Efficient Learning Using Designed Experiments

DREW BAUMGARTEL

STATISTICIAN AT INTEL CORPORATION, LOGIC AND TECHNOLOGY DEVELOPMENT

TECHCON, SEPTEMBER 8 2019

The Lady tasting Tea

R.A. Fisher, a British geneticist, largely developed the field of Experimental Design.

Inspired by true events: Fisher's co-worker, Muriel Bristol, claimed that she could discern between cups of tea where the milk was poured first and cups of tea where the tea had been poured first.

- 8 cups of tea, 4 with milk first and 4 with tea first, **randomized**
- Bristol was made aware of the experimental setup
- She needed to choose enough cups correctly to demonstrate she could tell the difference
- Prob of choosing ≥ 3 correctly by chance = 0.24 → not unlikely!
- Prob of choosing 4 correctly by chance = 0.01 → unlikely!

R.A. Fisher

A Brief History of DoE

1919 Fisher begins work at Rothamsted Experimental Station

1935 Fisher publishes 1st edition *Design of Experiments*

WWII George Box teaches himself statistics while enrolled in British army

1951 Box works with others to develop Response Surface Methods (RSM)

1950's W. Edwards Deming ignited application of statistical methods to manufacturing in Japan ("Quality Revolution")

- Genichi Taguchi develops Taguchi Arrays

Today Active research into DoE continues today and leverages the vast increase in computing power over the past few decades

George Box

What statistics is not, and what DoE is

- "There are lies, damn lies, and statistics." – Benjamin Disraeli
  - "…and someone statistically literate to know the difference."
- "The stats don't lie." – Unknown
  - "I shall try not to use statistics as a drunken man uses lampposts – for support rather than for illumination" – Andrew Lang
- "Love is never having to say you're sorry." – From the 1970 film *Love Story*
  - "Statistics is never having to say you're certain."
- Statistics is all about the average (mean)
  - **The purpose of DoE** is to know which factor(s) (X's) best explain the **variability** in the response variable (Y)

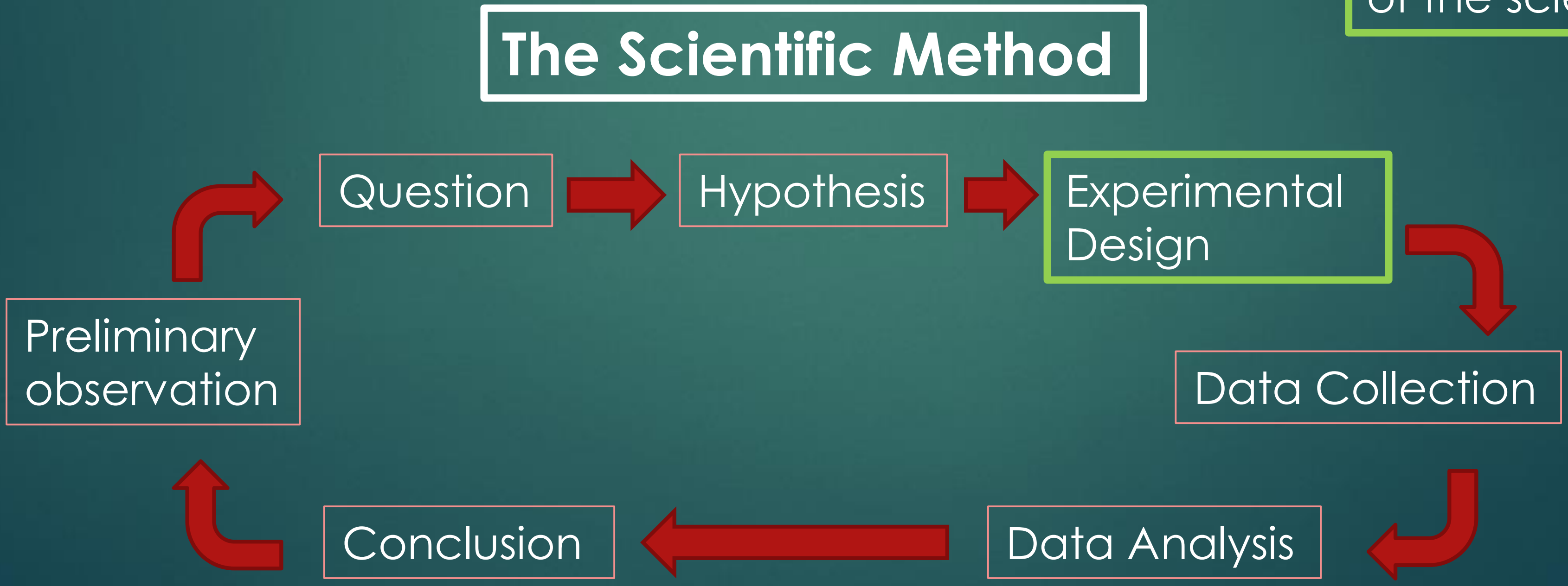
DoE, or designed experiments, create a detailed **plan** that can be used to answer a specific research question(s). This systematic plan considers sources of variability, constructs a statistical model, and seeks to <u>maximize learning in a resource constrained environment</u>. A successful DoE is <u>reproducible</u>, and results in <u>valid and defensible conclusions</u>.

DoE example: Shooting Koosh balls

A multitude of TechCon researchers want to know if there is a difference in the number of shots made when shooting a Koosh ball into a trash can with the left vs. right hand. Resources constrain the experiment to be run using a single individual, with three shots per hand.

How should the experiment be run?

Experimental design (DoE) touches every part of the scientific method



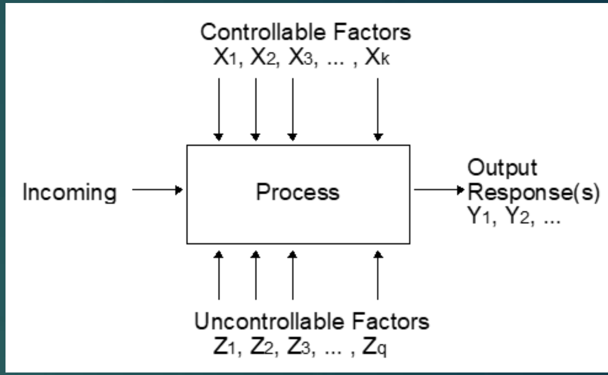
Scientific Method and DoE

1) **Clearly define the question and response variable**
   - Among SRC Conference attendees, is there a difference in the number of shots made when shooting a koosh ball into a trash can from a distance of 2 meters with the dominant vs. non-dominant hand? (dominant vs. non-dominant hand a much better question than R vs L)

2) State formal (statistical) hypotheses
   - Null hypothesis is no difference in average number of shots made between dominant and non-dominant hand
   - Alternative hypothesis is a difference exists in the average number
   - **Tests for equal variability** between dominant and non-dominant

Scientific Method and DoE

3) Design the Experiment
   - **What are the sources of variation and how will they be addressed?**
     - Control what you can, randomize what you can't, except…
     - …consideration needs to be given to realism of experiment/generalizability of results
     - Sources of variation present in the koosh ball experiment? How to address?
   - **What is the magnitude of the difference to be detected?**
     - This is called the test sensitivity
   - **What sample size is needed to reliably detect such a difference?**
     - Reliable, in this context, refers to the power of the test
   - How will the data be modeled/analyzed?
   - **Determine logistics of running experiment; standardize data collection**



As a rule of thumb, steps 1-3 should be ~60% of the researcher's work

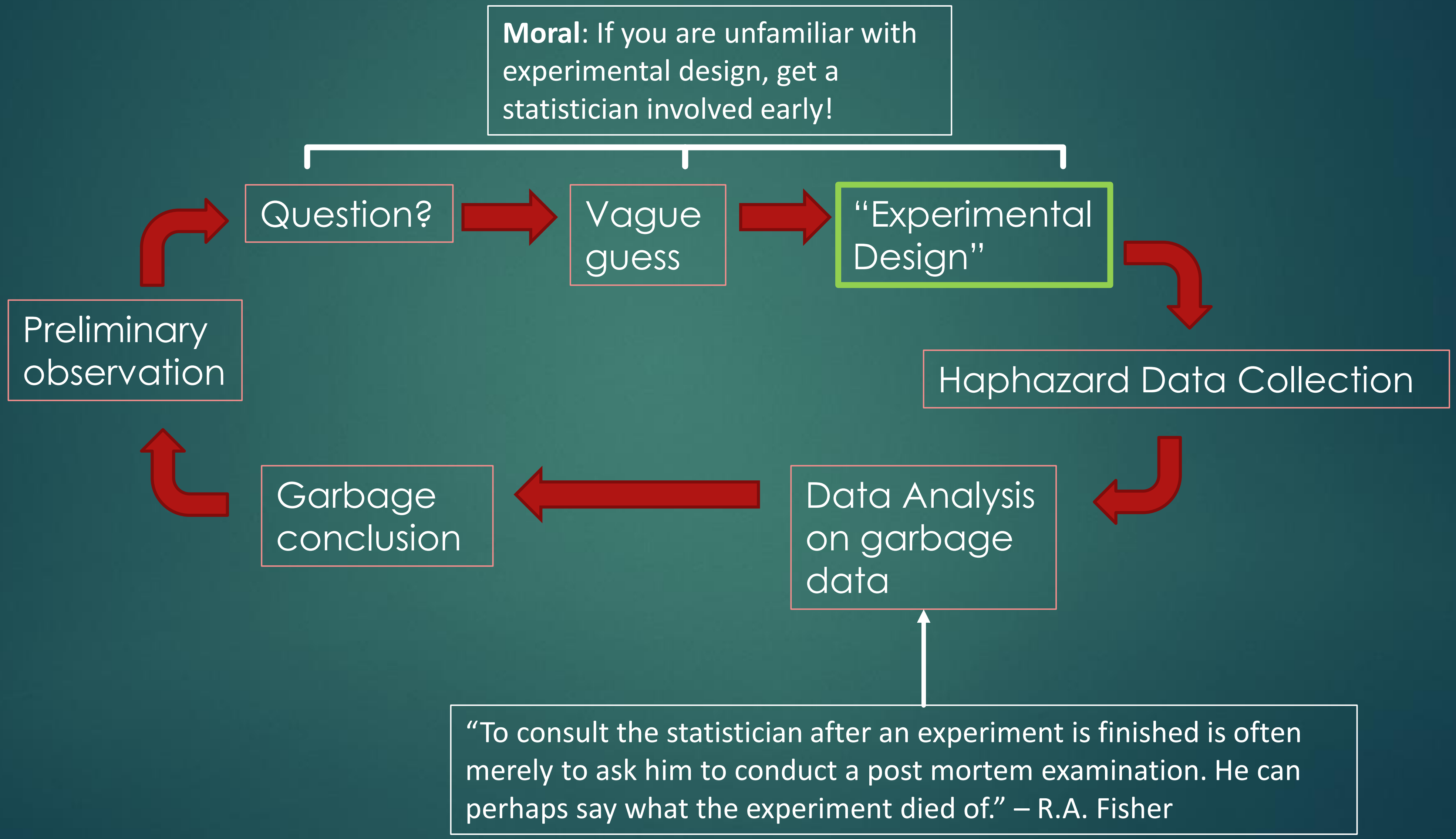
Scientific Method and DoE

4) Conduct the Experiment
    - Ensure data are being collected as specified in the experimental design

5) Analyze and interpret the data (this is the fun part!)
    - **Make sure model assumptions are met**; fit the model to the data, not the data to the model
        - "Far better an approximate answer to the right question than an exact answer to the wrong question" – John Tukey
    - **Look at the data**; do not rely solely on formal statistical analysis (p-values, etc.)

6) State conclusions
    - Stick with the null or go with the alternative?
    - What is the practical implication of the experiment?
    - **Results not turning out as expected/desired is OK!**
    - If results are counterintuitive from a first principles standpoint, review the data and the design to ensure there are no errors
    - **Even if these steps are followed carefully, validation experiments should be performed. Successful DoE's are reproducible!**

The (Un)scientific Method: Pitfalls



Common experimental designs

Single factor (one X), sometimes called OFAT (One Factor At a Time)

Two independent samples

Extension to more than two: ANOVA (Analysis of Variance)

Paired samples

Extension to more than two: Randomized Complete Block

Multi-factor (more than one X)

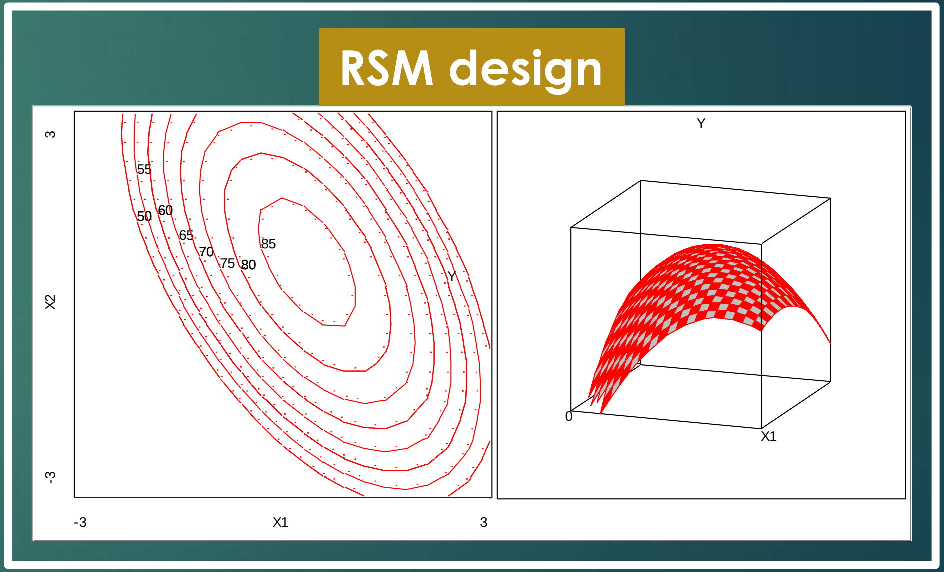
Two-way ANOVA, k-way ANOVA

Factorial, Fractional Factorial

Taguchi Arrays

Response Surface Methods (RSM)

- e.g. Central Composite Designs (CCD)



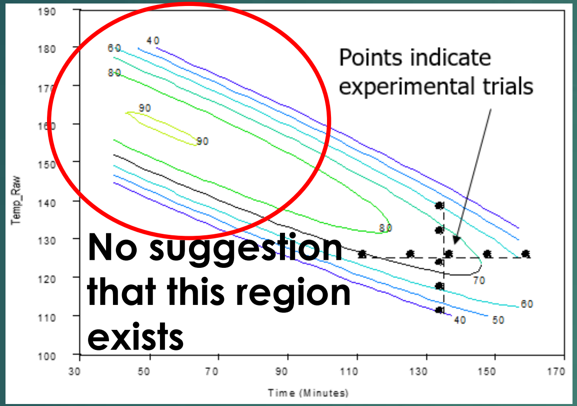
Others: Latin Square design, Incomplete Block, Split Plot, Definitive Screening Designs, Computer Experiments (e.g. climate change models)

OFAT or multi-factor experiments?

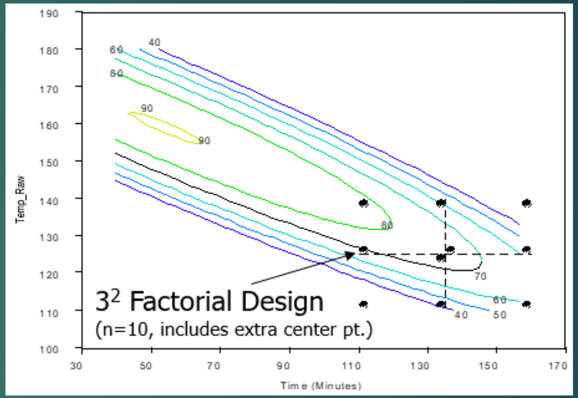
OFAT has its applications, but for most situations, multi-factor experiments are preferred. Why?

- Varying multiple X's simultaneously can show *interactions* among the X's

Two OFATs:



3^2 factorial:



- **Multi-factor DOE extracts more information than OFAT for a fixed amount of resources**
  - E.g. 2^k Full Factorial has a *smaller* standard error of the effect size as the # of factors *increases*. Compare the standard error of factorial to two independent sample t-test:

levels per factor

factors (#X's)

Full Factorial

$$\sigma_{effect} = \sqrt{\frac{\sigma_\varepsilon^2}{n2^{k-2}}}$$

Two sample t-test

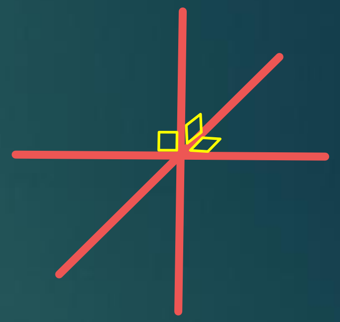
$$\sigma_{effect} = \sqrt{\frac{2\sigma_\varepsilon^2}{n}}$$

OFAT or multi-factor experiments?

So why aren't multi-factor experiments used more often?

- Lack of familiarity/comfort with multi-factor experimentation
- Misconceptions
  - "I don't have time to design an experiment"
  - "A designed experiment uses too many resources"
  - Mistaken belief that multi-factor experiments produce *hopelessly* confounded results
- Legitimate concerns
  - Multi-factor experiments require more pre-work than OFAT, more difficult logistically
  - Confounding will occur if design is not orthogonal; however, in many cases a proper DoE will provide a way to un-confound the X's with additional experimentation. If confounding is present, state the confounding structure and phrase conclusions accordingly

Orthogonal Vectors



Orthogonal: $\vec{A}^T \vec{B} = 0$

Shooting Koosh balls revisited

Data collection sheet (.jmp, .csv):

koosh_experiment_data_collection_sheet.jmp

koosh_experiment
lata_collection_she

Simulated Data (.jmp, .csv):

koosh_experiment_data.jmp

koosh_experiment
_data

Results (.jrn, .pdf):

Koosh_results

Koosh Results

Multi-factor DoE: Cake Baking

This is an example of a 2^2 factorial experiment.

levels/factor   # factors (#X's)

A manufacturer wants to know the time (X_1) and temperature (X_2) they should list on the box for their cake mix. A judge or judge(s) will rate the quality of the case on the scale described below.

Ranking Scale				
-10, -9,..-8,	…	-1, 0, 1,	…	..8, 9, 10
Soggy and Sticky		Just Right		Dry and Hard

N=12 cakes are used for the experiment (3 cakes per factor combination).

Sources of variability? (besides time and temperature)

Multi-factor DoE: Cake Baking

$$Model: \vec{Y} = X\vec{\beta} + \vec{\mathcal{E}}$$

Time	Time (coded)	Temp	Temp (coded)
25 min	-1	335	-1
35 min	1	365	1

X

Intercept (B0)	Time (B1)	Temp (B2)	Temp (B1*B2)
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

$\vec{\beta}$

Beta
B0
B1
B2
B12

		Baking Temperature	
		335°	365°
Baking Time	25 min	-3,-4,-2	-1,-3, 0
	35 min	-2, 0, 2	8, 8, 5

		Baking Temperature		
		335°	365°	Averages
Baking Time	25 min	-3	-1.33	-2.17
	35 min	0	7	3.5
	Averages	-1.5	2.84	

Figure 2.1 Average Rank vs. Baking Time

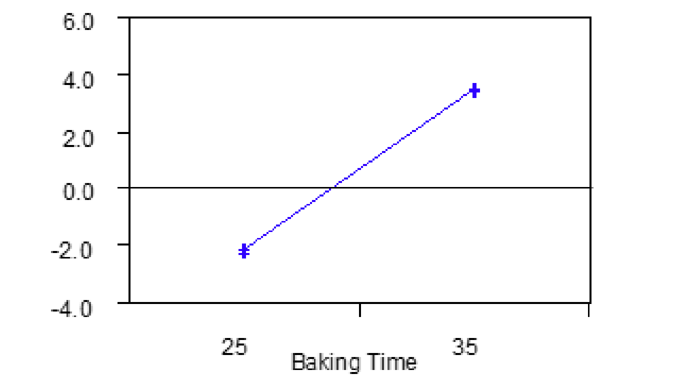


Figure 2.2 Average Rank vs. Baking Temperature

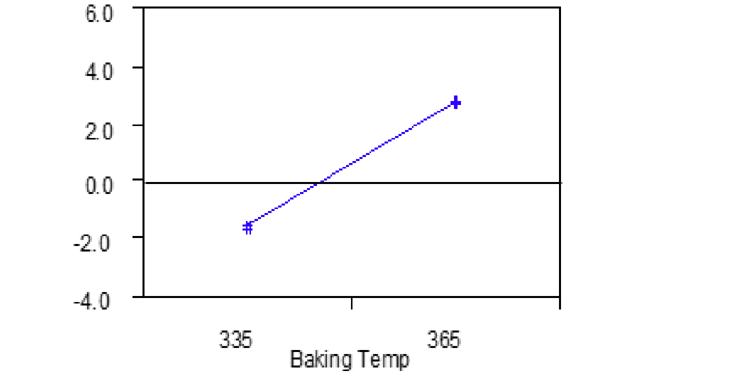
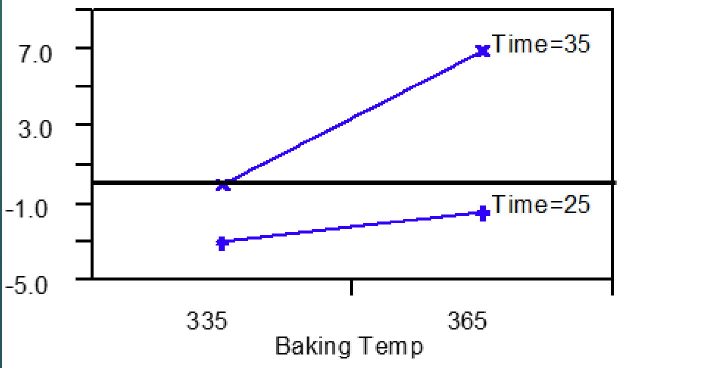


Figure 2.3 Average Rank vs. Time and Temp



$$Fitted\ Model: \hat{Y} = 0.67 + 2.83 * Time + 2.17 * Temp + 1.33 * Time * Temp$$

Concluding Remarks

- When planning an experiment, slow down and think it through carefully. If help from a statistician is needed, get him/her involved early!
- Educate your statistician on your topic. Statisticians can be more helpful with some background knowledge.
- "If you torture the data enough, it will confess". Look at the data curiously, without agenda, and let it tell the story
- Use Mutli-factor DoE's when possible (as opposed to OFAT)
- DoE is, and will continue to be, a cornerstone of the scientific method

Acknowledgements

My sincere thanks to Teri Utlaut, for sharing her expertise and providing valuable suggestions, to Scott Clendenning, for inviting me to speak at TechCon 2019 and helpful inputs, and to Larry Wrong, for his feedback.

I'd also like to acknowledge both Lisa Green and Leslie Faiers. Both were very welcoming and helpful for conference setup and accommodations.

References

Lady tasting tea

https://brainder.org/2015/08/23/the-lady-tasting-tea-and-fishers-exact-test/

https://en.wikipedia.org/wiki/Lady_tasting_tea

Ronald Fisher

https://en.wikipedia.org/wiki/Ronald_Fisher

https://pdfs.semanticscholar.org/a617/f75136d7645fdbcc60bac4ced1df19febedd.pdf

George Box

An Accidental Statistician, by Box, George. Published 2013 by Wiley Books

What statistics is not, and what DoE is

https://en.wikipedia.org/wiki/Lies,_damned_lies,_and_statistics

https://quoteinvestigator.com/2014/01/15/stats-drunk/

https://www.itl.nist.gov/div898/handbook/pmd/section3/pmd31.htm (definition of DoE)

OFAT and multi-factor DoE

http://mescal.imag.fr/membres/arnaud.legrand/teaching/2011/EP_czitrom.pdf