STA305/1004-Class 17

March 16, 2017

Today's Class

- Factorial designs at two levels
- Cube plots
- Calculation of factorial effects

- HW #3 extension until Monday, March 20 by 22:00
- TA office hours on Thursday, Friday see portal for exact times.

Difference between ANOVA and Factorial Designs

In ANOVA the objective is to compare the individual experimental conditions with each other. In a factorial experiment the objective is generally to compare combinations of experimental conditions.

Let's consider the food diary study above. What is the effect of keeping a food diary?

$2 \times 2 \times 2 = 2^3$

Expt condition	Keep food diary	Increase physical activity	Home visit	weight loss
1	No	No	No	y_1
2	No	No	Yes	y_2
3	No	Yes	No	y_3
4	No	Yes	Yes	y_4
5	Yes	No	No	y_5
6	Yes	No	Yes	y_6
7	Yes	Yes	No	y_7
8	Yes	Yes	Yes	y_8

We can estimate the effect of food diary by comparing the mean of all conditions where food diary is set to NO (conditions 1-4) and mean of all conditions where food diary set to YES (conditions 5-8). This is also called the **main effect** of food diary, the adjective *main* being a reminder that this average is taken over the levels of the other factors.

Difference between ANOVA and Factorial Designs

Expt condition	Keep food diary	Increase physical activity	Home visit	weight loss
1	No	No	No	y_1
2	No	No	Yes	y_2
3	No	Yes	No	y_3
4	No	Yes	Yes	y_4
5	Yes	No	No	y_5
6	Yes	No	Yes	y_6
7	Yes	Yes	No	y_7
8	Yes	Yes	Yes	y_8

The main effect of food diary is: $= \text{food diary} = \text{No}$ $= \text{food diary} = \text{Yes}$.

$$\frac{y_1 + y_2 + y_3 + y_4}{4} - \frac{y_5 + y_6 + y_7 + y_8}{4}.$$

The main effect of physical activity is: $= \text{phys activity} = \text{No}$ $= \text{phy activity} = \text{yes}$.

$$\frac{y_1 + y_2 + y_5 + y_6}{4} - \frac{y_3 + y_4 + y_7 + y_8}{4}.$$

The main effect of home visit is:

$$\frac{y_1 + y_3 + y_5 + y_7}{4} - \frac{y_2 + y_4 + y_6 + y_8}{4}.$$

Question

A chemical reaction experiment was carried out with the objective of comparing if a new catalyst B would give higher yields than the old catalyst A. The experiment was run on six different batches of raw material which were known to be quite different from one another.

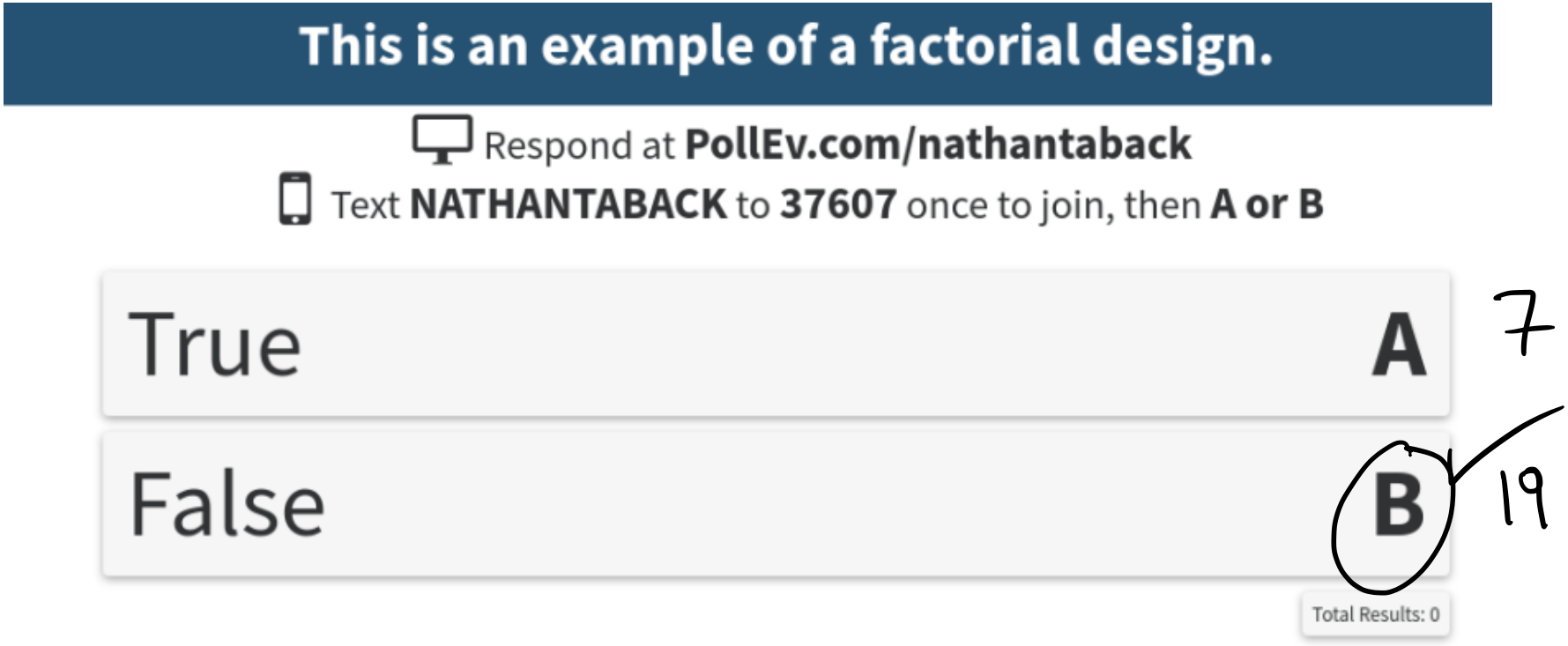


Figure 1:

Catalyst	Batch 1	2	3	4	5	6
A						
B						

Expt. Cond	Catalyst	Batch	Yield
1	A	1	y_1
2	A	2	y_2
3	A	3	
4	A	4	
5	A	5	
6	A	6	
7	B	1	
8	B	2	
9	B	3	
10	B	4	
11	B	5	
12	B	6	y_{12}

$2^1 \times 6^1$ factorial.

Factorial designs at two levels

To perform a factorial design:

1. Select a fixed number of levels of each factor.
2. Run experiments in all possible combinations.

Pilot plant investigation - example of factorial design

A pilot plant invsetiagtion employed a 2^3 factorial design (Box, Hunter, and Hunter (2005)) with

Factors	level 1	level 2
Temperature	160C°(-1)	180C°(+1)
Concentration	20% (-1)	40% (+1)
Catalyst	A (-1)	B(+1)

run	T	C	K	y
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

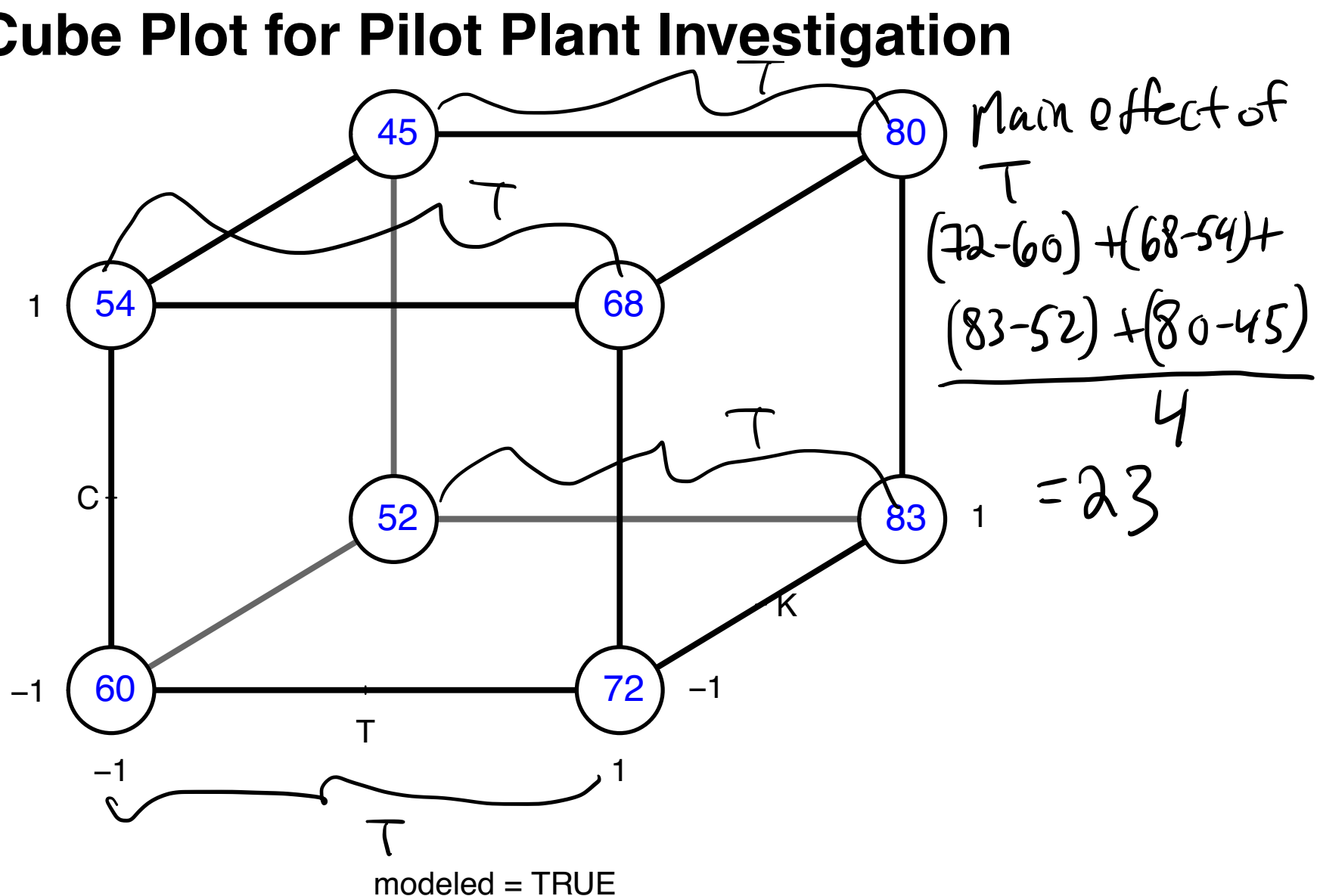
Main effect of T

$$= \frac{(72+68+83+80) - (60+54+52+45)}{4}$$

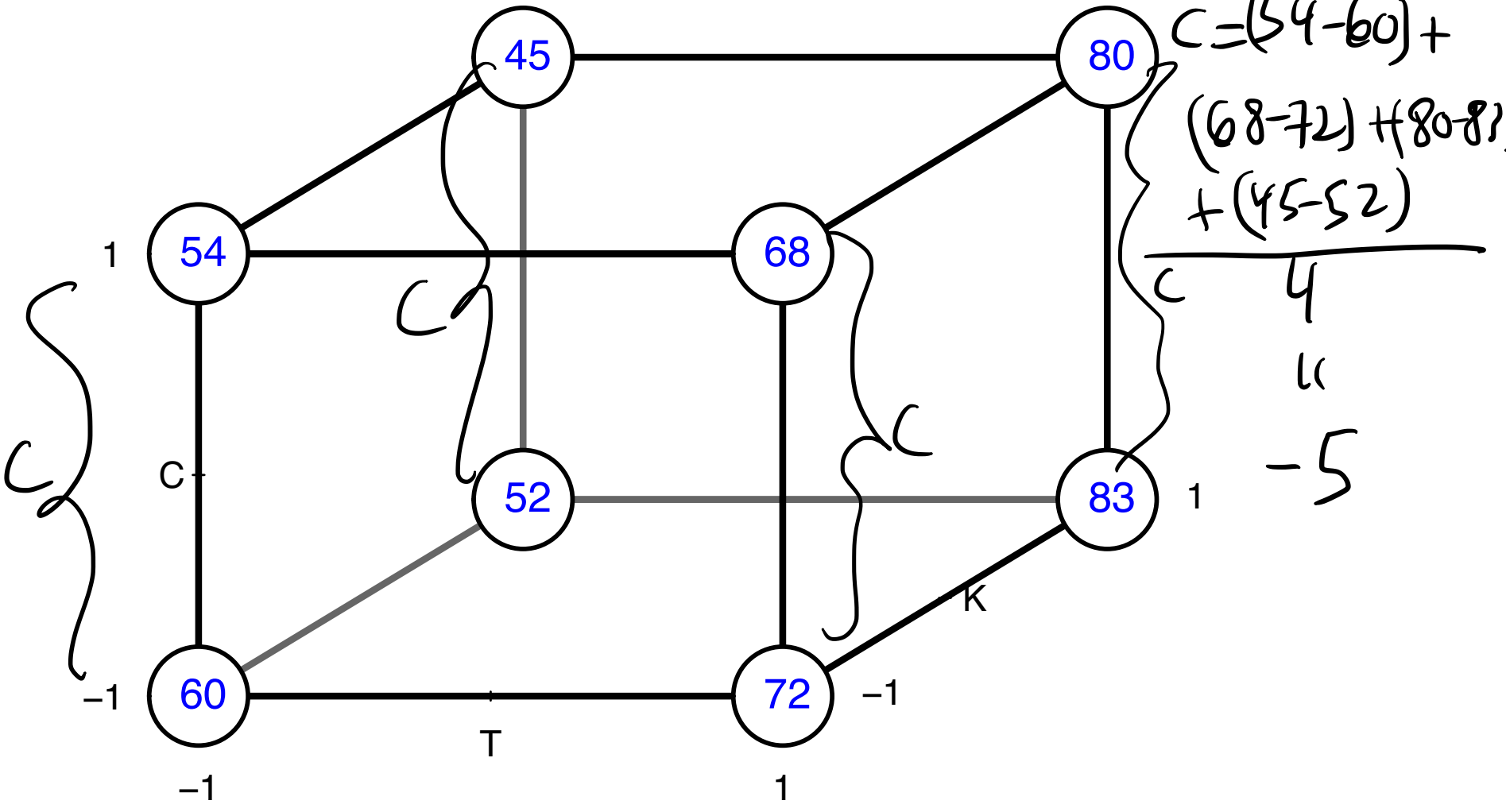
- Each data value recorded is for the response yield y averaged over two duplicate runs.

Cube plots

```
library("FrF2")
bhh54 <- lm(y~T*C*K,data=tab0502)
cubePlot(bhh54,"T","K","C",main="Cube Plot for Pilot Plant Investigation")
```



Cube plots

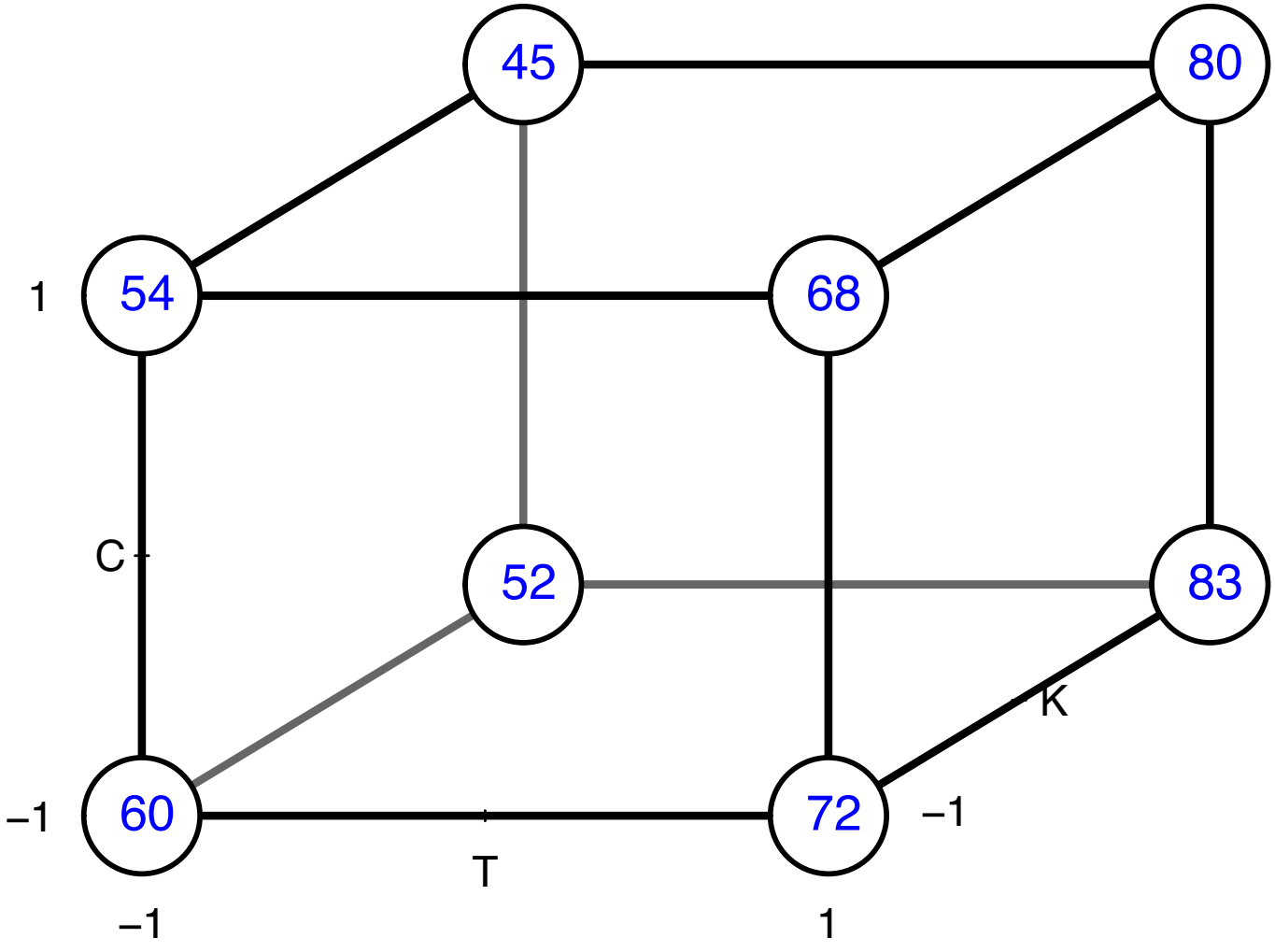


Main effect of C = (54−60) + (68−72) + (80−81) + (45−52) / 4

= −5

modeled = TRUE

Cube plots

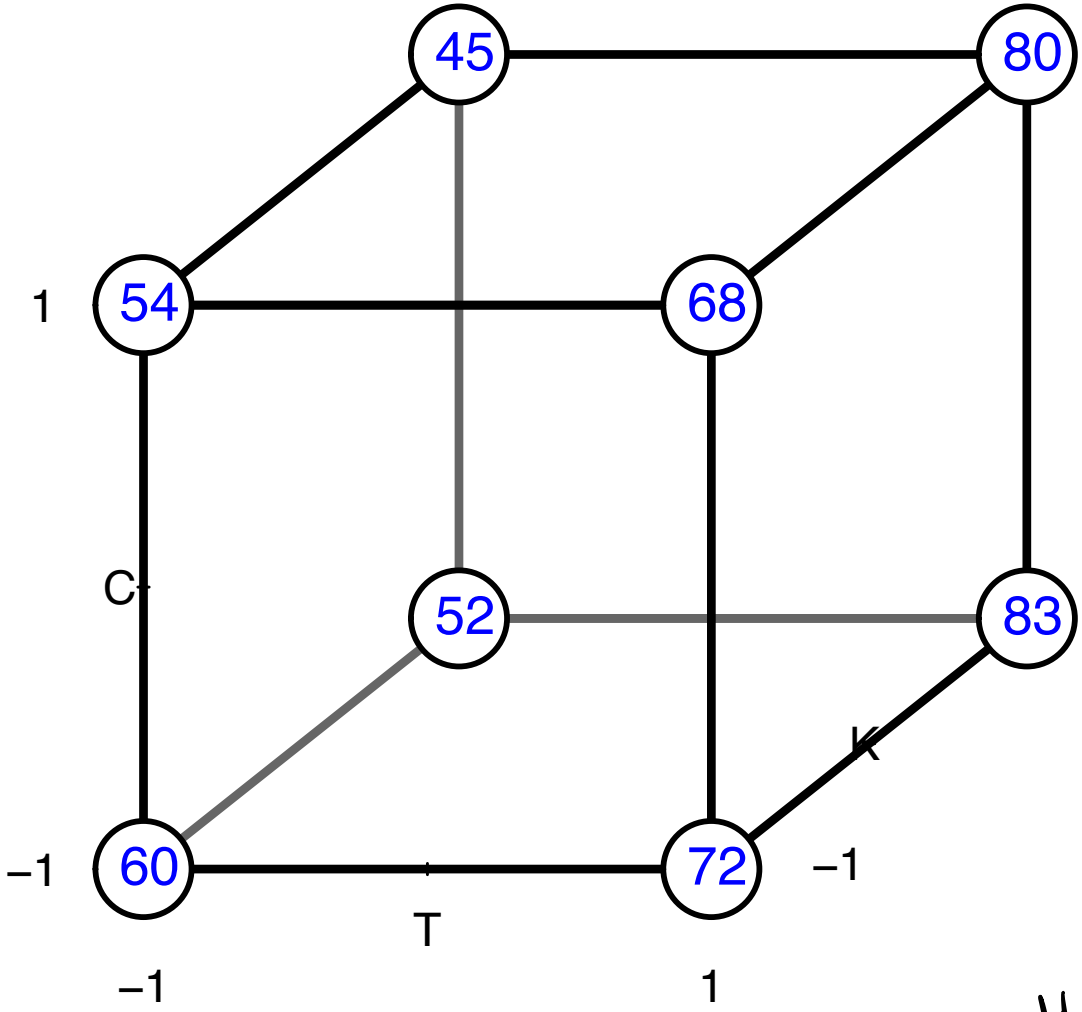


$K = \frac{(83-72)+(80-68)+(45-54)+(52-60)}{4}$

$=$

modeled = TRUE

Cube plots



modeled = TRUE

When $K=-1$

Main effect of T is:

$$\frac{(72-60)+(68-54)}{2}$$

When $K=+1$

Main effect of T is

$$\frac{(83-52)+(80-45)}{2}$$

ave.

then the difference between these two values is the interaction between TK.

Cube plots

- 8 run design produces 12 comparisons
- Each edge of cube only one factor changed while other 2 held constant.
- Therefore experimenter that believes in only changing one factor at a time is satisfied.

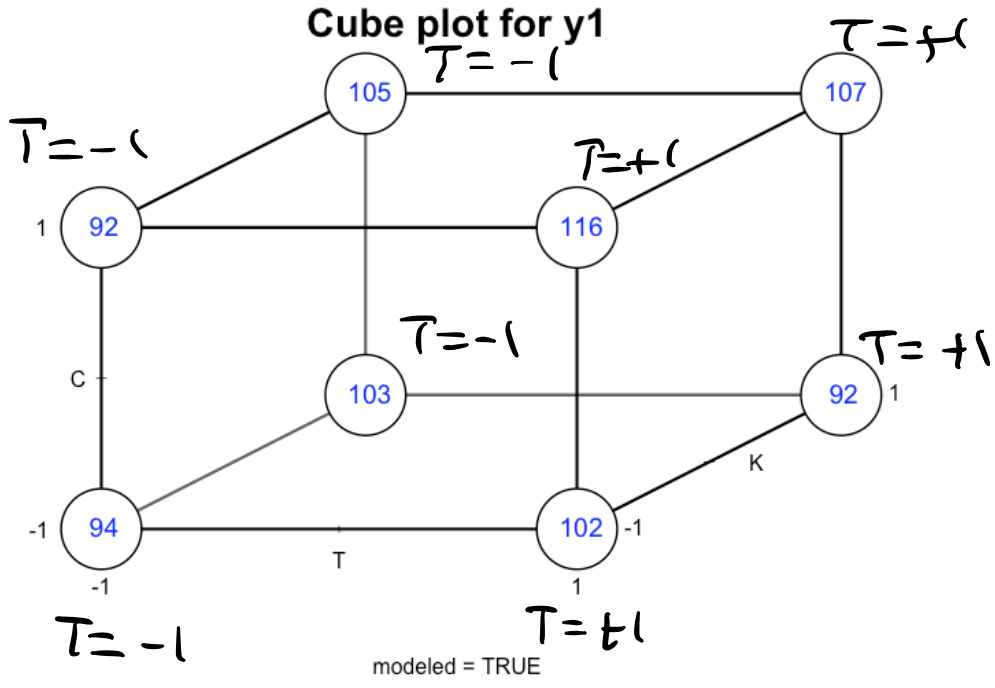
Cube plots

Using the cube plot below the main effects for T, C, K (respectively) are approximately:

Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A, B, or C**

T=2.88; C=3.63; K=0.38	**A**
T=5.75; C=7.25; K=0.75	**B** ✓
T=11.5; C=14.5; K=1.5	**C**



Why?

$$T = \frac{(102-94)+(116-92)+(107-105)+(92-103)}{4} = 5.75$$

Figure 2:

Interaction effects - two factor interactions

run	T	C	K	y
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

When K=-1

When K=+1

We would get the same value for the two-factor interaction if we hold T constant then do a similar calc. for K.

when K=-1

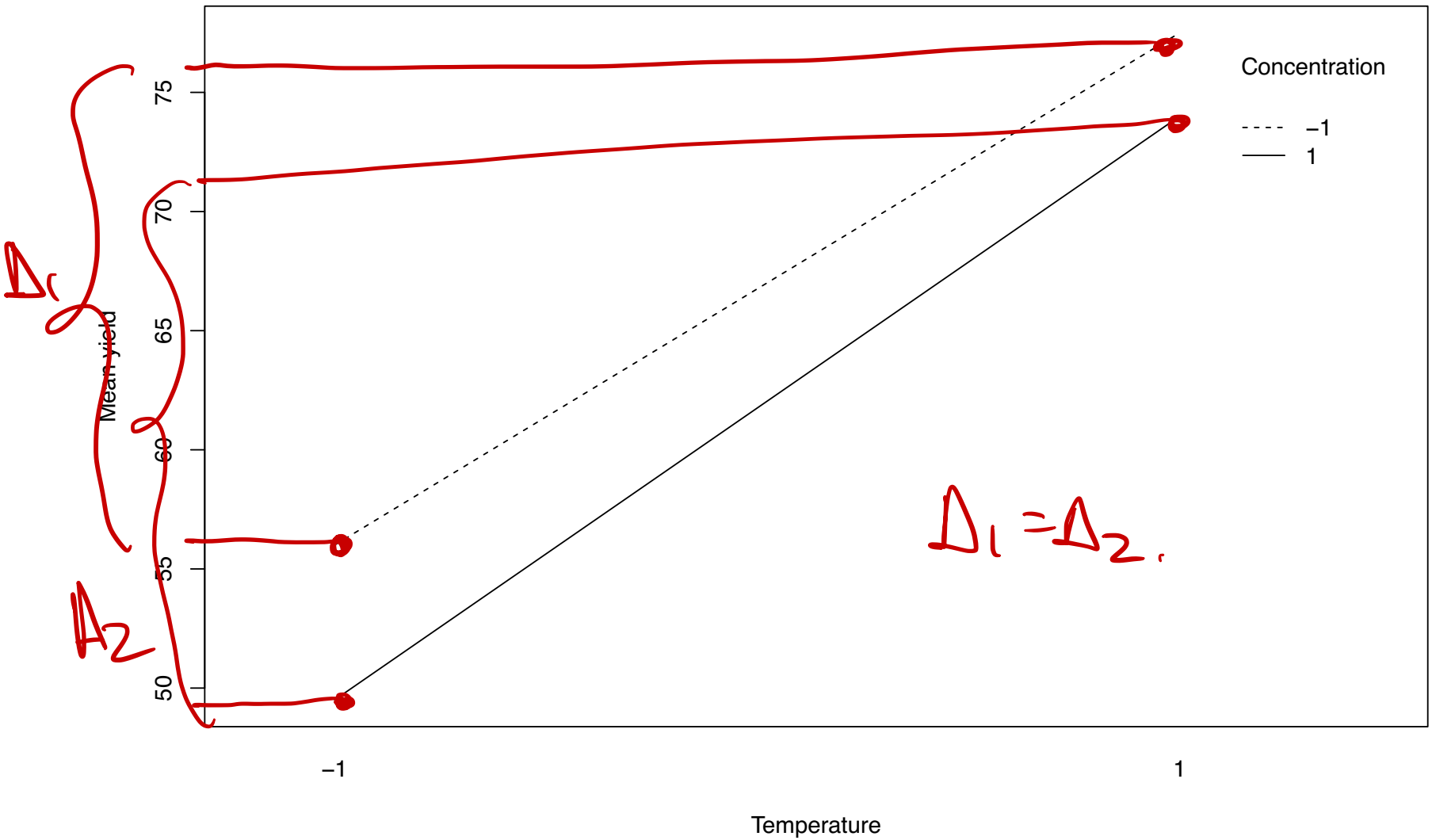
- When the catalyst K is A the temperature effect is: $\frac{68+72}{2} - \frac{60+54}{2} = 70 - 57 = 13$.
- When the catalyst K is B the temperature effect is: $\frac{83+80}{2} - \frac{52+45}{2} = 81.5 - 48.5 = 33$.
- The average difference between these two average differences is called the **interaction** between temperature and catalyst denoted by TK. This is the interaction between the two factors temperature and catalyst - the two factor interaction between temperature and catalyst.

$$TK = \frac{33 - 13}{2} = 10$$

.

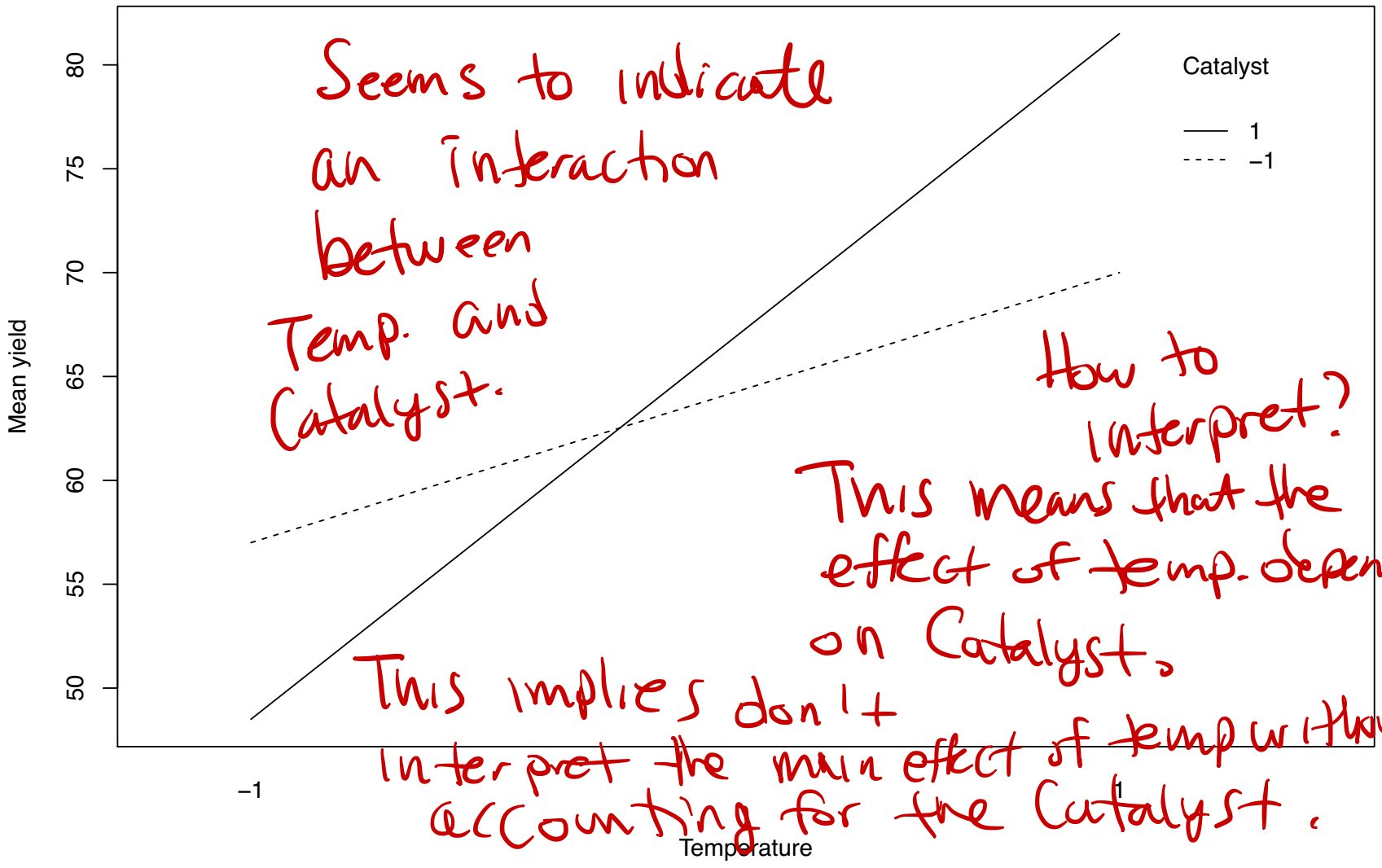
Interaction plots - Concentration by temperature

```
interaction.plot(tab0502$T,tab0502$C,tab0502$y, type="l",
                 xlab="Temperature",trace.label="Concentration",
                 ylab="Mean yield")
```



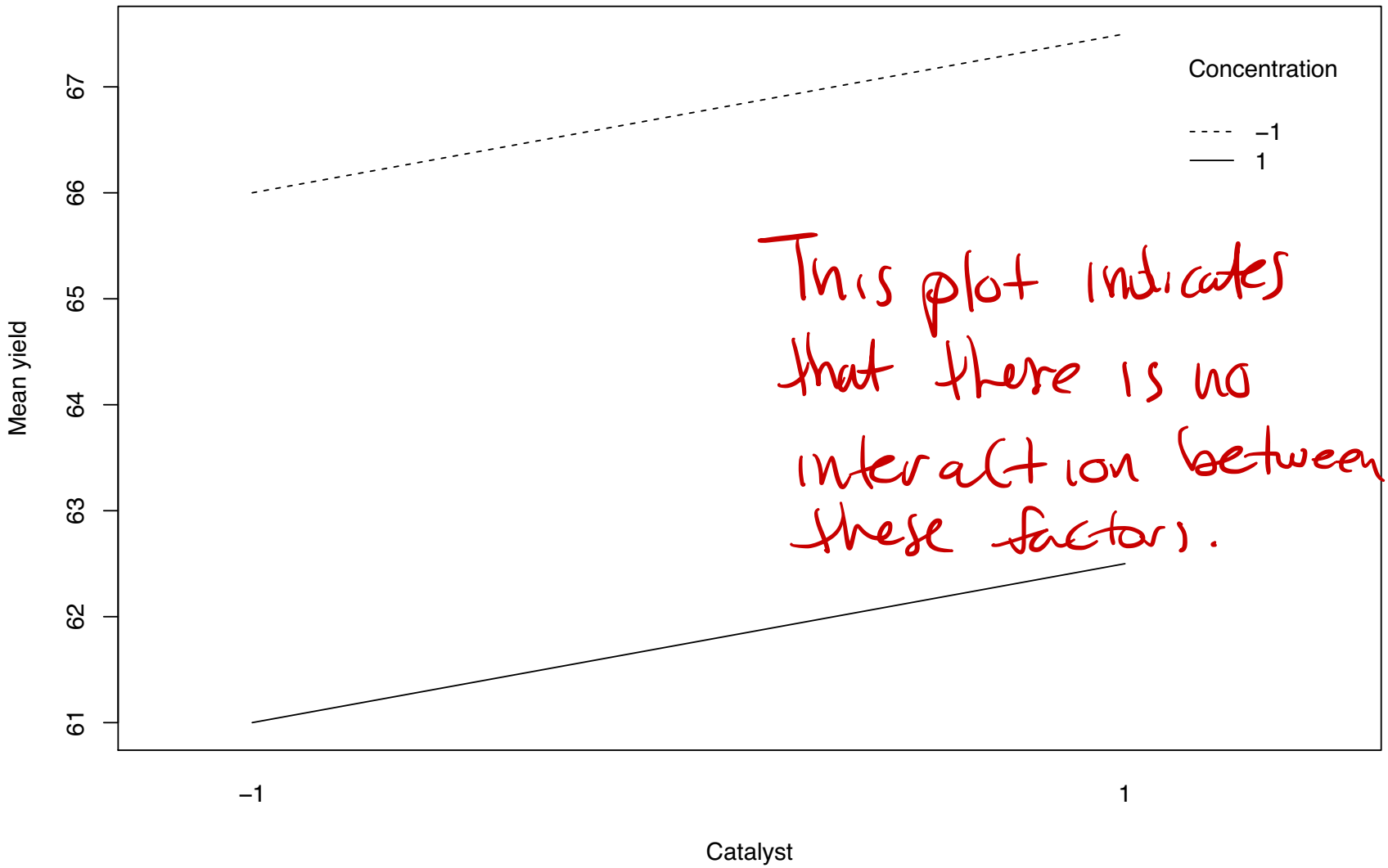
Interaction plots - Temperature by catalyst

```
interaction.plot(tab0502$T,tab0502$K,tab0502$y, type="l",
                 xlab="Temperature",trace.label="Catalyst",
                 ylab="Mean yield")
```



Interaction plots - Concentration by catalyst

```
interaction.plot(tab0502$K,tab0502$C,tab0502$y, type="l",
                 xlab="Catalyst",trace.label="Concentration",
                 ylab="Mean yield")
```



Three factor interactions

run	T	C	K	y
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

The temperature by concentration interaction when the catalyst is B (at it's $+1$ level) is:

$$\text{Interaction TC} = \frac{(y_8 - y_7) - (y_6 - y_5)}{2} = \frac{(80 - 45) - (83 - 52)}{2} = 2.$$

The temperature by concentration interaction when the catalyst is A (at it's -1 level) is:

$$\text{Interaction TC} = \frac{(y_4 - y_3) - (y_2 - y_1)}{2} = \frac{(68 - 54) - (72 - 60)}{2} = 1.$$

$$\text{TCK} = \frac{2 - 1}{2} = \frac{1}{2}.$$

Three factor interaction

- Interactions are symmetric in all factors.
- It could have been defined as half the difference between the temperature-by-catalyst interactions at each of the two concentrations.
- Mostly rely on statistical software such as R.

Replicate runs

- Each of the 8 responses in the table is the average of two (genuinely) replicated runs.
- Genuinely replicated run means that variation between runs made at same experimental conditions is a reflection of the total run-to-run variability.

run	T	C	K	y
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

Replicate runs

- Randomization of the run order for all 16 runs ensures the replication is genuine.
- `run1` is order of the first run and `run2` is order of the second run.

run1	run2	T	C	K	y1	y2	diff
6	13	-1	-1	-1	59	61	-2
2	4	1	-1	-1	74	70	4
1	16	-1	1	-1	50	58	-8
5	10	1	1	-1	69	67	2
8	12	-1	-1	1	50	54	-4
9	14	1	-1	1	81	85	-4
3	11	-1	1	1	46	44	2
7	15	1	1	1	79	81	-2

Replicate runs

- Replication not always feasible or easy.
- For the pilot plant experiment a run involved: cleaning the reactor; inserting the appropriate catalyst charge; and running the apparatus at a given concentration for 3 hours, and sampling output every 15 minutes.
- A genuine run involved taking all of these steps all over again!

Replicate runs

- There are usually better ways to employ 16 independent runs than by fully replicating a 2^3 factorial.
- Other designs can study four or five factors with a 16 run two-level design.

Estimate of error variance of the effects from replicated runs

run1	run2	T	C	K	y1	y2	diff
6	13	-1	-1	-1	59	61	-2
2	4	1	-1	-1	74	70	4
1	16	-1	1	-1	50	58	-8
5	10	1	1	-1	69	67	2
8	12	-1	-1	1	50	54	-4
9	14	1	-1	1	81	85	-4
3	11	-1	1	1	46	44	2
7	15	1	1	1	79	81	-2

$$s_i^2 = \frac{(y_{i1} - y_{i2})^2}{2},$$

to Show use formula for Sample Variance.

- y_{i1} is the first outcome from *ith* run.
- $\text{diff}_i = (y_{i1} - y_{i2})$.
- A pooled estimate of σ^2 is

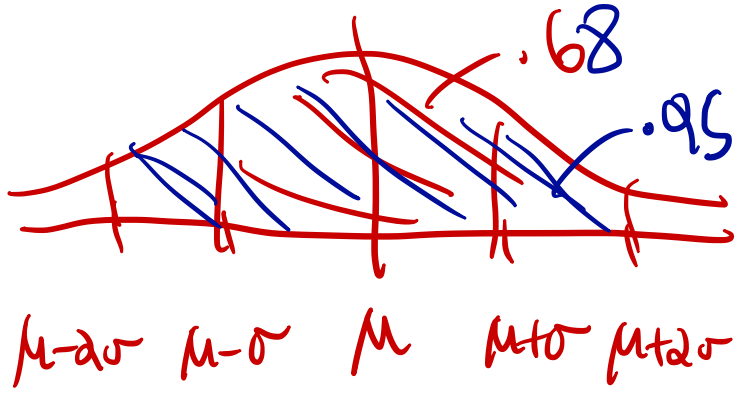
$$s^2 = \frac{\sum_{i=1}^{8} s_i^2}{8} = \frac{64}{8} = 8.$$

- The variance of an effect is:

$$Var(\text{effect}) = \left(\frac{1}{8} + \frac{1}{8}\right) s^2 = 8/4 = 2$$

Interpretation of results

68-95-99.7 for any normal dist.



- Which effects are real and which can be explained by chance?
- A rough rule of thumb: any effect that is 2-3 times their standard error are not easily explained by chance alone.

factorial effects are differences of means. If the sample size is large enough then the factorial effects will have a normal distn.

Interpretation of results

- Assume that the observations are independent and normally distributed then

$$\text{effect}/se\,(\text{effect}) \sim t_8.$$

- A 95% confidence interval can be calculated as:

$$\text{effect} \pm t_{8,.05/2} \times se\,(\text{effect})\,.$$

where $t_{8,.05/2}$ is the 97.5th percentile of the t_8. This is obtained in R via the `qt()` function.

```
qt(p = 1-.025,df = 8)
```

```
## [1] 2.306004
```

- In the pilot plant study

$$\text{effect} \pm 2.3 \times 1.4 = \text{effect} \pm 3.2.$$

Interpretation of results

- The main effect of a factor should be individually interpreted only if there is no evidence that the factor interacts with other factors.
- Which effects should be considered jointly and which independently?

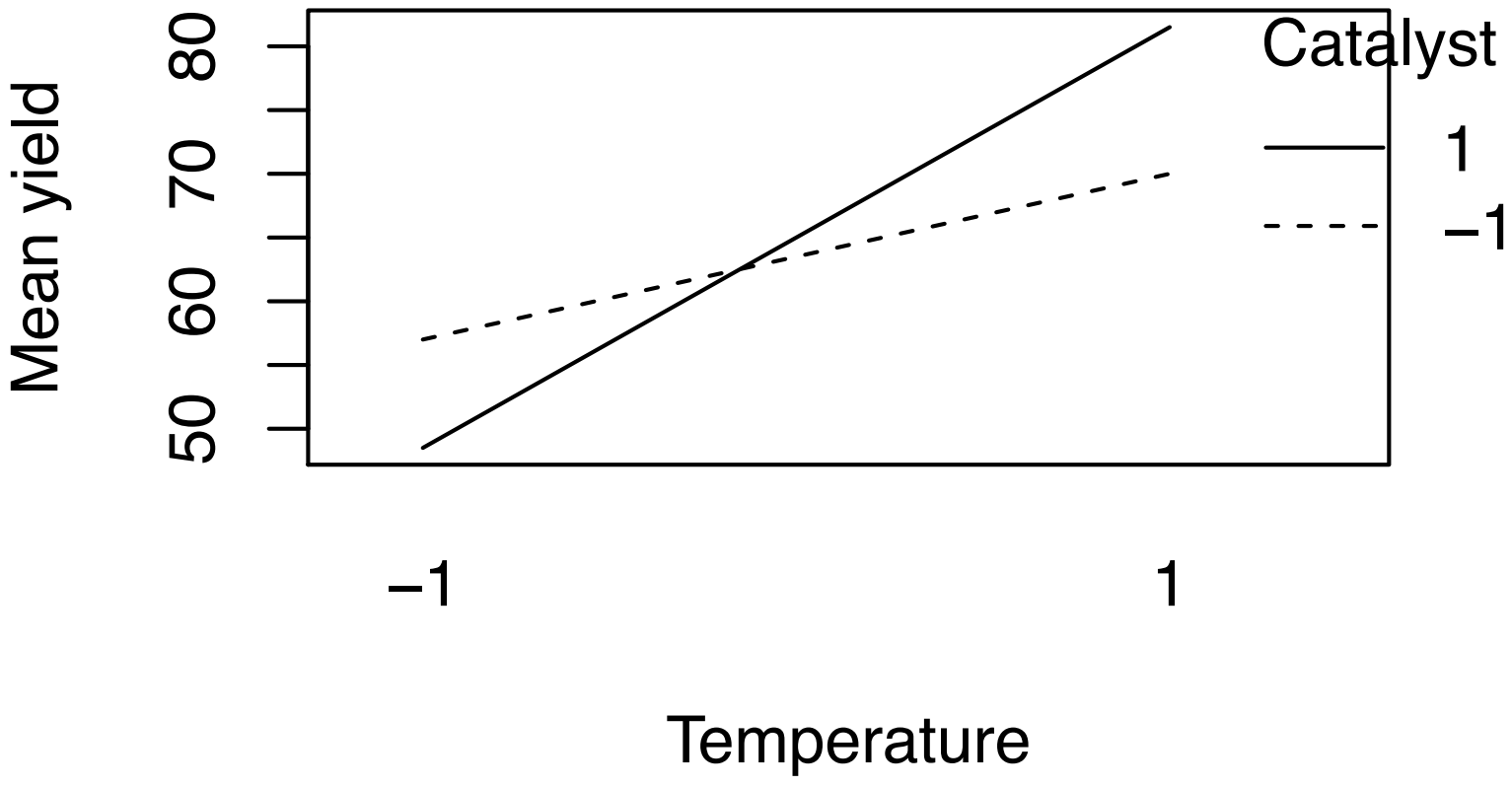
Effects	95% Confidence Interval
T	(19.8, 26.2)
C	(-8.2, -1.8)
K	(-1.7, 4.7)
TC	(-1.7, 4.7)
TK	(6.8, 13.2)
CK	(-3.2, 3.2)
TCK	(-2.7, 3.7)

The Significant (at $\alpha = .05$) effects are 95% CI that do not contain 0.

Interpretation of results

(main effect)

- The effect of changing concentration over the ranges studied is to reduce yield by about 5 units. This is irrespective of the tested level of other variables.
- The effects of temperature and catalyst cannot be interpreted separately because of the large TK interaction. With catalyst A the temperature effect is 13 units and with catalyst B it is 33 units.



Linear model for factorial design

Let y_i be the yield from the i^{th} run,

$$x_{i1} = \begin{cases} +1 & \text{if } T = 180 \\ -1 & \text{if } T = 160 \end{cases}$$

$$x_{i2} = \begin{cases} +1 & \text{if } C = 40 \\ -1 & \text{if } C = 20 \end{cases}$$

$$x_{i3} = \begin{cases} +1 & \text{if } K = B \\ -1 & \text{if } K = A \end{cases}$$

A linear model for a 2^3 factorial design is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \epsilon_i.$$

The variables $x_{i1}x_{i2}$ is the interaction between temperature and concentration, $x_{i1}x_{i3}$ is the interaction between temperature and catalyst, etc.

Linear model for factorial design

The table of contrasts for a 2^3 design is the design matrix X from the linear model above.

Mean	T	K	C	T:K	T:C	K:C	T:K:C	yield average
1	-1	-1	-1	1	1	1	-1	60
1	1	-1	-1	-1	-1	1	1	72
1	-1	-1	1	1	-1	-1	1	54
1	1	-1	1	-1	1	-1	-1	68
1	-1	1	-1	-1	1	-1	1	52
1	1	1	-1	1	-1	-1	-1	83
1	-1	1	1	-1	-1	1	-1	45
1	1	1	1	1	1	1	1	80

$$TMC = (72+54+52+80)/4 - (60+68+83+45)/4$$

- All factorial effects can be calculated from this table.
- Signs for interaction contrasts obtained by multiplying signs of their respective factors.
- Each column perfectly balanced with respect to other columns.
- Balanced (orthogonal) design ensures each estimated effect is unaffected by magnitude and signs of other effects.
- Table of signs obtained similarly for any 2^k factorial design.

Linear model for factorial design

$2^4 = 16$ Runs

What is the table of contrasts for a 2^4 factorial design?

	main				two-way						three way				four way
Run	1	2	3	4	12	13	14	23	24	34	123	124	134	234	1234
1	−	−	−	−	+										
2	+	−	−	−	−										
3	−	+	−	−	−										
4	+	+	−	−	+										
5	−	−	+	−	+										
6	+	−	+	−	−										
7	−	+	+	−	−										
8	+	+	+	−	+										
9	−	−	−	+	+										
10	+	−	−	+	−										
11	−	+	−	+	−										
12	+	+	−	+	+										
13	−	−	+	+	+										
14	+	−	+	+	−										
15	−	+	+	+	−										
16	+	+	+	+	+										

Linear model for factorial design - calculating factorial effects from parameter estimates

The parameter estimates are obtained via the `lm()` function in R.

- Estimated least squares coefficients are one-half the factorial estimates.
- Therefore, the factorial estimates are twice the least squares coefficients.

$$\hat{\beta}_1 = 11.50 \Rightarrow T = 2 \times 11.50 = 23.26$$

$$\hat{\beta}_2 = 0.75 \Rightarrow K = 2 \times 0.75 = 1.5$$

$$\hat{\beta}_4 = 5.00 \Rightarrow TK = 2 \times 5.00 = 10.00$$

$\hat{\beta}_i$ = factorial / 2

∴ multiply Least Squares estimate by 2 to get factorial effects.

```
fact.mod <-lm(y~T*K*C,data=tab0502)
round(summary(fact.mod)$coefficients,2)
```

only one obs. per expt. run.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.25        NaN     NaN      NaN
T               11.50        NaN     NaN      NaN
K                0.75        NaN     NaN      NaN
C               -2.50        NaN     NaN      NaN
T:K              5.00        NaN     NaN      NaN
T:C              0.75        NaN     NaN      NaN
K:C              0.00        NaN     NaN      NaN
T:K:C            0.25        NaN     NaN      NaN
```

Linear model for factorial design - significance testing

- When there are replicated runs we also obtain p-values and confidence intervals for the factorial effects from the regression model.
- For example, the p-value for β_1 corresponds to the factorial effect for temperature

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

If the null hypothesis is true then $\beta_1 = 0 \Rightarrow T = 0 \Rightarrow \mu_{T+} - \mu_{T-} = 0 \Rightarrow \mu_{T+} = \mu_{T-}$.

- μ_{T+} is the mean yield when the temperature is set at 180° and μ_{T-} is the mean yield when the temperature is set to 160°.

Linear model for factorial design - significance testing

To obtain 95% confidence intervals for the factorial effects we multiply the 95% confidence intervals for the regression parameters by 2. This is easily done in R using the function `confint.lm()`.

```
fact.mod <-lm(y~T*K*C,data=tab0503)
round(2*confint.lm(fact.mod),2)
```

```
             2.5 % 97.5 %
(Intercept) 125.24 131.76
T            19.74  26.26
K            -1.76   4.76
C            -8.26  -1.74
T:K           6.74  13.26
T:C          -1.76   4.76
K:C          -3.26   3.26
T:K:C        -2.76   3.76
```

this regression model used the duplicate runs so that an estimate of the Standard error Can be Calculated

Advantages of factorial designs over one-factor-at-a-time designs

- Suppose that one factor at a time was investigated. For example, temperature is investigated while holding concentration at 20% (-1) and catalyst at B (+1).
- In order for the effect to have more general relevance it would be necessary for the effect to be the same at all the other levels of concentration and catalyst.
- In other words there is no interaction between factors (e.g., temperature and catalyst).
- If the effect is the same then a factorial design is more efficient since the estimates of the effects require fewer observations to achieve the same precision.
- If the effect is different at other levels of concentration and catalyst then the factorial can detect and estimate interactions.

Suppose two factors A (levels a_1, a_2), B (levels b_1, b_2).

One factor at a time approach.

Fix level of B at b_1 then Compare levels of A.

Obs	Trt #1 $= a_1b_1$	$a_2b_1 =$ Trt #2
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
	$\bar{x}$	$\bar{y}$

Suppose $Var(x_i) = Var(y_i) = \sigma^2$

$$Var(\bar{x} - \bar{y}) = Var\left(\left(\frac{x_1+x_2+x_3+x_4}{4}\right) - \left(\frac{y_1+y_2+y_3+y_4}{4}\right)\right)$$

$$= \frac{1}{16}\left[Var(x_1+x_2+x_3+x_4) + Var(y_1+y_2+y_3+y_4)\right]$$

$$= \frac{1}{16} \cdot \left(4\sigma^2 + 4\sigma^2\right)$$

$$= \sigma^2/2.$$

Now if a_1 is better than a_2 then do the expt again and fix the level of a_1 then Compare b_1 to b_2 for $A = a_1$. Use 4 observations for each treatment.

∴ two Single factor expts. use 16 observations.

Replicated factorial design

A	B	Run 1	Run 2	
a_1	b_1	x_{11}	x_{12}	$\bar{x}_1$
a_2	b_1	x_{21}	x_{22}	$\bar{x}_2$
a_1	b_2	x_{31}	x_{32}	$\bar{x}_3$
a_2	b_2	x_{41}	x_{42}	$\bar{x}_4$

$$\text{Var}\left(\frac{\bar{x}_1 + \bar{x}_3}{2} - \frac{\bar{x}_2 + \bar{x}_4}{2}\right)$$

$$= \frac{1}{4}\left[\text{Var}(\bar{x}_1 + \bar{x}_3) + \text{Var}(\bar{x}_2 + \bar{x}_4)\right] = \frac{\sigma^2}{2}$$

- σ^2 8 run factorial expt. has the same variance vs. 16 run in single factor approach.

- factorial expt. provides an estimate of the interaction not provided by one-at-a time approach.