Model Construction Process Outline

The model construction process at its core is comprised of three key building blocks:

I. Sketching a ***Study Blueprint***
II. Utilizing a ***What Would Fisher Do (WWFD)*** table
III. Writing the ***Model and Model Conditions***

Building Block I: Study Blueprint

Part 1: Sketch relevant explanatory variables.
Part 2: Rough sketch of the organization of the units.
Part 3: Rough sketch of the "randomization" scheme.

Building Block II: WWFD Table

Step 1: Treatment Structure.
Objective: Identify any treatments (i.e. explanatory variables). These could be innate characteristics or attributes manipulated in a study. Determine the corresponding degrees of freedom using the blueprint layout. Any extra degrees of freedom that are not attributable to the treatment fall into a blanket category named leftovers, originally referred to as "parallels". [3] [4]
Step 2: Design Structure.
Objective: Determine the organizational structure of the units (including hierarchical structures) and the corresponding degrees of freedom (df). Pay close attention to the units where treatment(s) will be assigned (or were observed). This interplay becomes especially important for hierarchical structures.
Step 3: Combined Structure.
Objective: Combine the treatment and design structures used in the study and account for corresponding degrees of freedom (df). Unlike the first two steps, in which degrees of freedom are assigned following the study blueprint directly, the combined degrees of freedom have to account for both the design and treatment structures. A generic template of a WWFD table is shown below.

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
Total		**Total**		**Total**	

Building Block III: Model and Model conditions

What terms should be in the model? All the terms from the combined column of the WWFD table. **There is a one-to-one correspondence between the terms in the combined column and model terms!**
Terms that are a part of the Treatment Structure will always be fixed. Terms that are a part of the Design Structure may be fixed or random, depending on the study. Additionally, we emphasize that the role of the error term is flexible depending on the distribution that the data follows. This allows for a smoother transition beyond the Gaussian distribution.

Terminology

- ☐ **Study Blueprint:** visualization of a layout of experimental units and treatment assignments
- ☐ **Experimental Study:** A study where the researcher actively manipulates one or more variables (factors) to observe effects on the outcome, allowing for stronger causal inferences.
- ☐ **Observational Study**: A study where the researcher observes and measures variables without manipulating them, often used to find associations but not to establish causality.
- ☐ **Units:** The individual entities or cases (people, animals, objects, etc.) on which data are collected in a study.

  Important notes:
  - A unit is where the treatment is independently applied, an important distinction between replication—the repetition of independent units to assess variability—and pseudoreplication
  - There can be more than one size of unit in the same study. For example, in split-plot designs, different sizes of units receive different treatment factors.
- ☐ **Explanatory Variable/Treatment:** A variable(s) that is manipulated or categorized to determine if it has an effect on the response variable. Other names used depending on the context: treatment, covariate, independent variable. Can be categorical or numerical.
- ☐ **Random Sample:** A sample selected from a population where each unit has an equal chance of being chosen, ensuring representativeness.
- ☐ **Random Assignment:** Assigning units to treatment groups by randomization, reducing bias, and helping establish causality.
- ☐ **Blocks:** Groups of units that are similar in some way, that are expected to affect the response; blocking accounts for extraneous variability.
  - ☐ **Why do we block?** To control for variability among groups, improving the accuracy and efficiency of the experiment.
  - ☐ **Benefit:** Reduces confounding, increases precision of comparisons.
  - ☐ **Cost:** Adds complexity to the design and analysis.
- ☐ **Fixed Effects:** Variables that capture the impact of specific, non-random characteristics—typically those that do not vary across units
- ☐ **Random Effects:** Variables that capture the impact of specific, random characteristics— these may vary across units. These are considered a random sample from a larger population.
- ☐ **What Would Fisher Do** (revisioned ANOVA table)- A framework that characterizes sources of variation attributable to explanatory variables vs the study design before building an ANOVA table or writing a model.
- ☐ **Degrees of Freedom:** Number of independent pieces of information from the study.

Terminology Continued

- ☐ **One-way Treatment Structure:** Investigating the effect of a single independent variable/factor on a dependent variable. This single factor has multiple levels or groups that are compared to see if they have a significant impact on the outcome.
- ☐ **Two-way Treatment Structure:** Investigating the effects of two or more factors/independent variables on a response variable, where each factor has at least two levels.
- ☐ **Interaction Effect:** When the effect of one explanatory variable on the response may change depending on another explanatory variable.
- ☐ **Main Effect:** Impact of a single independent variable on a dependent variable, averaged across all levels of other independent variables. Essentially, examines the effect of one factor while ignoring the influence of other factors in the experiment.
- ☐ **CRD (Completely Randomized Design):** A design structure in which treatments are randomly assigned to experimental units.
- ☐ **RCBD (Randomized Complete Block Design):** A design structure in which experimental units are grouped into blocks, and each block receives all treatment conditions.
- ☐ **Split Plot:** A design structure in which factors are applied to different-sized experimental units within the same experiment. It involves whole plots (larger experimental units) and sub-plots (smaller units within the whole plots).
- ☐ **BIBD (Balanced Incomplete Block Design):** A design structure in which treatments are arranged into blocks, with each block containing a subset of the total treatment/treatment combinations. Key features include each treatment appearing an equal number of times (replication) and each pair of treatments appearing together in a specific number of blocks (balance).

References

[1] Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2018). Introduction to statistical investigations (2nd ed.). Wiley.

[2] Milliken, G.A., & Johnson, D.E. (2009). Analysis of Messy Data Volume 1: Designed Experiments, Second Edition (2nd ed.). Chapman and Hall/CRC.

[3] Stroup, W., Ptukhina, M., Garai, J., 2024. Generalized linear mixed models. Boca Raton, FL. CRC Press.

[4] Stroup, W., Milliken, G., Claassen, E. & Wolfinger, R. 2018. SAS for Mixed Models: Introduction and Basic Applications. Cary, NC. SAS Institute Inc.

Contact Information (please reach out with any questions)

Dr. Marina Ptukhina	Associate Professor at Whitman College	ptukhim@whitman.edu
Dr. Julie Garai	Senior Data Scientist at Incredible Health	julie.garai@incrediblehealth.com

Close Friends[1]

A random representative sample of adult Americans was asked:

> *Looking back over the last six months, who are the people with whom you discussed matters important to you? Just tell me their first names or initials.*

The interviewer then recorded how many names each person gave, along with demographics of the respondent (e.g. sex, location, age group). Note for students that sex here may be binary because other sex options were limited in sample size.

Research Question:

Do men and women tend to differ on the number of close friends?

Study Results:

Summary Statistics

	n	Mean	SD
Women	814	2.09	1.76
Men	654	1.86	1.78
pooled	1468	1.99	1.77

Observed Difference

-0.228

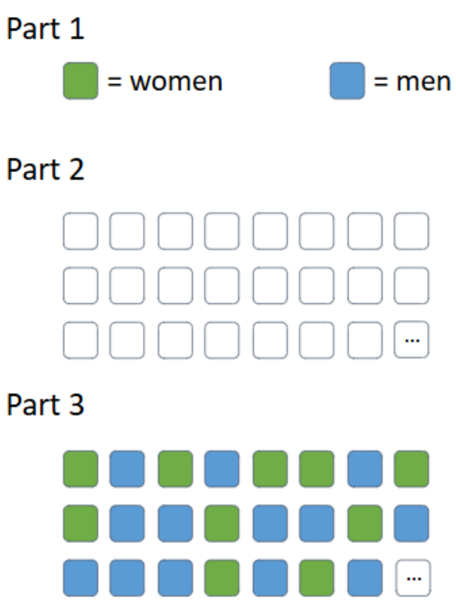
Close Friends: Scenario 1A

Research Question:

Do men and women tend to differ on the number of close friends?

Analysis

Scenario 1A: **Study Blueprint**



Part 1: Talking Points
- Explanatory variables… assign colors / patterns

Part 2: Talking Points
- Draw the units… Who or what are we gathering data on. Here it's just the survey respondents…
- No need to draw all 1468 boxes
- Let's just assume this is the order in which respondents are surveyed, though it doesn't really matter

Part 3: Talking Points
- Record the values of the explanatory variable

Reference Definitions
Explanatory variable(s), treatments, random sampling, random assignment

WWFD: Arrow means that we observed(applied) treatments (explanatory variables) on the units to which the arrow is pointing.
Vertical bar | means "after accounting for"

Model in words

Scenario 1A: **WWFD**

Treatment Structure		**Design Structure**		**Combined Structure**	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
Sex	2 - 1 = 1			Sex	2 - 1 = 1
leftovers	1467	Participants	1468 - 1 = 1467	Paricipants|Sex	1467-1=1466
Total	**1468-1 = 1467**	**Total**	**1467**	**Total**	**1467**

Scenario 1A: **Model**

Number of close friends = Sex + error

Think about mentioning using error / residual. Here they mean the same thing. Be consistent with your students!

Close Friends: Scenario 1B

Research Question:

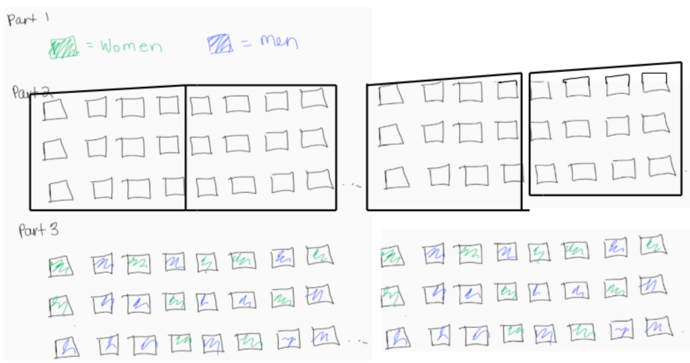
Do men and women tend to differ on the number of close friends?

Additional Context:

- Suppose we have additional information about the respondents, specifically what region of the US they live in: East Coast, West Coast, Midwest, South.

Analysis

Scenario 1B: **Study Blueprint**



Talking points: show how to incorporate information from the description of the study in the study blueprint.

To Do: Draw four "squares" around part 2 and part 3

Reference Definitions: Blocks

Talking points: Follow up… what are we assuming about region? Is this reasonable? Do we want to change our research question? Does it do students a disservice to make them the same size?

If we care about region, how does the blueprint change? (Part 1!)

What assumptions do we make if we only include region in Part 2?

For now: Let's keep it in Part 2/3 only.

Scenario 1B: **WWFD**

Treatment Structure		**Design Structure**		**Combined Structure**	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
		US Region	4-1=3	US Region	4-1=3
Sex	2 - 1 = 1			Sex	2 - 1 = 1
leftovers	*1466*	Participants	1464	Paricipants|Sex	1464 - 1=1463
Total	**1468-1 = 1467**	**Total**	**1467**	**Total**	**1467**

Scenario 1B: **Model**

Number of close friends = *Region* + Sex + error

Close Friends: Scenario 1C

Research Question:

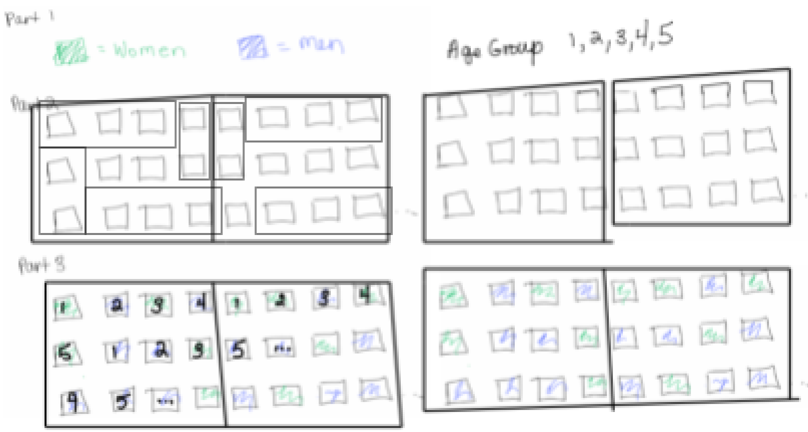
Do men and women tend to differ on the number of close friends?

Additional Context:

- Suppose in addition to region (East Coast, West Coast, Midwest, South), we also have 5 age groups for the respondents: 18-22, 23-30, 31-45, 46-65, 65+

Analysis

Scenario 1C: **Study Blueprint**



Talking points:
Should age be a block or a treatment? What does each one allow for? We want to have this discussion with students as researchers rather than just coders.

- Age group is a block: Age group is just noise. We are not interested in measuring it. We don't anticipate there will be an interaction between the treatment (sex) will the age group. Add simple human non stat words here. (Core assumption of blocking, no treatment by block interaction). Add rectangles to part 3 to illustrate blocking by age.
- Reduction in Variability… Why is this beneficial? More powerful tests, easier to detect differences that are small.

Follow-up question… is that a fair assumption?

- Assume no. We think there could be some interaction between sex and age group impacting the number of close friends.
- NOTE: Having this conversation with students now highlights best practices of deciding on how to not model-cheat.
- Age group is a treatment: Allows us to test a sex*age group interaction. This allows us to possible test higher order terms, quadratic?
- Do the highest age groups suddenly socialize more in retirement?
- Also opens the door for analysis of covariance if we want to include a more linear relationship for age.

Scenario 1C: **WWFD**

Age as a block

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
		US Region	4-1=3	US Region	4-1=3
		Age	5-1=4	*Age*	5-1=4
Sex	2 - 1 = 1			Sex	2 - 1 = 1
leftovers	*1466*	Participants	1460	Paricipants|Sex	1460-1=1459
Total	**1468-1 = 1467**	**Total**	**1467**	**Total**	**1467**

Age as a treatment + interaction

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
		US Region	4-1=3	US Region	4-1=3
Sex	2 - 1 = 1			Sex	2 - 1 = 1
Age	5-1=4			*Age*	5-1=4
*Sex*Age*	(2-1)(5-1)=4			*Sex*Age*	(2-1)(5-1)=4
leftovers	1462	Participants	1464	Paricipants|Sex,Age, Sex*Age	1464-9=1455
Total	**1468-1 = 1467**	**Total**	**1467**	**Total**	**1467**

Scenario 1C: **Models**

Age as a block: **Number of close friends = *Region* + *Age* + Sex + error**

Age as a treatment: **Number of close friends = *Region* + Age + Sex + Age*Sex + error**

Baking Bread[2]: Scenario 2A

Research Question:

A baker wants to create the fluffiest possible loaves of bread, measured by overall volume, by investigating the impact of fat and yeast.

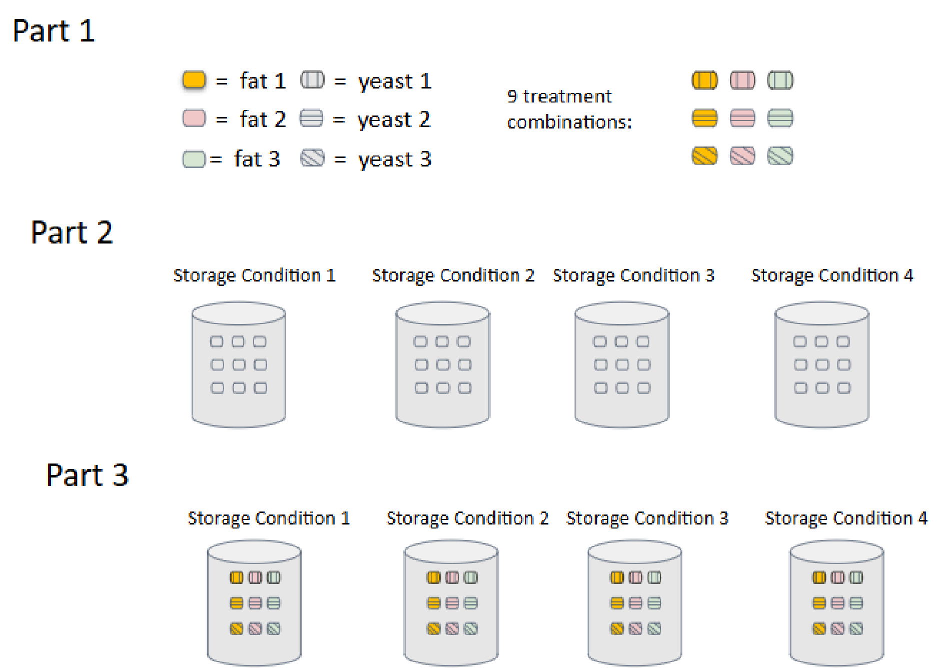
Additional Context:

They decide to study the effects of combining three different fats with three different brands of yeast, resulting in nine total combinations. The baker is using one type of flour, but needs to take into account that it was stored under four different conditions. This could impact the rise and volume of the final bake. The baker uses the flour with each combination of fat and yeast and bakes all loaves at the same temperature in the same oven. All loaves are given the exact same amount of time to rise and are baked under the same conditions. Study results are shown in the table below.

Fat	Yeast	Flour Storage Condition			
		1	2	3	4
1	1	6.7	4.3	5.7	7
	2	7.1	4.8	5.9	5.6
	3	6.8	5.5	6.4	5.8
2	1	6.5	5.9	7.4	7.1
	2	6.9	5.6	5.8	6.8
	3	6.4	5.1	6.2	6.3
3	1	7.1	5.9	7.9	6.5
	2	7.3	6.6	8.1	6.8
	3	7.2	7.5	9.1	6.2

Analysis:

Scenario 2A: **Study Blueprint**



Good time to discuss random assignment with students we don't expect the same order of treatment combinations

Scenario 2A: **WWFD**

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
–	–	Storage Condition	4-1=3	Storage Condition	4-1=3
Fat	3-1 = 2	–	--	Fat	3-1 = 2
Yeast	3-1 = 2			Yeast	3-1 = 2
Fat*Yeast	(3-1)(3-1) = 4			Fat*Yeast	(3-1)(3-1) = 4
leftovers		Loaf(SC))	(9-1)*4=32	Loaf(SC) | Fat,Yeast,F*Y	32-8=**24**
Total	**36-1=35**	**Total**	**35**	**Total**	**35**

Scenario 2A: **Model**

Volume = *Storage condition* + Fat + Yeast + Fat*Yeast + error

Potential Follow-up
Note: If we didn't change anything could we answer the question of what is the optimal storage for each fat*yeast combination? (We can't because it becomes a treatment factor and we don't have enough replication of a three-way effect) . This is an exercise for practicing the difference between treatment factors and blocks. How does this table change if we are interested in storage as a factor?

Baking Bread: Scenario 2B

Research Question:

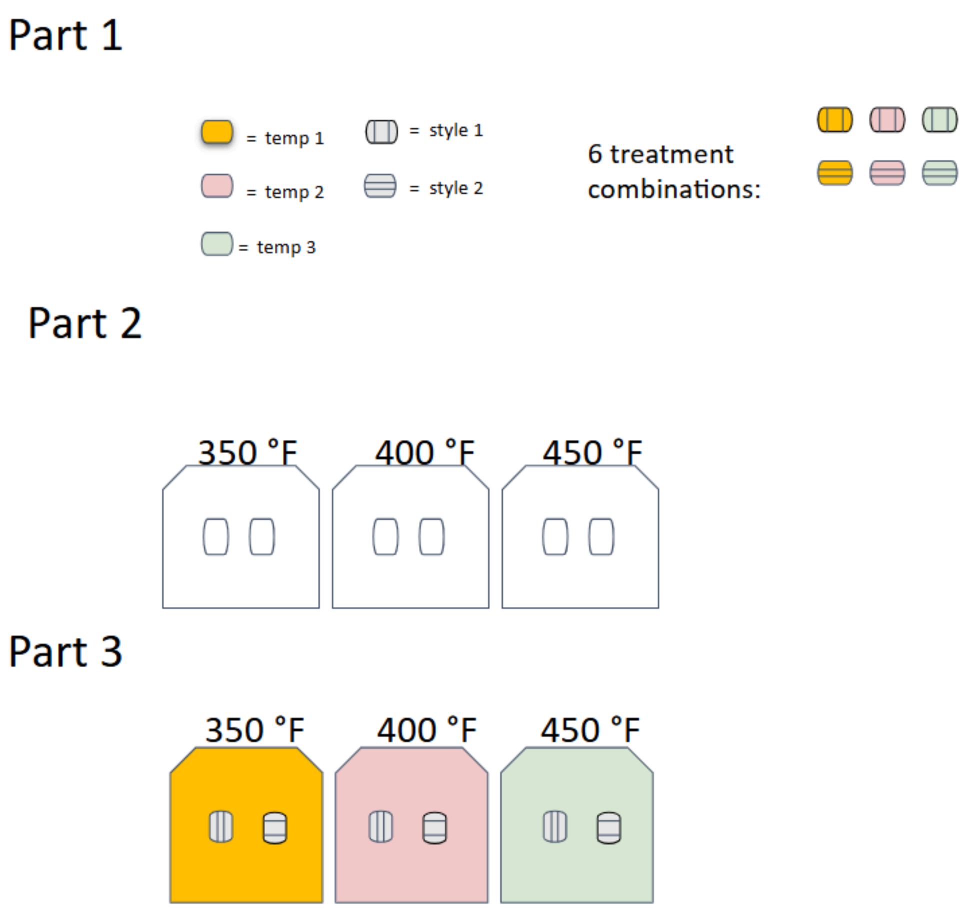
The baker now wants to determine the optimal baking conditions: temperature and style, to continue having the fluffiest bread with a **crispy crust**.

Additional Context:

The baker has been making fluffy loaves for a few months using the results of the last study. They just moved to a commercial kitchen with 3 ovens. The baker is considering three different temperatures: 350°F, 400°F, and 450°F, and two different baking styles: Dutch oven and hot stone. They decide to bake two different loaves in each of the three ovens, which are randomly assigned to the three temperatures. For example, the first oven is assigned 350°F, and one of the loaves is in a Dutch oven, and the other is on a hot stone. They bake all loaves according to the assigned temperature for the same amount of time, and then measure the crust crispiness for each loaf.

Analysis:

Scenario 2B: **Study Blueprint**



Scenario 2B: **WWFD**

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
Temperature	3-1 = 2	–	--	Temperature	3-1 = 2
		Oven	3-1=2	Oven(Temperature)	2-2=0
Baking style	2-1=1	-	-	*Baking style*	2-1=1
Temperature*Baking Style	(3-1)*(2-1)=2	-	-	Temperature*Baking Style	(3-1)*(2-1)=2
leftover		Loaf(Oven)	(2-1)*3=3	Loaf(Oven)	3-3=0
Total	**6-1=5**	**Total**	**6-1=5**	**Total**	**6-1=5**

Scenario 2B: **Model**

Crispiness = Temperature + larger-unit error + Baking style + Temperature*Baking style + error

Baking Bread: Scenario 2C

Research Question:

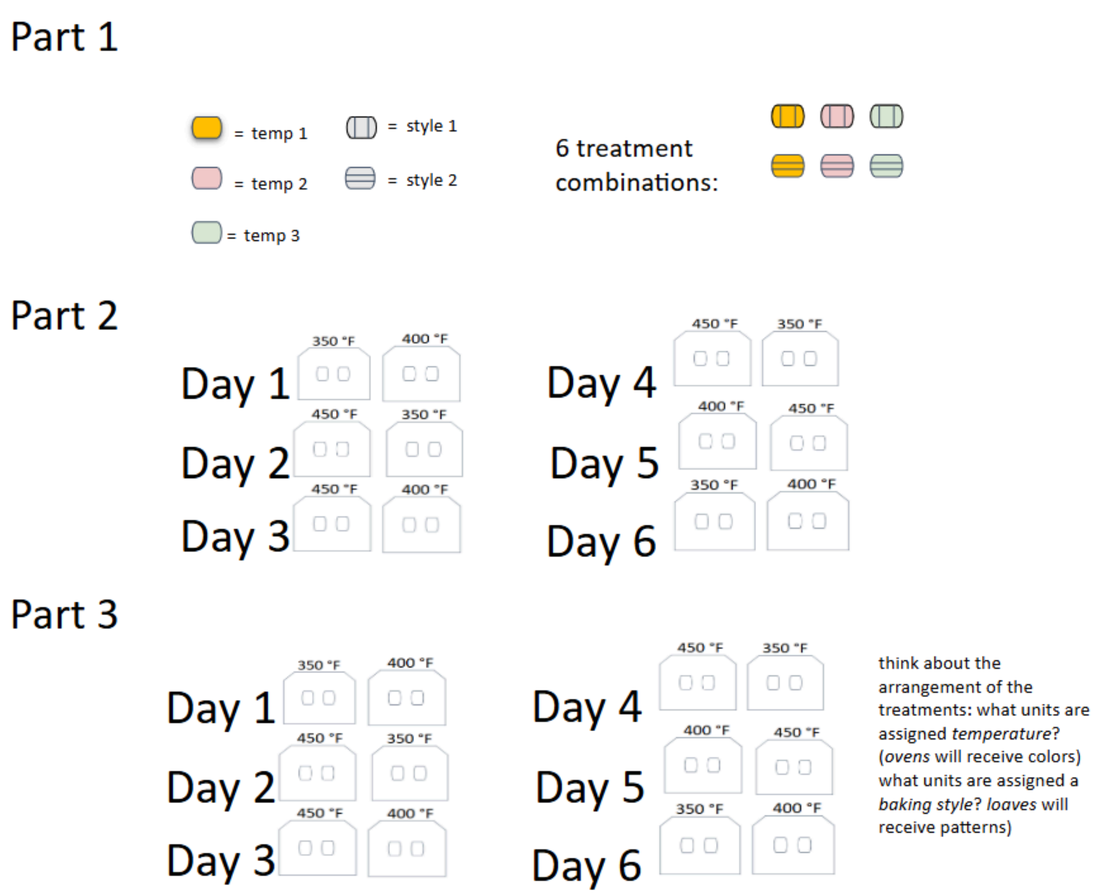
The baker still wants to determine the optimal baking conditions: temperature and style, to continue having the fluffiest bread with a **crispy crust**.

Additional Context:

After consulting with you, the baker acknowledges that they need additional replication. They agree to use 2 of the 3 ovens for the experiment, one time every day. They still want to use two different baking styles and determine the optimal baking temperature. The baker decides to bake two different loaves in each of the two ovens available on a given day over 6 different days. The weather and humidity of each day vary. They bake all loaves according to the assigned temperature for the same amount of time, and then measure the crust crispiness for each loaf.

Analysis:

Scenario 2C: **Study Blueprint**



Scenario 2C: **WWFD**

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
		Day	6-1=5	Day	6-1=5
Temperature	3-1 = 2	–	--	Temperature	3-1 = 2
		Oven(day)	(2-1)*6=6	Oven(Day)|Temperature	6-2=4
Baking style	2-1=1	-	-	Baking style	2-1=1
Temperature*Baking Style	(3-1)*(2-1)=2	-	-	Temperature*Baking Style	(3-1)*(2-1)=2
leftover		Loaf(Oven(Day))	(2-1)*2*6=12	Loaf(Oven(Day))|BS, T*BS	12-3=9
Total	**24-1=23**	**Total**	**24-1=23**	**Total**	**24-1=23**

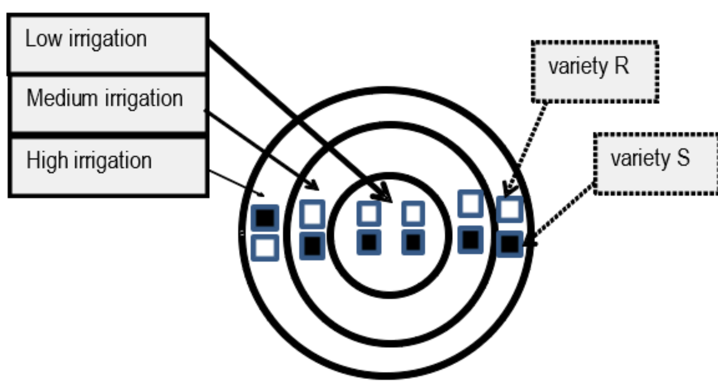
Scenario 2C: **Model**

Crispiness = *Day* + Temperature + larger-unit error + Baking style + Temperature*Baking style + error

Scenario 3: Practice

A researcher comes to you for help analyzing an experiment to assess the effects of irrigation and variety on the response of a crop (generically referred to as "Y" in this problem). There are three irrigation levels: low, medium, and high. There are two varieties of a crop, denoted here as R and S.
The objective of the experiment is to characterize the effects of irrigation and variety.
In particular, researchers want to understand how the varieties perform under different irrigation regimes. The experiment design used center-pivot irrigation. The nozzles on the irrigator can be set so that concentric rings in the field receive different levels of irrigation. The design used four fields, each with one center-pivot irrigator. Three irrigation levels were included per field, using a randomized complete block design. Within each irrigation regime, four plots were planted, two to variety R, the other two to variety S (assignment was random, of course!). The figure below shows a study blueprint for one of the fields to help you visualize the layout of the experiment. Remember that there are four fields in total.

Scenario 3: **Study Blueprint (Part 3)**



Scenario 3: **WWFD**

Treatment Structure		Design Structure		Combined Structure	
Source of Variation	**df**	**Source of Variation**	**df**	**Source of Variation**	**df**
		Field	4-1=3	Field	4-1=3
Irrigation	3-1 = 2	–	--	Irrigation	3-1 = 2
		Circle(field)	(3-1)*4=8	Circle(field)|Irrigation	8-2=6
Variety	2-1=1	-	-	Variety	2-1=1
Irrigation*Variety	(3-1)*(2-1)=2	-	-	Irrigation*Variety	(3-1)*(2-1)=2
leftover		Square(Circle(Field))	(4-1)*3*4=36	Square(Circle(Field))|I,V,I*V	36-3=33
Total	**48-1=47**	**Total**	**48-1=47**	**Total**	**48-1=47**

Scenario 3: **Model**

Crop response= *Field* + Irrigation + larger-unit error + Variety + Variety*Irrigation + error