ENGR 599 Homework 2

Adrian Henle

```julia
using DataFrames   , Distributions   , PlutoUI
```

Data

Table 3.1

```julia
begin
    # transcribed data from Table 3.1
    local table3_1_raw ="40 60 40 60
A A B B
57 92 55 66
61 88 53 70
59 90 54 68"

    # split by line
    local table3_1_data = split(table3_1_raw, "\n")

    table3_1 = DataFrame(
        :temperature => [parse(Float64, x) for x in split(table3_1_data[1])],
        :catalyst => split(table3_1_data[2]),
        :yield1 => [parse(Float64, x) for x in split(table3_1_data[3])],
        :yield2 => [parse(Float64, x) for x in split(table3_1_data[4])],
        :avg_yield => [parse(Float64, x) for x in split(table3_1_data[5])]
    )
end;
```

Table 3.2

```julia
table3_2 = Dict(
    :grand_average_μ => 67.75,
    :grand_average_σ => 0.9,
    :effect_T_μ => 22.5,
    :effect_T_σ => 1.8,
    :effect_C_μ => -13.5,
    :effect_C_σ => 1.8,
    :effect_TC_μ => -8.5,
    :effect_TC_σ => 1.8
);
```

Table 3.3

```
begin
    local factors = Dict(
        1 => Dict(:- => 40., :+ => 60.),
        2 => Dict(:- => "A", :+ => "B"),
        3 => Dict(:- => 1.0, :+ => 1.5)
    )

    local runs = DataFrame(
        :factor1 => [:-, :+, :-, :+, :-, :+, :-, :+],
        :factor2 => [:-, :-, :+, :+, :-, :-, :+, :+],
        :factor3 => [:-, :-, :-, :-, :+, :+, :+, :+],
        :yield1 => [56., 85., 49., 64., 65., 92., .57, .70],
        :yield2 => [52., 88., 47., 62., 61., 95., 60., 74.],
        :avg_yield => [54., 86.5, 48., 63., 63., 93.5, 58.5, 72.]
    )

    table3_3 = Dict(
        :factors => factors,
        :runs => runs
    )
end;
```

Table 3.8

```
table3_8 = Dict(
    :average => 67.188,
    1 => 22.875,
    2 => -14.125,
    3 => 8.875,
    4 => 0.875,
    12 => -8.625,
    13 => -0.625,
    14 => 0.875,
    23 => -0.625,
    24 => 0.875,
    34 => 0.375,
    123 => 0.875,
    124 => -0.125,
    134 => -0.625,
    234 => 0.375,
    1234 => 0.375
);
```

Exercises

Exercise 3.1

Think of an experiment, preferably in your area of research interest, with a qualitative response. Which factors would you like to examine, in order to determine their possible influences on the response? Which factors could be confounding? Which factors could contribute to noise—that is, random fluctuations — in the response?

Determining the effect of a specific gene on eye color (can't think of one from my AI research...)

The most likely factors would be the presence of specific allele combinations (blue/blue, blue/green, green/brown, etc.)

The presence of specific versions of other genes (i.e. ones that are different from the classic "eye color" gene) could be confounding variables. Epigenetic differences could contribute to noise.

Exercise 3.2

Besides the temperature and catalyst, at the levels described above, our chemist would like to study the effect of pressure at three levels: 1, 5 and 10 atm at the same time using a full factorial design. What is the minimum number of runs that must be carried out?

$$2^2 * 3 = 12$$

Exercise 3.3

We have calculated a measure of the interaction between the T and C factors using the difference in the temperature effects. It might be asked why we did this instead of calculating the difference between the catalyst effects at the two different temperature levels. Use algebraic arguments to show that these two measures are identical. Remember that, with the sign convention we adopted, the calculation that you should make is ((Catalyst effect at 60 1C)–(Catalyst effect at 40 1C)), and not the contrary.

$$T = \frac{\bar{y}_2 + \bar{y}_4}{2} - \frac{\bar{y}_1 + \bar{y}_3}{2}$$

$$C = \frac{\bar{y}_3 + \bar{y}_4}{2} - \frac{\bar{y}_1 + \bar{y}_2}{2}$$

$$\mathbf{TC} = \frac{(\bar{y}_4 - \bar{y}_3) - (\bar{y}_2 - \bar{y}_1)}{2} = \frac{\bar{y}_1 + \bar{y}_4 - \bar{y}_2 - \bar{y}_3}{2}$$

$$\mathbf{CT} = \frac{(\bar{y}_4 - \bar{y}_2) - (\bar{y}_3 - \bar{y}_1)}{2} = \frac{\bar{y}_1 + \bar{y}_4 - \bar{y}_2 - \bar{y}_3}{2}$$

$$\therefore \mathbf{TC} = \mathbf{CT}$$

Exercise 3.4

Demonstrate that for any pair of numerical values $s^2 = d^2/2$; where d is the difference between the two values. Use this result and show that for a set of n duplicate runs (i.e., each run repeated only once, as in Table 3.1) the pooled estimate of the variance is $s^2 = \sum \frac{d_i^2}{2n}$

	temperature	catalyst	yield1	yield2	avg_yield
1	40.0	"A"	57.0	61.0	59.0
2	60.0	"A"	92.0	88.0	90.0
3	40.0	"B"	55.0	53.0	54.0
4	60.0	"B"	66.0	70.0	68.0

- `table3_1`

$$s^2 = \frac{\sum_i (x_i - \bar{x}_i)^2}{n-1}$$

$$\bar{x} = \frac{1}{2}(x_1 + x_2)$$

$$n - 1 = 1$$

$$s^2 = \sum_i \left(x_i - \frac{1}{2}(x_1 + x_2)\right)^2 = \left(x_1 - \frac{1}{2}(x_1 + x_2)\right)^2 + \left(x_2 - \frac{1}{2}(x_1 + x_2)\right)^2$$

$$= \left(\frac{1}{2}(x_1 - x_2)\right)^2 + \left(\frac{1}{2}(x_1 - x_2)\right)^2$$

$$d^2 = (x_1 - x_2)^2$$

$$s^2 = \frac{d^2}{2}$$

Pooled estimate:

$$s^2 = \sum_i \frac{d_i^2}{2n}$$

Exercise 3.5

According to Table 3.2, the standard error of the grand average is half the standard error of the effects. Use Eq. (2.15) to show that this has to be true.

```
  Dict{Symbol, Float64}(
      :effect_C_σ ⇒ 1.8
      :grand_average_μ ⇒ 67.75
      :effect_T_σ ⇒ 1.8
      :effect_C_μ ⇒ -13.5
      :grand_average_σ ⇒ 0.9
      :effect_TC_σ ⇒ 1.8
      :effect_TC_μ ⇒ -8.5
      :effect_T_μ ⇒ 22.5
)
```

```
table3_2
```

$$\sigma_{\bar{x}}^2 = \sum_i \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

$$\sigma_e^2 = \sigma^2(\bar{y}_+) + \sigma^2(\bar{y}_-) = 4\frac{\sigma^2}{n}$$

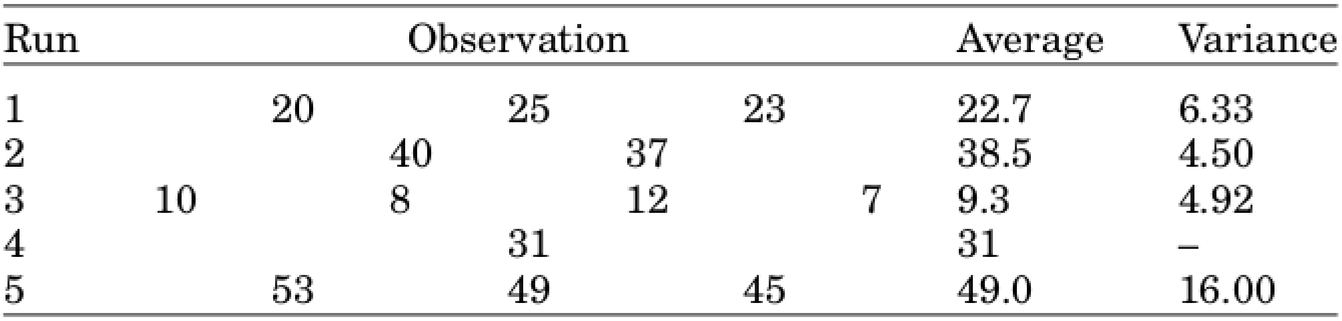
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_e = 2\frac{\sigma}{\sqrt{n}}$$

$$\therefore \sigma_e = 2\sigma_{\bar{y}}$$

Exercise 3.6

The observations listed below were recorded in genuine replicates of the different runs. Calculate a pooled estimate of the experimental error associated with these observations. How many degrees of freedom does this estimate have?



Run	Observation				Average	Variance
1	20	25	23		22.7	6.33
2	40	37			38.5	4.50
3	10	8	12	7	9.3	4.92
4	31				31	–
5	53	49	45		49.0	16.00

$$\sigma^2 = \frac{1}{R}\sum_r \sigma_r^2$$

where R is the number of different runs and r the specific observation.

$$\sigma^2 = \frac{3*6.33+2*4.5+4*4.92+3*16.00}{3+2+4+3} = 7.97$$

$$\sigma = 2.83$$

$$v_T = 12$$

Exercise 3.7

The data below were obtained in a study of the influence of two factors on the initial curing time of plaster of Paris, that is, the time when the plaster of Paris starts to harden after the powder is mixed with water (M.F. Pimentel and B.B. Neto, Anais do XXXI Congresso Brasileiro de Quimica, Recife, 1991). The runs were carried out in duplicate and in random order. Calculate all effects and their standard errors. Interpret the results.

$$\mathbf{F_1} = \frac{-\bar{y}_1+\bar{y}_2-\bar{y}_3+\bar{y}_4}{2} = -1.57$$

$$\mathbf{F_2} = \frac{-\bar{y}_1-\bar{y}_2+\bar{y}_3+\bar{y}_4}{2} = -2.45$$

$$\mathbf{F_{12}} = \mathbf{F_2} - \mathbf{F_1} = 0.88$$

Exercise 3.8

If we include the divisors, the matrix for calculating the effects becomes:

$$\mathbf{A} = \begin{bmatrix} +1/4 & +1/4 & +1/4 & +1/4 \\ -1/2 & +1/2 & -1/2 & +1/2 \\ -1/2 & -1/2 & +1/2 & +1/2 \\ +1/2 & -1/2 & -1/2 & +1/2 \end{bmatrix},$$

such that

$$\mathbf{Ay} = \begin{bmatrix} +1/4 & +1/4 & +1/4 & +1/4 \\ -1/2 & +1/2 & -1/2 & +1/2 \\ -1/2 & -1/2 & +1/2 & +1/2 \\ +1/2 & -1/2 & -1/2 & +1/2 \end{bmatrix} \times \begin{bmatrix} 59 \\ 90 \\ 54 \\ 68 \end{bmatrix} = \begin{bmatrix} 67.75 \\ 22.5 \\ -13.5 \\ -8.5 \end{bmatrix} = \mathbf{e},$$

where e *is a column vector containing the grand average and the original effects before they are divided by* 2. *Multiplying this vector on the left by the inverse of* A, *we get back our original observations, that is, the* y *vector* $A^{-1}e = A^{-1}Ay = I_4 y = y$ *where* I_4 *is the identity matrix of order 4. Determine* A^{-1} *(recall that the rows of the matrix of contrast coefficients are orthogonal and use your good sense; the calculation is in fact very simple) and show that the* $A^{-1}e$ *product is identical to the* Xb *product of Eq. (3.12). To understand why this is so, compare the* A^{-1} *and* X *matrices and the* b *and* e *vectors.*

```
begin
    local A = [
        .25 .25 .25 .25;
        -.5 .5 -.5 .5;
        -.5 -.5 .5 .5;
        .5 -.5 -.5 .5
    ]

    local e = [67.75, 22.5, -13.5, -8.5]

    A⁻¹ = inv(A)

    A⁻¹e = A⁻¹*e
end;
```

```
4×4 Matrix{Float64}:
 1.0  -0.5  -0.5   0.5
 1.0   0.5  -0.5  -0.5
 1.0  -0.5   0.5  -0.5
 1.0   0.5   0.5   0.5
```

```
A⁻¹
```

```
[59.0, 90.0, 54.0, 68.0]
```

```
A⁻¹e
```

Exercise 3.9

For each of the two levels of variable 3 there is a complete 22 factorial in variables 1 and 2. Calculate, from the values in Table 3.3, the 12 interaction at both levels of variable 3. Take the difference between these two values, divide by 2, and call the result the interaction of factor 3 with the 12 interaction. Repeat the whole process, starting from the values of the 23 interactions at the two levels of factor 1. You will then have the interaction of factor 1 with the 23 interaction. Compare the results obtained in these two calculations with the value of the 123 interaction given in the text.

	factor1	factor2	factor3	yield1	yield2	avg_yield
1	:-	:-	:-	56.0	52.0	54.0
2	:+	:-	:-	85.0	88.0	86.5
3	:-	:+	:-	49.0	47.0	48.0
4	:+	:+	:-	64.0	62.0	63.0
5	:-	:-	:+	65.0	61.0	63.0
6	:+	:-	:+	92.0	95.0	93.5
7	:-	:+	:+	0.57	60.0	58.5
8	:+	:+	:+	0.7	74.0	72.0

$$\mathbf{12}(-) = \frac{1}{2}(y_1 - y_2 - (y_3 - y_4)) = -8.75$$

$$\mathbf{12}(+) = \frac{1}{2}(y_5 - y_6 - (y_7 - y_8)) = -8.50$$

$$\mathbf{123} = \frac{1}{2}(\mathbf{12}(+) - \mathbf{12}(-)) = 0.125$$

Exercise 3.10

Use Eq. (2.15) to calculate the variance of the effects of a 23 factorial design without replicates, starting from Eq. (3.2).

$$\sigma_y^2 = \sum_i a_i^2 \sigma_i^2$$

$$\mathbf{T} = \bar{y}_+ - \bar{y}_-$$

$$\mathbf{V} = V(\bar{y}_+ - \bar{y}_-) = \frac{4s^2}{n}$$

$$n = 2^3$$

$$\mathbf{V} = \frac{1}{2}s^2$$

Exercise 3.11

As an exercise in a chemometrics course, M.R. Vallim and V.F. Juliano analyzed data obtained by a research worker in a series of experiments involving the synthesis of polypyrrole in an EPDM matrix. Three factors were studied: reaction time (t), oxidant concentration (C) and particle size (P). The response observed was the reaction yield. Using the data presented below, calculate the values of the effects and their standard errors. Before doing this, however, carefully examine the set of response values, taking into account the signs of the design matrix. Is it possible to anticipate which variable has the largest influence on the yield?

$$\mathbf{t} = \frac{1}{4}(5.98 + 20.34 + 3.27 + 18.69) - \frac{1}{4}(4.56 + 13.98 + 2.01 + 12.23) = 3.88$$

$$\mathbf{C} = 12.36$$

$$\mathbf{P} = -2.17$$

$$\mathbf{tC} = \frac{1}{2}(18.69 – 2.01) - \frac{1}{2}(20.34 – 4.56) = 0.45$$

$$\mathbf{tP} = –0.02$$

$$\mathbf{CP} = 2.54$$

Exercise 3.12

What conclusions can you draw from Fig. 3.5?

Increasing temperature is good. Increasing catalyst is bad. The negative effect of increasing catalyst is more severe at higher temperature.

Exercise 3.13

What is the geometric interpretation of the 123 interaction in the 2^3 factorial design?

A contrast between tetrahedra

Exercise 3.14

Our analysis of the results of the 2^3 factorial design indicates that the interactions 13, 23 and 123 can be neglected. Exclude the terms corresponding to these interactions from Eq. (3.15) and then estimate the yields for the eight runs. Calculate the differences between the estimated and observed values, and compare these differences — which are the residuals of the simplified model — with the observed average values.

```
[54.1, 85.5, 48.9, 63.1, 62.9, 94.3, 57.7, 71.9]
```

```
begin
    local C = [1 1 1 1 1 1 1 1; -1 1 -1 1 -1 1 -1 1; -1 -1 1 1 -1 -1 1 1; -1 -1 -1 -1 1 1 1 1; 1 -1 -1 1 1 -1 -1 1; 1 -1 1 -1 -1 1 -1 1; 1 1 -1 -1 -1 -1 1 1; -1 1 1 -1 1 -1 -1 1]'

    ŷ = C * [67.3; 11.4; -6.9; 4.4; -4.3; 0; 0; 0]
end
```

```
[-0.1, 1.0, -0.9, -0.1, 0.1, -0.8, 0.8, 0.1]
```

```
[54; 86.5; 48; 63; 63; 93.5; 58.5; 72] - ŷ
```

The residuals are very small, so this simplified model is still quite good.

Exercise 3.15

Write the full equation of the statistical model corresponding to the 2^4 factorial design.

$$\hat{y}(x_1x_2x_3x_4) =$$

$$b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{14}x_1x_4 + b_{23}x_2x_3 + b_{24}x_2x_4$$

$$+b_{34}x_3x_4 + b_{123}x_1x_2x_3 + b124x_1x_2x_4 + b134x_1x_3x_4b_{234}x_2x_3x_4 + b_{1234}x_1x_2x_3x_4$$

Exercise 3.16

Interpret the values in Table 3.8, in the light of the error estimates we have just made.

```
  Dict{Any, Float64}(
      1234 ⇒ 0.375
      123 ⇒ 0.875
      12 ⇒ -8.625
      24 ⇒ 0.875
      1 ⇒ 22.875
      23 ⇒ -0.625
      234 ⇒ 0.375
      14 ⇒ 0.875
      3 ⇒ 8.875
      34 ⇒ 0.375
      4 ⇒ 0.875
      13 ⇒ -0.625
      2 ⇒ -14.125
      124 ⇒ -0.125
      :average ⇒ 67.188
      134 ⇒ -0.625
)
```

```
table3_8
```

$$s = 0.54$$

0.95 CI, 5 DoF:

$$t = 2.571$$

$$st = 1.388$$

Significant effects:

```
[12, 1, 3, 2]
```

```
[key for key in keys(table3_8) if isa(key, Int) && abs(table3_8[key]) > 1.388]
```

Exercise 3.17

Suppose that x is a normalized standard variable. What are the cumulative probabilities corresponding to:

(a) $x_1 = 0$

```
0.5
```

```
0.5 # value is the mean
```

(b) $x_1 = 1$

```
0.84
```

```
round(0.68 + (1 - 0.68)/2, digits=2) # 68 % + tail of distribution
```

(c) $x_1 = 1.96$

```
0.9750021048517795
```

```
cdf(Normal(), 1.96) # cumulative density function on normal distribution
```

Exercise 3.18

Use the values of all the effects that fall on the line in Fig. 3.10 to calculate an estimate of the variance of an effect with 11 degrees of freedom. Use an F test to show that this estimate and the estimate obtained from the ternary and quaternary effects (with 5 degrees of freedom) can be considered as belonging to the same population.

$$V_{11} = \frac{3(-0.625)^2 + (-0.125)^2 + 3(0.325)^2 + 4(0.8755)^2}{11} = 0.425$$

$$V_5 = 0.291$$

$$\frac{V_{11}}{V_5} = 1.459$$

$$F_{11/5,95\%} = 4.71$$

Since the F-test value is much higher than the ratio of variances, we can say the estimates come from the same population.

Exercise 3.19

Suppose that the 123 interaction effect does not really exist for the experiment we are discussing, even though its numerical value is found to be relatively high. How can we interpret this value?

The $\mathbf{123}$ interaction is confounded with the batch effect.

Exercise 3.20

A 2^3 factorial was performed in two blocks. The runs of the second block were executed one month after those of the first one, and contain a contribution, h, caused by systematic errors that were absent from the response values of the first block. Show that the presence of this systematic difference in the second block does not affect the value calculated for the 23 interaction effect.

Without systematic error:

$$\mathbf{23} = \frac{1}{4}(y_1 + y_2 + y_7 + y_8 - y_3 - y_4 - y_5 - y_6)$$

With systematic error affecting runs $2, 3, 5, 8$:

$$\mathbf{23} = \frac{1}{4}(y_1 + y_2 + h + y_7 + y_8 + h - y_3 - h - y_4 - y_5 - h - y_6)$$

So the h terms cancel and the value of $\mathbf{23}$ is unaffected.