-antenna" for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;
- explained that 41 treatment means were almost useless without some reference to the treatment structure and the models we were comparing;
- explained that the "error bars" were misleading, because they ignored the fact that the experiment had been conducted in blocks;
- suggested that it would be better for us to make the complete data set available in a web appendix.

Dealing with a referee for an ecology journal

We submitted a paper to an ecology journal. The referee demanded that we include a diagram giving a "bar-and-antenna" for each of the 41 treatments.

In our response, we

- included such a diagram, to show how uninformative it was;
- explained that 41 treatment means were almost useless without some reference to the treatment structure and the models we were comparing;
- explained that the "error bars" were misleading, because they ignored the fact that the experiment had been conducted in blocks;
- suggested that it would be better for us to make the complete data set available in a web appendix.

The editor accepted our arguments.

What models did we fit?

The biologist fitted the model ‘Richness’ with 3 parameters,
one for each level of richness,
and found no evidence of any differences between the levels.

What models did we fit?

The biologist fitted the model ‘Richness’ with 3 parameters,
one for each level of richness,
and found no evidence of any differences between the levels.

I suggested the model ‘Type’ with 6 parameters $\alpha_A, \ldots, \alpha_F$:

monoculture A	α_A
duoculture AB	$\dfrac{\alpha_A + \alpha_B}{2}$
triculture ABC	$\dfrac{\alpha_A + \alpha_B + \alpha_C}{3}$

What models did we fit?

The biologist fitted the model 'Richness' with 3 parameters,
one for each level of richness,
and found no evidence of any differences between the levels.

I suggested the model 'Type' with 6 parameters $\alpha_A, \ldots, \alpha_F$:

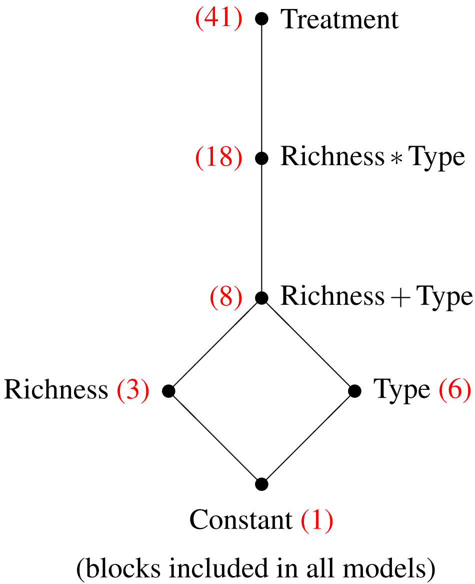
monoculture A	α_A
duoculture AB	$\dfrac{\alpha_A + \alpha_B}{2}$
triculture ABC	$\dfrac{\alpha_A + \alpha_B + \alpha_C}{3}$

In other words, if there are x_i shrimps of type i then

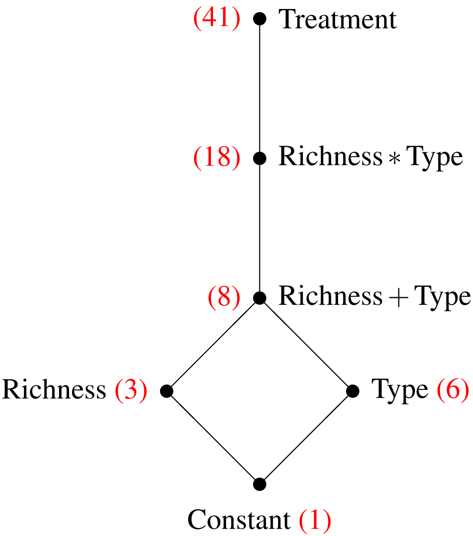
$$\mathbb{E}(Y) = \sum_{i=1}^{6} a_i x_i \qquad \text{where } 12a_i = \alpha_i.$$

($\sum x_i = 12$ always, so no need for intercept.)

The Hasse diagram of our family of expectation models



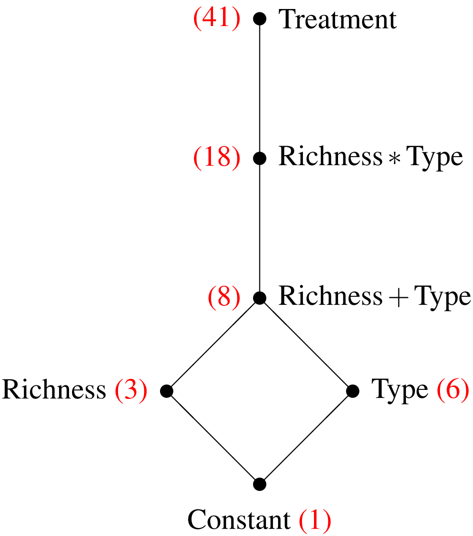
The Hasse diagram of our family of expectation models



(add a different constant for each level of richness)

(blocks included in all models)

The Hasse diagram of our family of expectation models

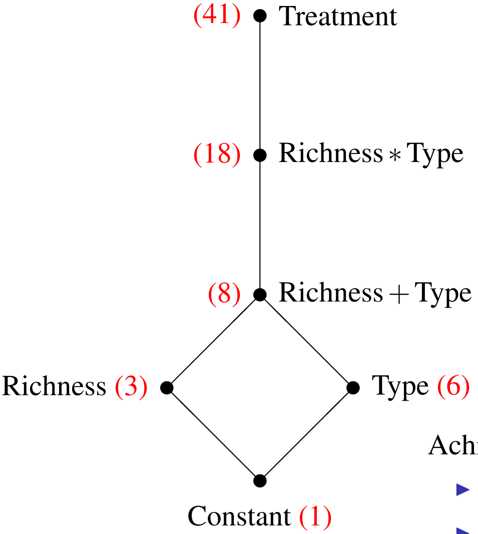


(a_i can change with each level of richness but does not depend on what else is present)

(add a different constant for each level of richness)

(blocks included in all models)

The Hasse diagram of our family of expectation models



(blocks included in all models)

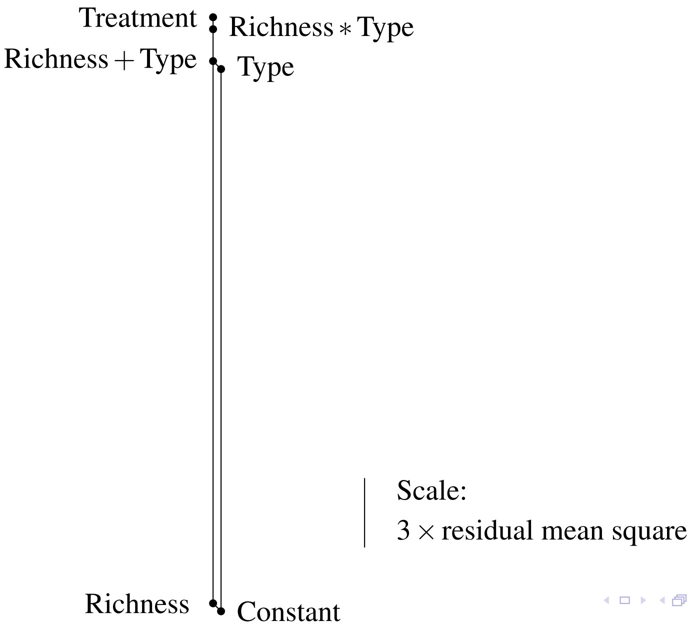
(a_i can change with each level of richness but does not depend on what else is present)

(add a different constant for each level of richness)

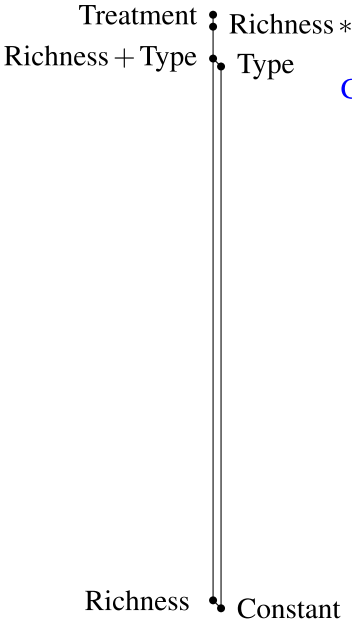
Achievement: an ecology journal published

- diagram of family of models
- statement that each row of an ANOVA table is for a difference between models.

What the data showed: lengths are mean squares



What the data showed: lengths are mean squares

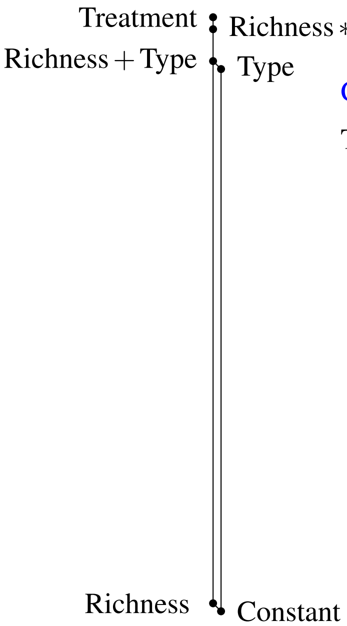


Conclusions:

Scale:

$3 \times$ residual mean square

What the data showed: lengths are mean squares



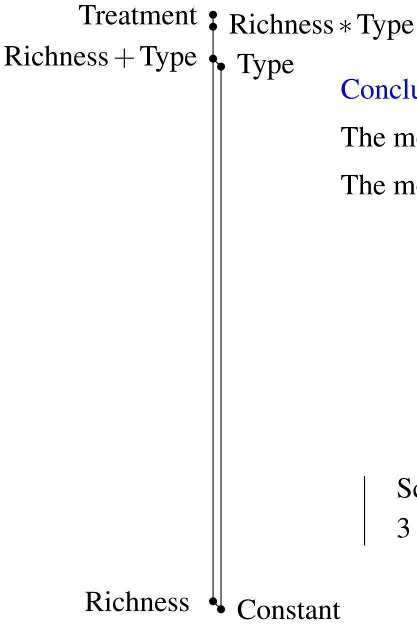
Conclusions:

The model Richness does not explain the data.

Scale:

$3 \times$ residual mean square

What the data showed: lengths are mean squares



Conclusions:

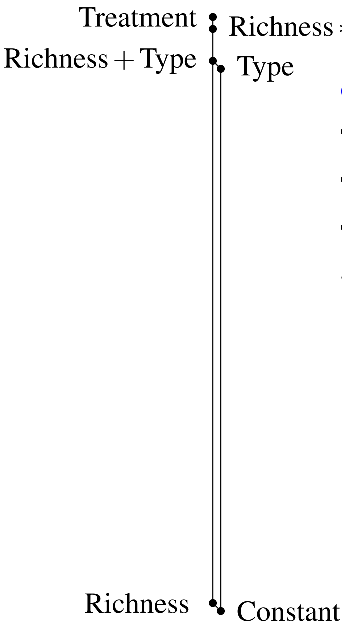
The model Richness does not explain the data.

The model Type explains the data well.

Scale:

$3 \times$ residual mean square

What the data showed: lengths are mean squares



Conclusions:

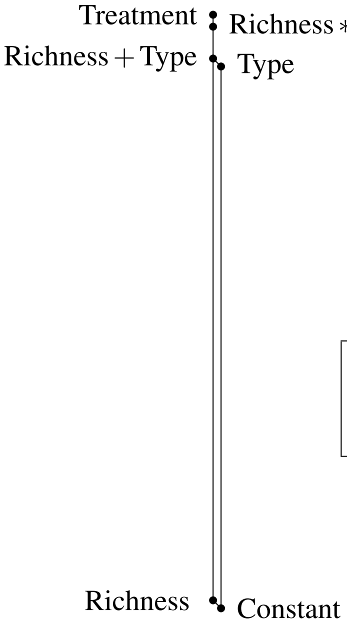
The model Richness does not explain the data.

The model Type explains the data well.

There is no evidence that any larger model does any better.

Scale:
$3 \times$ residual mean square

What the data showed: lengths are mean squares



Is such a scaled Hasse diagram a good way of displaying the ANOVA table when there is only one relevant residual mean square?

Scale:

$3 \times$ residual mean square