Section 1: Introduction to Experimental Design
(Section 1.1 & Chapter 4)

William Christensen and H. Dennis Tolley

Introduction

We have reviewed the following:

- Testing hypotheses about a single mean.
- Forming a Confidence Interval about a single mean.
- Analyses of paired observations.
- Testing hypotheses about two means.
- Forming a Confidence interval about two means.

NOW, What about more than two means???

Popcorn Example

Consider a popcorn experiment comparing two levels of salt (low and high), two levels of oil (canola and buttery), and two levels of brand (no-name and Orville). Several means of interest:

- Mean of each of two salt levels (low and high)
- Mean of each of two oil levels (canola and buttery)
- Mean of each popcorn brand (no-name and Orville)
- Mean of two salt levels for canola oil batches only.
- Mean of salt level for batches with canola oil made from Orville popcorn.
- ETC...

3-Factor Popcorn Experiment

Prefer:

- low salt or high salt?
- canola oil or buttery oil?
- Orville or no-name brand?
- Experimental Set-Up:

SALT	OIL	BRAND	SCORE
low	canola	orville	
low	butter	orville	
high	canola	orville	
high	butter	orville	
low	canola	noname	
low	butter	noname	
high	canola	noname	
high	butter	noname	

2-Factor Popcorn Experiment

Prefer:

- low salt or high salt?
- canola oil or buttery oil?
- experimental set-up

SALT	OIL	SCORE
low	canola	
low	canola	
low	butter	
low	butter	
high	canola	
high	canola	
high	butter	
high	butter	

3 Kinds of Variability

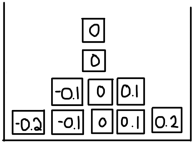
1. Planned, systematic variability
   - differences in treatments $\longleftarrow$ We want!
   - differences in material (blocks) $\longleftarrow$ We can manage
2. Chance-like variability
   - measurement error $\longleftarrow$ We can live with
     - use means of multiple measurements to minimize their effect
     - can be characterized with a box model

$$x_i = \mu + \varepsilon_i$$

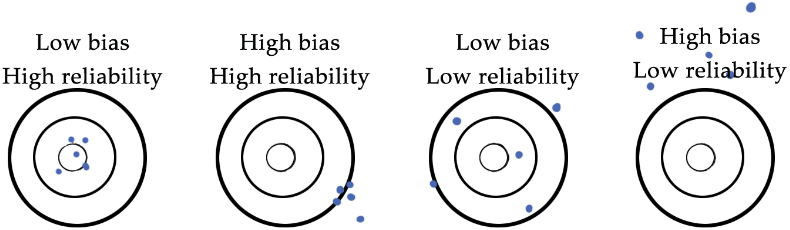
measured ht to nearest tenth of an inch

actual ht to nearest tenth of an inch

drawn from:



3. Unplanned systematic variation ⟵ This causes nightmares
   - Main threat: bias
     - Definition: An effect which deprive a statistic or observation of representativeness by systematically distorting it.
     - Contrast with a random error which may distort on any one occasion, but balances out on average.



Biases to beware

- Selection bias
  - Ensure randomization of treatments (and everything else)
- Bias due to placebo (and other) psychological effects
  - Blind subjects
  - Blind evaluators (double blind)
  - Control groups
- Measurement bias
  - Objective response variables
    - e.g., physiological measures vs. self-assessment

Approach to Experimental Design "Max-Min-Con" (Kerlinger, 1973)

- Seek to maximize the systematic variation of factors of interest and to minimize error variance by controlling the systematic variation of factors that are not of interest.

3 initial choices

1. What to measure (the response)
2. What conditions to study (the treatments)
   - Balanced designs aren't necessary, but are convenient

Example of unbalanced design:

		SALT		
		Low	Medium	High
OIL	Canola	Batch 1	Batch 2	Batch 3
	Butter	X	Batch 4	Batch 5

3. What experimental material to use (the units)
   - Remember, the structure in the material matters!

Terminology

- response: the output of a process or experiment that is measure for analysis
  - e.g., popcorn taste score
- factor: variable that can cause changes in the output (response). Experimental factors are manipulated by the researcher. Observational factors (like gender) are not.
  - e.g., "type of oil," which has factor levels "canola" and "buttery"
- treatment: a combination of factor levels (ie. a unique condition) whose effect is compared with other treatments
  - e.g., low salt with canola oil
- experimental unit: the chunk of experimental material that is assigned a treatment
  - e.g., ______________________________

Randomization

- Replace haphazard assignment with randomization when planning and executing a designed experiment.

  Use a random device (eg. random number generator in computer software)

- Why?
  1. Protect against bias
  2. Allows us to use probability and sampling distributions when analyzing the data

- The design you choose determines the way you carry out the randomization

  Eg. blocking involves restricted randomization

- The nature of a study's randomization affects the nature of the conclusions one can draw

		Treatments assigned at random?	
		Yes	No
Subjects selected at random from population of interest	Yes	Experiment: causation and inference to population	Observational Study: no causation, but inference to population
	No	Experiment: causation but limited inference	Observational Study: No causation and limited inference

Examples?

Replication

- "Replicate" = An independent run of an experimental condition so that variability can be estimated
  - Replication is <u>not</u> the same as repeated measurements on the same experimental unit. Replication means assigning multiple experimental units to the same treatment.
  - Eg. Three classrooms are used, each one assigned to a different teaching method (A, B, or C). 30 students in each classroom. Each student takes 2 different assessments of mastery. How many replicates (experimental units) do we have for each treatment?

Decisions:

- What measurement to make (response)
- What conditions to compare (factors, levels, treatments)
- What material to use (what are the experimental units and how units are structured)

$\boxed{\text{EX}}$ Compare long-term and short-term memory for verbal and visual information

- Response:
- Conditions:
- Material:

1. Choosing the Response

3 criteria to consider

1. Reliability (repeatability)
2. Validity (relevance)

   EX Consider the quality of these responses:

   - popcorn tastiness score?
   - attendance at concert (via visual inspection of aerial photo)?
   - Rorschach (ink blot) score for evaluating neurosis?
   - Self-evaluation exam (eg. Minnesota Multiphasic Personality Inventory) for evaluating neurosis?
   - Non-contextual word frequencies for evaluating unique authorship?

3. Scale of measurement
   Some types of variables are more amenable for ANOVA than others.
   Steven's Scales of Measurements
   1. Nominal: response is a non-numerical category (never use ANOVA) $\longleftarrow$ Never use ANOVA
   2. Ordinal: responses can be ordered, but there is no meaningful notion of distance between categories $\leftarrow$ Sometimes use ANOVA
   3. Interval: responses are numbers with meaningful notion of distance between values$\longleftarrow$ ANOVA generally appropriate
   4. Ratio: responses are numbers with meaningful notions of distance and relative size (eg. "twice as large") $\longleftarrow$ ANOVA generally appropriate
      - For ratio data, there is a meaningful zero in the scale

$\boxed{\text{EX}}$ What is scale?

1. Nominal
2. Ordinal
3. Interval
4. Ratio

- popcorn tastiness score?
- attendance at concert (via visual inspection of aerial photo)?
- Rorschach (ink blot) score for evaluating neurosis? (1=normal, 2= borderline, 3= neurotic, 4= psychotic)
- Self-evaluation exam (eg.Minnesota Multiphasic Personality Inventory) for evaluating neurosis? (suppose: scores from 27 to 108)
- Non-contextual word frequencies for evaluating unique authorship?

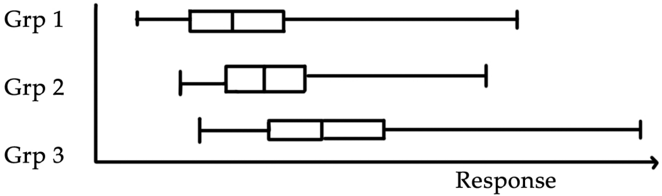
- Additional note: Sometimes it makes sense—either scientifically or to facilitate the analysis—to consider the following as response variables
  - Change in measurement:

    $$y_{post} - y_{pre}$$

  - Percent change in measurement

    $$\frac{y_{post} - y_{pre}}{y_{pre}} \cdot 100\%$$

  - log(y) - good idea when y is heavily right-skewed



2. Choosing the Conditions

The main point here is to isolate the effects of interest from all other possible variables that might be affecting the response.

- Use control groups
  - EX Gastric freezing ⟵ “placebo effect” is the trouble maker
  - EX Chiropractic and alternative medicine

$\boxed{\text{EX}}$ Memory and Meaning (ex. 4.7, Cobb)

- Does meaning affect memory?

Sentence Pair A (similar meaning)

1. When you score your results, do nothing to correct your answers but **mark carefully** those answers which are wrong.
2. When you score your results, do nothing to correct your answers but **carefully mark** those answers which are wrong.

Sentence Pair B (different meaning)

1. When you score your results, do nothing to **correct your answers** but mark carefully those answers which are wrong.
2. When you score your results, do nothing to **your correct answers** but mark carefully those answers which are wrong.

Experimental vs. Observational Studies

- **Experiment:** compares treatments
  - Start with one set of subjects/material
  - Assign treatments randomly
  - Easier to isolate effects of interest—either hold constant or randomize all factors other than the factors of interest)
- **Observational Study:** compares populations
  - Start with several populations with conditions already built in
  - Take random sample from each population
  - Conditions you want to study will almost never be the only thing that differentiates populations

    $\boxed{\text{EX}}$ Bladder cancer in Mormons: Teetotalers?

Confounding

- Two influences on the response are **confounded** if the design makes it impossible to isolate the effects of one from the effects of the other.
  - What effects are confounded in
    - Early gastric freezing studies
    - Bladder cancer in Mormons study
    - Gender differences in spatial ability
    - Early proposal for polio vaccine study: compare polio rates for those who choose vaccination vs. those who abstain
  - **Selection bias**: process of selecting groups confounds the effects of interest with other effects

3. Choosing the Material

Experiment: Unit is a chunk of material that receives the assignment of a treatment

- The information you get from a study depends more on the number of units than the number of response values you have

Observational Study: Unit is an item or individual in the population of interest

Ideally, units used for an experiment will be:

- representative
  - ...but representativeness is difficult and expensive
- uniform (yielding small chance errors)
  - In contrast with uniformity, if there's lots of natural variability in your experimental material, that variability will show up in your response

    $\Rightarrow$ Large chance errors

    $\Rightarrow$ Hard to show significant differences in treatment groups

- Don't force uniformity when it's not there - choosing units all of one type will yield units that are not representative
  - $\boxed{\text{EX}}$ lab rats vs. wild rats
- if lots of natural variability, use blocking to better isolate the effects of interest

,

- Compare: (1) a random sample of U.S. 20-year-olds with (2) a collection of LDS missionaries as subjects in a study evaluating response to visual images (violent, beautiful etc.)
  - Representative?
  - Uniform?
  - Expected conclusions from experiments?

Overview of Types of Designs

Analysis Approaches:

		RESPONSE VARIABLE **Continuous**	**Categorical**
EXPLAN.	**Continuous**	Regression	Logistic
VARI-	**Mix**	Regression/ANCOVA	Regression
ABLES	**Categorical**	ANOVA	L.R. or χ^2

Basic Factorial Design (BF or "Completely Randomized Design")

- Randomly assign treatments to the experimental units
- Good choice for your design when your experimental material is reasonably uniform
- If there is only one factor $\Rightarrow$ BF[1]
  - analysis method called "one-way ANOVA"

- Multiple factors $\Rightarrow$ BF[2], BF[3],...
  - With multiple factors that we want to study, we cross the factors and consider all the treatment combinations as our conditions

    E.g.,

		Oil	
		Canola	Buttery
Salt	Low	Trt # 1	Trt # 3
	High	Trt # 2	Trt # 4

  - Often interested in the interaction of factors
    - Definition: The failure of one factor to produce the same effect on the response at different levels of another factor
    - E.g., "There is an interaction between salt and oil because high salt yields much higher scores with canola oil, but only slightly higher scores with buttery oil."
    - Generic interpretation of a significant interaction between factor A and factor B: "The effect of [factor A] on [the response variable] is different for differing levels of [factor B]" –OR– "A change in [factor A] produces a different change in [the response variable] at two levels of [factor B]."

- One-factor-at-a-time experiments can yield incorrect decisions when interactions are present.
- Factor structure for BF[2] design (CRD with 2 factors)
  - EX Imagine a popcorn experiment with two factors (oil and salt) and 2 reps of each treatment.

		Salt	
		Low	High
Oil	Canola	6.2 7.1	6.9 7.5
	Buttery	8.0 7.6	8.1 7.9

$\longleftarrow$ scores for the 8 batches

Visualize the main effects for salt and oil and the interaction by sorting 8 cards with numbers on them—then comparing means

- Assess main effect for salt by considering...

		Salt: Low	Salt: High
Oil	Canola	6.2 7.1	6.9 7.5
	Buttery	8.0 7.6	8.1 7.9

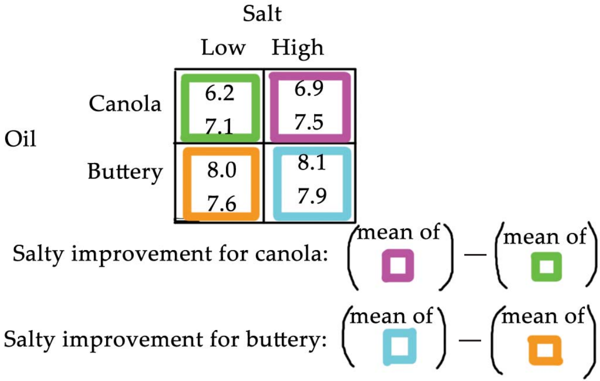
– Mean of minus mean of

- Assess main effect for oil by considering...

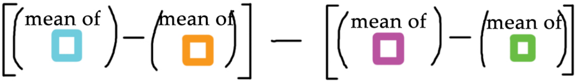
		Salt: Low	Salt: High
Oil	Canola	6.2 7.1	6.9 7.5
	Buttery	8.0 7.6	8.1 7.9

– Mean of minus mean of

- Interaction between oil and salt (“oil*salt”)
  - Consider the change in score when we move from low salt to high salt. We’ll call that change the “salty improvement.”
  - Question: Is the salty improvement the same for canola and battery oil? If NO, we have an interaction



- Assess oil×salt interaction (“oil*salt”) by considering...



OR

		Salt	
		Low	High
Oil	Canola	6.2 7.1	6.9 7.5
	Buttery	8.0 7.6	8.1 7.9

2 * [Mean of minus mean of]

If statistically significant, we say there’s a significant interaction and use caution in interpreting main effects

- In addition to the main effects and interactions of interest, this (and every) design will have an "effect" or "factor" due to the grand (overall) mean, and the residual error (which has as many levels as there are observations in the study).
  - In this sense, "factor" is used as a component of a statistical model that is associated with a useful sorting of observations into groups
- “Factor Diagram”:

Randomized Complete Block Design (CB)

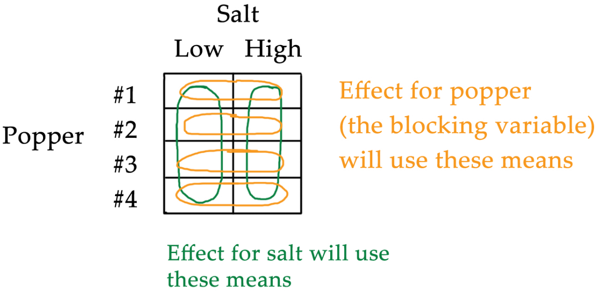
Blocking:

- Converts unplanned, systematic variation into planned, systematic variation
- Involves "restricted randomization" to explicitly control for known nuisance factors
- Good strategy if you can sort your experimental material into groups of units that are:
  - similar within groups
  - different across groups (where "similar" and "different" refer to the likely values for the response)

- Eg. Two researchers decide to save time and money by recruiting the same 30 persons as subjects for their studies. The subjects are grouped into 3 groups according to Body Mass Index (low, medium, and high).
    - Researcher 1 is studying the impact of protein supplements as a means for increasing muscle mass.
    - Researcher 2 is studying the impact of violent movie images on aggressive driving
    - Critique the designs with respect to the choice of blocks.

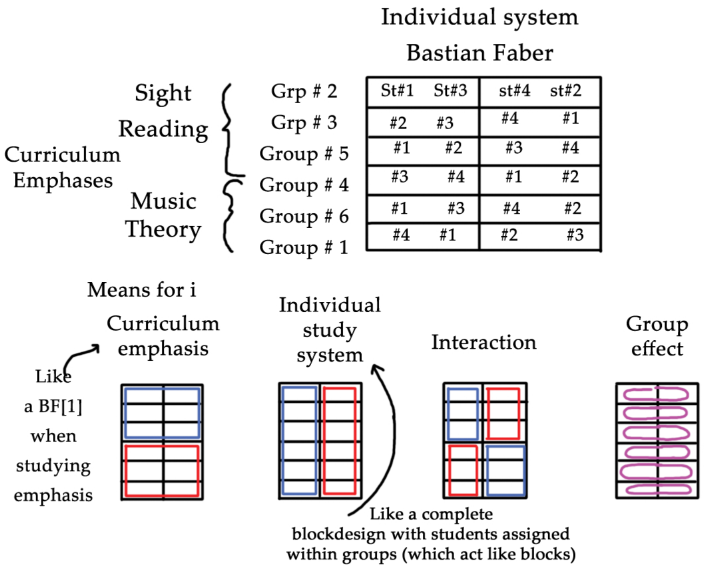
- Factor Structure for CB[1]

  EX Imagine 4 different popcorn poppers, each used to compare low salt to high salt.



Split Plot Design

- $\boxed{\text{EX}}$ A local piano teacher incorporates group lessons as a central component of her instruction. She has 6 groups of intermediate students that she will use to compare 2 group-study curriculum emphases ("sight-reading emphasis" vs "music theory emphasis") AND 2 individual study systems ("Bastian" vs. "Faber"). Each group is randomly assigned a curriculum emphasis (3 groups to each emphasis) and the four students within each group are randomly assigned to one of the individual study systems (2 to each system). Response is player improvement score after 1 year.
  - What is the experimental unit?
  - How many experimental units are available to study each main effect?


Individual system
Bastian Faber
Sight
Reading
Curriculum
Emphases
Music
Theory
Grp # 2
Grp # 3
Group # 5
Group # 4
Group # 6
Group # 1
St#1 St#3 st#4 st#2
#2 #3 #4 #1
#1 #2 #3 #4
#3 #4 #1 #2
#1 #3 #4 #2
#4 #1 #2 #3
Means for i
Curriculum
emphasis
Like
a BF[1]
when
studying
emphasis
Individual
study
system
Interaction
Group
effect
Like a complete
blockdesign with students assigned
within groups (which act like blocks)