Experimental Design

Motivational example: forecasting temp./precip/low-level wind...
(NRL, Sandgathe)

Jargon
y = Continuous variable
x_1 = categorical variable with 3 levels
+ pressure levels. (x_7)

categ. var. = factor

y = performance (e.g. skill) score

NWP = x_1 = NCEP, WRF, COAMPS

x_2 = land, water

Season = x_3 = Autumn, Winter, Spring, Summer

x_4 = tropical, midlat, polar

x_5 = Flat, mountainous

x_6 = Grassland, forest, desert, tundra

x_7 = Level (2m, 500hPA, 200hPA)

x_8 = Station1, Station2, ..., Station5

x_8 = a 5-level factor.

PART I: How to compare multiple (≥ 2) scores.

Questions: Score(x_1=NCEP) = Score(x_1=WRF) = Score(x_1=COAMPS)?
& x_2=land ?
& x_2=water ?

To answer these questions, we need 6 runs (for 6 scores).

To make more comparisons, we need $l_1 \times l_2 \times l_3 \times \cdots l_8$
where l_i = #of levels of i^{th} factor. $3\times2\times4\times3\times2\times4\times3\times5$
= 8,640 runs !!

PART II: How to reduce the number of runs.
Not only we can reduce the number of runs, the methods of Experimental Design even tell us which runs to perform, e.g.

run 1: x_1=NCEP, x_2=water, x_3=Summer ... → Score 1
run 2: x_1=WRF, x_2=land, x_3=Autumn ... → Score 2
...

More detail

PART I: How to compare multiple (≥ 2) scores

Let μ_y = true/population mean of y

$$\begin{cases} H_0: \mu_y(x_2 = \text{land}) = \mu_y(x_2 = \text{water}) \\ H_1: \quad \neq \end{cases} \Rightarrow \text{(paired) 2-sample t-test}$$

(See Ian Jolliffe lecture)

$$\begin{cases} H_0: \mu_y(x_1 = \text{NCEP}) = \mu_y(x_1 = \text{WRF}) = \mu_y(x_1 = \text{COAMPS}) \\ H_1: \text{At least 2 of the 3 means are different} \end{cases} \Rightarrow$$

1-way ANOVA F-test (Below)

Test of needle in a haystack

The problem of multiple hypothesis testing:

When there are too many levels, it's best to not perform pair-wise 2-sample t-tests because the prob. of false alarm $\to 1$.

Type I error = Concluding there is a diff. when there is not.

If/when this single test suggests that there is a diff., then (and only then) you can ask which means are different. Even then, control family-wise error rate (future lecture?)

PART II: How to reduce the number of runs.

PART I is all about significance tests involving comparisons of multiple mean scores. In experimental design, each comparison is called an effect. So PART I is about the statistical significance of effects.

PART II focuses on the magnitude of effect (not the statistical significance of differences/effects), because doing tests requires some estimate of variability, which requires even more runs (e.g., repeated runs). Instead, the question we want to answer is whether we can reduce the number of runs but at least be able to have some point estimates (i.e., no confidence intervals, no p-values) of all the score comparisons we care about.

Example: Magnitude of effect of x_2 (on y)

= (score on water) − (score on land)

If close to zero, we'll conclude "No land/water effect".

Again, here we will not address "close."

E.g. Score = Mean Square Error

observed temp.	o_1	o_2	---	o_n
predicted temp.	p_1	p_2	---	p_n

y, above

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2$$

Ideally, for every one of the (8,640) scores, y, we also have an s_y.

Example 1: 3 models (x_1), each with 3 scores. See R code:

NCEP	WRF	COAMPS
(130, 138, 188)	(34, 174, 122)	(20, 96, 70)

$\Rightarrow$ 1-way ANOVA

"There is no NWP effect" $\Leftarrow$ There is no evidence that the 3 NWP models are different.

Example 2: x_1 and x_2, each cell with 3 scores

	NCEP	WRF	COAMPS
Land	(130, 138, 188)	(34, 174, 122)	(20, 96, 70)
Water	(150, 155, 110)	(136, 40, 120)	(25, 70, 104)

a cell.

$\Rightarrow$ 2-way ANOVA

There is an x_1 (NWP) effect

But no evidence for an x_2 (land/water) effect.

Example 3: x_1 and x_2, but 1 score in each cell.

	NCEP	WRF	COAMPS
Land	130	34	20
Water	150	136	25

$\Rightarrow$ 2-way ANOVA

Inference not possible, because there is no estimate of within-cell variability.

Example 4: Don't give-up!

(Assume there is no land/water diff.)

NCEP	WRF	COAMPS
(130, 150)	(34, 136)	(20, 25)

$\Rightarrow$ 1-way ANOVA

There is no x_1 (NWP) effect.

Exercise:

Assume there is no NWP effect.

Is there evidence for a land/water effect?

```
##################### Example 1 ################################

X1 = as.factor(c("NCEP", "NCEP", "NCEP",
                 "WRF", "WRF", "WRF",
                 "COAMPS", "COAMPS", "COAMPS"))
y = c(130, 138, 188,  34, 174, 122,  20, 96, 70)
data.frame(X1,y).            # Always good to confirm the correct matching of y to level combinations.
#      X1     y
# 1  NCEP 130
# 2  NCEP 138
# 3  NCEP 188
# 4  WRF  34
# 5  WRF 174
# 6  WRF 122
# 7  COAMPS  20
# 8  COAMPS  96
# 9  COAMPS  70

lm.1 = lm(y ~ X1)
summary.aov(lm.1)

#            Df Sum Sq Mean Sq F value Pr(>F)
# X1          2  12168    6084   2.437  0.168    No evidence that the 3 NWPs are different. There is no NWP effect.
# Residuals   6  14976    2496

##################### Example 2 ################################

x1 = as.factor(c("NCEP", "WRF", "COAMPS"))
x2 = as.factor(c("land","water"))

# mumbo-jumbo for formating the data. The details are irrelevant; just learn to use it.
temp = expand.grid(x1,x2)                # expand.grid = useful in factorial designs.
design = rbind(temp, temp, temp)         # 3 times because of 3 repeated obs.
X1 = as.factor(design[,1]) ; X2 = as.factor(design[,2])

y = c(130, 34, 20,    150, 136, 25,
      138, 174, 96,  155,  40, 70,
      188, 122, 70,  110,  120, 104)       # Note the format.

data.frame(X1,X2,y)                        #Again, always check format.
#      X1          X2   y
# 1   NCEP        land 130
# 2   WRF         land  34
# 3   COAMPS      land  20
# 4   NCEP        water 150
# 5   WRF         water 136
# 6   COAMPS      water  25
# 7   NCEP        land 138
# 8   WRF         land 174
# 9   COAMPS      land  96
# 10  NCEP        water 155
# 11  WRF         water  40
# 12  COAMPS      water  70
# 13  NCEP        land 188
# 14  WRF         land 122
# 15  COAMPS      land  70
# 16  NCEP        water 110
# 17  WRF         water 120
# 18  COAMPS      water 104

lm.1 = lm(y ~ X1 + X2 + X1:X2)
summary.aov(lm.1)

#             Df SumSq MeanSq Fvalue Pr(>F)
# X1          2  19683   9842  4.796  0.0295    # There is evidence that at least two of the NWPs is/are different.
# X2          1    214    214  0.104  0.7525    # No evidence that skill over land and water are different.
# X1:X2       2    287    144  0.070  0.9327     # Later (below).
# Residuals  12  24624   2052
```

```
##################### Example 3 ############################

X1 = X1[1:6] ; X2 = X2[1:6] ; y = y[1:6]
data.frame(X1,X2,y)

#      X1          X2   y
# 1   NCEP       land 130
# 2   WRF        land  34
# 3   COAMPS  land  20
# 4   NCEP       water 150
# 5   WRF        water 136
# 6   COAMPS  water  25

lm.1 = lm(y ~ X1 + X2 + X1:X2)
summary.aov(lm.1)

#            Df Sum Sq Mean Sq
# X1          2  13825    6912    # there is no p-value
# X2          1   2688    2688    # there is no p-value
# X1:X2       2   2726    1363    # there is no p-value

########################  Example 4 ##########################

X = as.factor(c("NCEP", "NCEP", "WRF", "WRF", "COAMPS", "COAMPS"))
y = c(130, 150,   34, 136,   20, 25)

data.frame(X,y)

# 1   NCEP 130
# 2   NCEP 150
# 3   WRF  34
# 4   WRF 136
# 5   COAMPS  20
# 6   COAMPS  25

lm.1 = lm(y ~ X)
summary.aov(lm.1)

#              Df  SumSq MeanSq Fvalue Pr(>F)
# X            2  13825    6912     3.83    0.149    # There is no evidence that the three NWPs are different,
# Residuals  3   5414    1805                        # assuming there is no land/water difference.

########################  Example 5  ############################
library(AlgDesign)                                                                  # For gen.factorial()
design = gen.factorial(c(2,2,2),varNames=c("X3","X2","X1"), factors="all")  # Play to see what this does!
attach(design)                                                                      # Allows us to use X1, X2, X3 in code.
y = c( 130, 120, 150, 110, 34, 40, 136, 90)
data.frame(X1,X2,X3,y)            # You know what I'm going to say!
# X1 X2 X3   y        # Here, factors are shown as having levels 1 and 2,
#  1  1  1 130         # instead of -1 and 1 (as in the lecture). But,
#  1  1  2 120         # keep in mind that factors are categorical, and
#  1  2  1 150         # so, in reality 1 and 2 really represent things
#  1  2  2 110         # like NCEP and WRF (for X1), or
#  2  1  1  34         # land and water (for X2).
#  2  1  2  40
#  2  2  1 136
#  2  2  2  90

mean(y[X1==2]) - mean(y[X1==1])                                  # -52.5 = x1 (NWP) effect
mean(y[X2==2]) - mean(y[X2==1])                                  #  40.5 = x2 (land/water) effect
mean(y[X3==2]) - mean(y[X3==1])                                  # -22.5 = x3 (Season) effect

(mean(y[X1==2 & X2==2]) - mean(y[X1==1 & X2==2]) ) -
(mean(y[X1==2 & X2==1]) - mean(y[X1==1 & X2==1]) )        # 71 = x1*x2 interaction effect   Etc.
```

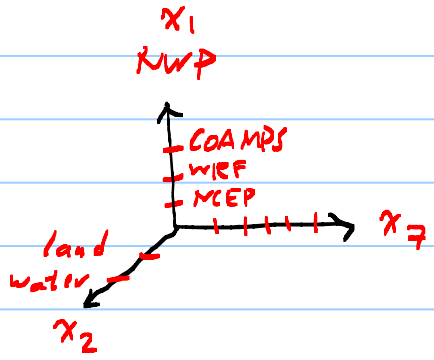
Interaction

In example 2 we found evidence that the 3 NWP models are different, but no evidence that skill score on land is different from that on water. I.e., we found an x1 effect, but not an x2 effect. Is it possible that there is an x1 effect on land but not on water? I.e., is it possible that the three NWP models are different only on land? If the answer is Yes, then one says that there is an interaction between x1 and x2. More generally, an interaction between two factors refers to the situation when the effect (on y) of one factor depends on the levels of the other factor. Higher order interactions are also possible. Mathematically, interactions appear in the model as "products," e.g., y ~ x1 + x2 + x3 + x1:x2 + x1:x2:x3 ...

We can actually answer that question by looking at the x1:x2 line of the ANOVA table. In this case, there is no evidence of an interaction effect because the p-value is large.

PART II

Look at example 3 again. Even in this "worse" of situations where we don't have multiple scores in each cell, there are 6 runs.
With 3 factors, e.g. x_1, x_2, x_7
there are $3 \times 2 \times 5 = 30$ runs/scores.



With k factors, $x_1 \dots x_k$
there are $l_1 \times l_2 \times \dots \times l_k$ runs.

→ Mathematically, No problem.
We can estimate all effects with a single k-way ANOVA model.

$$y \sim x_1 + x_2 + \dots + x_k + x_1x_2 + \dots + x_1x_2 \dots x_k$$

p-value p-value p-value ↳ p-value.

↳ And get p-values, if we have enough runs to estimate s_y (see part I)

Note: Eventhough it may look like we are doing regression, we are not, because the $x_1, x_2, \dots, x_k$ are factors (ie. categorical variables).

→ Practically, we have a problem. Too many runs!

But, as mentioned in the intro, we can reduce the number of runs and still be able to estimate the effects of interest.

For simplicity, "reduce" each factor to binary:

$x_1 = (-1, +1)$, $x_2 = (-1, +1)$, ..., $x_k = (-1, +1)$ $\Rightarrow$ k-way ANOVA

Such designs are called 2^k Factorial Designs.

This reduction is NOT the reduction being advertised. Even with $k = 8$, There are $2 \times 2 \times \cdots 2 = 2^8 = 64$ runs. Still, too many!

Example 5: 2^3 Design

	x_1	x_2	x_3	y
NCEP	-1	-1 (land)	-1	130
	-1	-1	+1	120
	-1	+1 (water)	-1	150
	-1	+1	+1	110
WRF	+1	-1	-1	34
	+1	-1	+1	40
	+1	+1	-1	136
	+1	+1	+1	90

With a single y on each row, we cannot do inference (see example 3). But we can compute the magnitude of all the effects:

(*) x_1 effect — How much does score vary across NWP?
x_2 effect — " " " " " " land/water?
x_3 effect — etc.

(†) $x_1 x_2$ effect — How much interaction is there?
$x_1 x_3$ " "
$x_2 x_3$ " "
$x_1 x_2 x_3$ " etc.

(*) = main effects
(†) = "interaction effects

Magnitude of effects:

x_1	x_2	x_3	y
-1	-1	-1	130
-1	-1	+1	120
-1	+1	-1	150
-1	+1	+1	110
+1	-1	-1	34
+1	-1	+1	40
+1	+1	-1	136
+1	+1	+1	90

mean of scores when $x_1 = -1$ (land)

$$x_1 \text{ effect} = \bar{y}(x_1 = +1) - \bar{y}(x_1 = -1) = -52.5$$

↑ See R code.

x_1	x_2	x_3	y
−1	−1	−1	130
−1	−1	+1	120
−1	+1	−1	150
−1	+1	+1	110
+1	−1	−1	34
+1	−1	+1	40
+1	+1	−1	136
+1	+1	+1	90

$$x_2 \text{ effect} = \bar{y}(x_2=+1) - \bar{y}(x_2=-1) = 40.5$$

x_1	x_2	x_3	y
−1	−1	−1	130 (*)
−1	−1	+1	120
−1	+1	−1	150
−1	+1	+1	110
+1	−1	−1	34
+1	−1	+1	40
+1	+1	−1	136
+1	+1	+1	90

$$x_1x_2 \text{ effect} = \left[\bar{y}(x_1=+1, x_2=+1) - \bar{y}(x_1=-1, x_2=+1)\right] - \left[\bar{y}(x_1=+1, x_2=-1) - \bar{y}(x_1=-1, x_2=-1)\right]$$

(*)

$$= 71.0$$

[x_1 effect when $x_2=+1$] ← cond'l effects

$-$ [x_1 effect when $x_2=-1$] ←

So:

x_1 effect $= \bar{y}(x_1=+1) - \bar{y}(x_1=-1) = -52.5$

x_2 effect $= \bar{y}(x_2=+1) - \bar{y}(x_2=-1) = 40.5$

x_3 effect $= \ldots = -22.5$

x_1x_2 effect $= \ldots = 71.0$

$\vdots$

$x_1x_2x_3$ effect $= \ldots$

There exist software for computing all of these.

Recall: Each of these effects is a difference/comparison of scores.

2^{k-p} Designs
Aliasing

Now, suppose we run only $\frac{1}{2}$ of the 2^k runs, i.e. 2^{k-1} runs.
Or, $\frac{1}{4}$ of the 2^k runs, i.e. 2^{k-2} runs.

These are called Incomplete 2^k Designs, or 2^{k-p} Designs.

$p=1 \Rightarrow$ 1/2 of the runs
$p=2 \Rightarrow$ 1/4 of the runs

$\Rightarrow$ Which $\frac{1}{2}$ (or $\frac{1}{4}$) should we run?

Play!

Suppose we run only this half:

x_1	x_2	x_3	y
-1	-1	-1	y_1
-1	-1	+1	y_2
-1	+1	-1	
-1	+1	+1	
+1	-1	-1	
+1	-1	+1	
+1	+1	-1	y_3
+1	+1	+1	y_4

above formulas $\rightarrow$

$$x_1 \text{ effect} = \frac{1}{2}(y_3+y_4) - \frac{1}{2}(y_1+y_2)$$
$$x_2 \text{ effect} = \frac{1}{2}(y_3+y_4) - \frac{1}{2}(y_1+y_2)$$
$$x_1x_2 \text{ effect} = \frac{1}{2}(y_3+y_4) - \frac{1}{2}(y_1+y_2)$$

$\therefore$ x_1 effect = x_2 effect = x_1x_2 effect.

One says "x_1, x_2, and x_1x_2 are aliased"

Alias structure: "$x_1 = x_2 = x_1x_2$"

Theorem:
If U, V, W are aliased, then the U(=V=W) effect estimated from the incomplete design = (U+V+W) effect in the complete design.

So, when we estimate the x_1 effect from this specific $\frac{1}{2}$ of the runs, we are actually estimating the $(x_1 + x_2 + x_1x_2)$ effect.

BAD News: we cannot estimate x_1, x_2 effects, separately.

Good News: Different halves have different alias structures.

Example: With a different $\frac{1}{2}$ of runs we can estimate

$x_1 + x_1x_2x_3$ and $x_2 + x_1x_2x_3$, separately.

Theorem (Hierarchy principle):

Higher-order effects are generally smaller (in magnitude).

So, $x_1 + x_1x_2x_3 \sim x_1$ $\quad$ $x_2 + x_1x_2x_3 \sim x_2$

$\Rightarrow$ We just estimated (approximately) x_1 and x_2 with $\frac{1}{2}$ of the runs!

There exist methods for deriving the alias structure for any $\frac{1}{2}$ of the runs, $\frac{1}{4}$ of the runs, ... These results are tabulated in books/software on Experimental Design (e.g. Montgomery).

Examples:

→ If you want the x_1, x_3, x_8 main effects, and the x_1x_3 effect, perform the runs ...

→ If you care only about main effects, perform these runs:

NWP land/water	X1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	
Season	X2	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	
	X3	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	
	X4	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	
	X5	−1	−1	1	1	1	1	−1	−1	1	1	−1	−1	
Station	X6	−1	1	−1	1	1	−1	1	−1	1	−1	1	−1	
	X7	−1	1	1	−1	1	−1	−1	1	−1	1	1	−1	

run 1: set all factors to "low" → score 1

run 2: set x_2–x_5 to "low", and rest to "High" → score 2

run 16 = set all to "high" → score 16

$\Rightarrow$ We need only 16 runs/scores, instead of $2^8 = 256$ runs!

The "return" ("punch") from these design ideas is larger, for larger k:

more factors

k	8	8	8	10	10	11	12	13	14	15
p	4	3	2	6	5	7	8	9	10	11
2^(k-p)	16	32	64	16	32	16	16	16	16	16

# of factors →

$\frac{1}{2^p}$ fraction →

# of necessary runs →

Our example

we can estimate all 8 effects (for $x_1, \ldots x_8$)
from only 16 runs (instead of $2^8 = 256$ runs)

With 15 factors ($x_1, \ldots, x_{15}$), we can estimate all main effects from only 16 runs (instead of $2^{15} = 32,768$ runs)

Conclusion:

Part I: BEFORE you start performing multiple comparisons (e.g. with t-test) between performance scores, make generous use of the k-way ANOVA F-test.
(statistical significance of effects)

Part II: BEFORE you start comparing performance scores, put thought into which comparisons you really care about (and which you care less about). Then, consult Exp. Design books/software to find the specific, necessary runs. (Magnitude of effects)

Part III: (Future talk!) After determining which runs are sufficient for estimating the magnitude of effects, there are techniques for performing tests of statistical significance, as well.