Randomization based perspectives of randomized block designs and a new test statistic for the Fisher randomization test

Design of Experiments: New Challenges, CIRM, France

Tirthankar Dasgupta
Department of Statistics and Biostatistics, Rutgers University
(joint with Peng Ding, UC Berkeley)

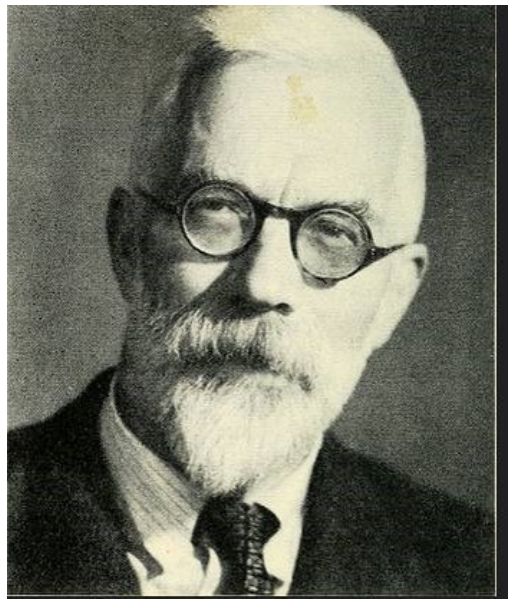
The randomization-based perspective of experimental design

- Long served as the foundation of experimental design.
- Seminal work by two stalwarts (Neyman 1923, Fisher 1925, 1935).
- Connection with survey sampling.
- Died down in the later half of the twentieth century
  - Lack of computational resources
  - Complicated asymptotics for complex designs

Potential outcomes - a framework that facilitates understanding the randomization perspective

- A framework for drawing statistical inference on causal effects of interventions.
- Widely used in the medical, biomedical, social and behavioral sciences.
- Also referred to as "the counterfactuals model" or the "Neyman-Rubin causal model" (Sekhon 2007) or simply the "Rubin causal model' (Holland 1986) in the field of causal inference.
- Notation introduced by Neyman (1923) and adopted by Rubin (1974) into a general framework for causal inference.

Historical perspectives



- Jerzy Neyman: originated the concept (1923) and introduced the first formal notation
- R. A. Fisher (1919): If we say this boy is tall because he has been well fed, we are suggesting that he might quite probably have been worse fed, and that in this case he would been shorter.

The potential and observed outcomes

Unit (i)	$Y_i(0)$	$Y_i(1)$	W_i	$Y_i^{\text{obs}} = W_i Y_i(1) + (1 - W_i) Y_i(0)$
1	$Y_1(0)$	$Y_1(1)$	W_1	$Y_1^{\text{obs}} = Y_1(W_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$Y_N(0)$	$Y_N(1)$	W_N	$Y_N^{\text{obs}} = Y_N(W_N)$
Average	$\bar{Y}(0)$	$\bar{Y}(1)$		

- Potential outcomes assumed to be fixed.
- Assignment vector $\boldsymbol{W} = (W_1, \ldots, W_N)^\top$ (binary stochastic, generated from a *known* probability distribution).

The potential and observed outcomes

Unit (i)	$Y_i(0)$	$Y_i(1)$	W_i	$Y_i^{\text{obs}} = W_i Y_i(1) + (1 - W_i) Y_i(0)$
1	?	$Y_1(1)$	$W_1 = 1$	$Y_1^{\text{obs}} = Y_1(1)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
N	$Y_N(0)$	?	$W_N = 0$	$Y_N^{\text{obs}} = Y_N(0)$
Average	$\bar{Y}(0)$	$\bar{Y}(1)$		

- For each unit, one potential outcome observed and the other missing.
- For each unit, can we "impute" the missing outcome under certain hypotheses?

Fisher randomization test (FRT) in a completely randomized design (CRD) setup

Unit	Treatment 1	Treatment 2	Treatment 3
1	8	?	?
2	?	18	?
3	?	12	?
4	10	?	?
5	?	?	9
6	?	?	5
Mean	$\overline{Y}_{.1}^{\text{obs}} = 9$	$\overline{Y}_{.2}^{\text{obs}} = 15$	$\overline{Y}_{.3}^{\text{obs}} = 7$

- Fisher's sharp null hypothesis is one of "no treatment difference for any individual."

$$H_0 : Y_i(1) = Y_i(2) = Y_i(3), \ i = 1, \ldots, N.$$

- FRT is a "stochastic proof by contradiction" (Rubin 2004) of Fisher's sharp null hypothesis.

FRT: Test statistic and its computation

Unit	Treatment 1	Treatment 2	Treatment 3
1	8	?	?
2	?	18	?
3	?	12	?
4	10	?	?
5	?	?	9
6	?	?	5
Mean	$\overline{Y}_{.1}^{\text{obs}} = 9$	$\overline{Y}_{.2}^{\text{obs}} = 15$	$\overline{Y}_{.3}^{\text{obs}} = 7$

- Observed $\boldsymbol{W} = (1, 2, 2, 1, 3, 3)$.
- Let's use $F = MS_T / MS_R$ statistic. Observed value from data is 3.71.
- Next step: derive randomization distribution of F.

FRT: Impute missing potential outcomes under H_0

Unit	Tr 1	Tr 2	Tr 3
1	**8**	8	8
2	18	**18**	18
3	12	**12**	12
4	**10**	10	10
5	9	9	**9**
6	5	5	**5**
Mean	$\overline{Y}_{.1}^{\text{obs}} = 9$	$\overline{Y}_{.2}^{\text{obs}} = 15$	$\overline{Y}_{.3}^{\text{obs}} = 7$

Computing the F statistic for all possible randomizations

$\mathbf{W} = (1, 1, 2, 2, 3, 3)'$

Unit	Tr 1	Tr 2	Tr 3
1	**(8)**	8	8
2	**(18)**	18	18
3	12	**(12)**	12
4	10	**(10)**	10
5	9	9	**(9)**
6	5	5	**(5)**
Mean	$\overline{Y}_{.1}^{\text{obs}} = 13$	$\overline{Y}_{.2}^{\text{obs}} = 11$	$\overline{Y}_{.3}^{\text{obs}} = 7$

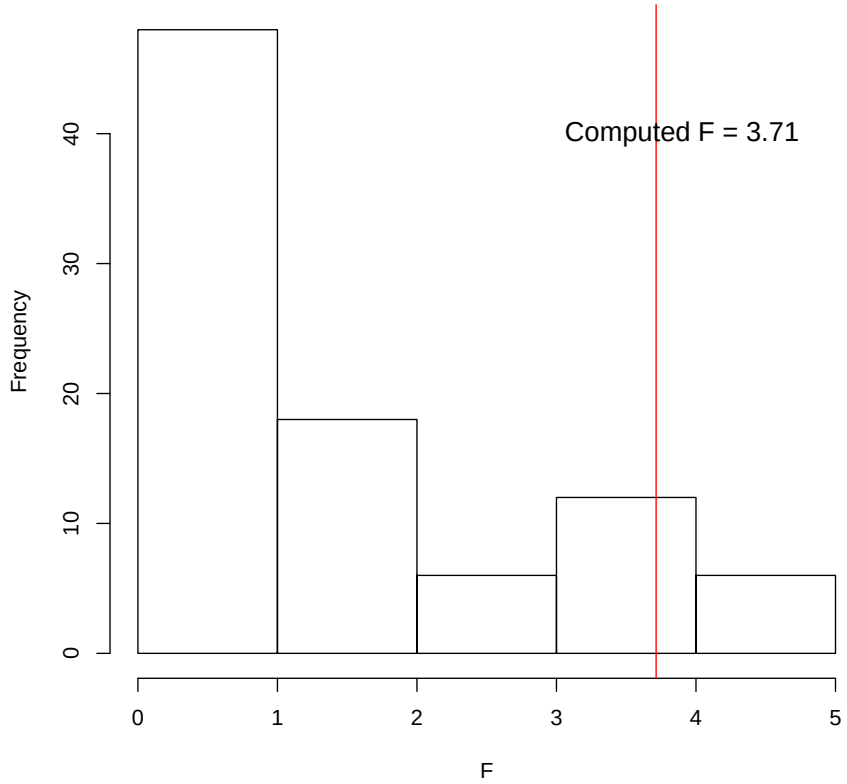
$F = 0.9333.$

$\mathbf{W} = (3, 3, 2, 2, 1, 1)'$

Unit	Tr 1	Tr 2	Tr 3
1	8	8	**(8)**
2	18	18	**(18)**
3	12	**(12)**	12
4	10	**(10)**	10
5	**(9)**	9	9
6	**(5)**	5	5
Mean	$\overline{Y}_{.1}^{\text{obs}} = 7$	$\overline{Y}_{.2}^{\text{obs}} = 11$	$\overline{Y}_{.3}^{\text{obs}} = 13$

$F = 0.9333.$

Randomization distribution of F (F_{rand}) under sharp null



- Computed p-value $= Pr(F_{rand} \geq F_{obs}) = 0.10$.

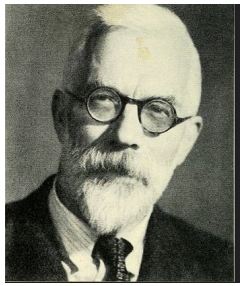
Why I like FRT

- Intuitive - analyze as you randomize. Easy to teach.
- Flexibility and broad applicability
  - continuous/binary response
  - any test statistic
  - a broad class of assignment mechanisms
  - multiple factors
- Always "valid" under the sharp null

Fisher's null and Neyman's null

- Fisher's null: Unit-level treatment effects are equal; Neyman's null: Average-level treatment effects are equal.
- Question: Randomization test is interesting, but the sharp null is boring. A more interesting hypothesis is the comparison of average treatment effects.
- Answer: Well ... maybe ... but randomization test works only under the sharp null.
- Question: Can the randomization test be used to test Neyman's null? Is F still an appropriate test statistic?

The infamous debate at the RSS in 1935

- Neyman: So long as the average yields of any treatments are identical, the question as to whether these treatments affect separate yields on single plots seems to be uninteresting and academic, and certainly I did not consider methods for its solution,
- Fisher: It [the null hypothesis that "the treatments were wholly without effect"] may be foolish, but that is what the Z-test was designed for, and the only purpose for which it has been used . . . I hope he [Neyman] will invent a test of significance, and a method of experimentation, which will be as accurate for questions he considers to be important [testing the average treatment effect being zero] . . .

Can FRT be used to test Neyman's hypothesis? Some findings for CRD

- Under Fisher's sharp null, under regularity conditions, the F statistic is approximately distributed as $F_{J-1,N-J}$ (known result).

Can FRT be used to test Neyman's hypothesis? Some findings for CRD

- Under Fisher's sharp null, under regularity conditions, the F statistic is approximately distributed as $F_{J-1,N-J}$ (known result).
- Under Neyman's null,
  - For balanced designs ($N_1 = \cdots = N_J$), $E(MS_R) \geq E(MS_T)$ [Using FRT with F may be conservative].
  - Special case: For balanced designs with *strict additivity* (all effects equal for all units) $E(MS_R) = E(MS_T)$ (Kempthorne 1955) - heuristic justification for using F (required?)
  - For unbalanced designs, $E(MS_R)$ may be larger or smaller than $E(MS_T)$ depending on the degree and nature of imbalance and the variances of the potential outcomes under treatments $1, \ldots, J$.

Can FRT be used to test Neyman's hypothesis? Some findings for CRD

- Under Fisher's sharp null, under regularity conditions, the F statistic is approximately distributed as $F_{J-1,N-J}$ (known result).
- Under Neyman's null,
  - For balanced designs ($N_1 = \cdots = N_J$), $E(MS_R) \geq E(MS_T)$ [Using FRT with F may be conservative].
  - Special case: For balanced designs with *strict additivity* (all effects equal for all units) $E(MS_R) = E(MS_T)$ (Kempthorne 1955) - heuristic justification for using F (required?)
  - For unbalanced designs, $E(MS_R)$ may be larger or smaller than $E(MS_T)$ depending on the degree and nature of imbalance and the variances of the potential outcomes under treatments $1, \ldots, J$.
- Thus using FRT with the F statistic to test Neyman's null may not control Type-I error.

Can FRT be used to test Neyman's hypothesis? Some findings for CRD

- Under Fisher's sharp null, under regularity conditions, the F statistic is approximately distributed as $F_{J-1,N-J}$ (known result).
- Under Neyman's null,
  - For balanced designs ($N_1 = \cdots = N_J$), $E(MS_R) \geq E(MS_T)$ [Using FRT with F may be conservative].
  - Special case: For balanced designs with *strict additivity* (all effects equal for all units) $E(MS_R) = E(MS_T)$ (Kempthorne 1955) - heuristic justification for using F (required?)
  - For unbalanced designs, $E(MS_R)$ may be larger or smaller than $E(MS_T)$ depending on the degree and nature of imbalance and the variances of the potential outcomes under treatments $1, \ldots, J$.
- Thus using FRT with the F statistic to test Neyman's null may not control Type-I error.
- For details, see Ding and Dasgupta (2018), *Biometrika*.

Some findings for CRD (contd.)

- Take weighted average of the sample means of the treatment groups

$$\bar{Y}_w^{\text{obs}} = \sum_{j=1}^{J} \frac{N_j}{s^2_{\text{obs}(j)}} \bar{Y}_\cdot^{\text{obs}}(j) / \sum_{j=1}^{J} \frac{N_j}{s^2_{\text{obs}(j)}}$$

- Consider the test statistic

$$X^2 = \sum_{j=1}^{J} \frac{N_j}{s^2_{\text{obs}(j)}} \left\{ \bar{Y}_\cdot^{\text{obs}}(j) - \bar{Y}_w^{\text{obs}} \right\}^2 .$$

- Sharp null: randomization distribution known; asymptotically $X \dot{\sim} \chi^2_{J-1}$.
- Average null: sampling distribution is complex; asymptotically

$$\Pr(X^2 \geq a) \leq \Pr(\chi^2_{J-1} \geq a).$$

- Randomization test using X^2 is exact under sharp null and conservative for the average null.

Unreplicated balanced RBD

Blocks	Treatment levels 1	2	3	$\cdots$	J	row mean	row variance
1	$Y_{.1}^{obs}(1)$	$Y_{.1}^{obs}(2)$	$Y_{.1}^{obs}(3)$	$\cdots$	$Y_{.1}^{obs}(J)$	$\bar{Y}_{.1}^{obs}(\cdot)$	s_1^2
2	$Y_{.2}^{obs}(1)$	$Y_{.2}^{obs}(2)$	$Y_{.2}^{obs}(3)$	$\cdots$	$Y_{.2}^{obs}(J)$	$\bar{Y}_{.2}^{obs}(\cdot)$	s_2^2
3	$Y_{.3}^{obs}(1)$	$Y_{.3}^{obs}(2)$	$Y_{.3}^{obs}(3)$	$\cdots$	$Y_{.3}^{obs}(J)$	$\bar{Y}_{.3}^{obs}(\cdot)$	s_3^2
$\vdots$							$\vdots$
K	$Y_{.K}^{obs}(1)$	$Y_{.K}^{obs}(2)$	$Y_{.K}^{obs}(3)$	$\cdots$	$Y_{.K}^{obs}(J)$	$\bar{Y}_{.K}^{obs}(\cdot)$	s_K^2
column mean	$\bar{Y}_{..}^{obs}(1)$	$\bar{Y}_{..}^{obs}(2)$	$\bar{Y}_{..}^{obs}(3)$	$\cdots$	$\bar{Y}_{..}^{obs}(J)$	$\bar{Y}_{..}^{obs}(\cdot)$	$s^2 = K^{-1}s_k^2$

- No replications, I units per block, J treatments, K blocks, $N = IK = JK$.
- Under null hypothesis of equality of average treatment effects, Neyman (1935):

$$E\{MS_R - MS_T\} = 0,$$

Sabbaghi and Rubin (2014):

$$E\{MS_R - MS_T\} = \Delta_F \geq 0.$$

New results

- The asymptotic randomization distribution of F is a linear combination of $J-1$ IID χ^2 variables and there is no guarantee that it is stochastically dominated by χ^2_{J-1}.
- Define:

$$\widehat{\boldsymbol{\delta}}_k = \left(\widehat{\tau}_{\cdot k}(1,2) \ldots \widehat{\tau}_{\cdot k}(1,J)\right)^\top$$

and the statistic:

$$Q = \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\delta}}_k\right)^\top \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\delta}}_k \widehat{\boldsymbol{\delta}}_k^\top\right)^{-1} \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\delta}}_k\right).$$

- Let $K \to \infty$ and assume certain regularity conditions. Under Neyman's null, the asymptotic distribution of Q is stochastically dominated by a χ^2_{J-1} random variable; under Fisher's null, the asymptotic distribution of Q is χ^2_{J-1}.
- Therefore, the Fisher randomization test using Q will guarantee the right probability of type-I error for Fisher's null and will be asymptotically conservative for Neyman's null.

Recap: advantages of randomization-based inference

- Intuitive - analyze as you randomize.
- Flexibility.
- Model free, but can be extended to model-based inference using a Bayesian approach (Key idea: obtain a probabilistic imputation model $p\left(\boldsymbol{Y}^{\text{mis}} | \boldsymbol{Y}^{\text{obs}}\right)$ and test model using posterior predictive checks).
- Bayesian extension permits super-population as well as finite-population inference.
- Reviving recondite connection with survey sampling.
- Excellent tool for analyzing data from complex modern BIG experiments, e.g., online experiments on social networks.

Some of my recent work on randomization-based inference

- Dasgupta, T., Pillai, N. and Rubin, D.R. (2015), "Causal Inference for 2^K factorial designs by using potential outcomes," *Journal of the Royal Statistical Society, Series B*, 77(4), 727–753.
- Espinosa, V., Dasgupta, T. and Rubin, D. B. (2016), "A Bayesian perspective on the analysis of unreplicated factorial designs using potential outcomes," *Technometrics*, 58, 62–73
- Ding, P. and Dasgupta, T. (2016), "A Potential Tale of Two by Two Tables from Completely Randomized Experiments", *Journal of the American Statistical Association*, 111, 157–168.
- Ding, P. and Dasgupta, T. (2018) "A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity," *Biometrika*, 105, 45–56.
- Zhao, A., Ding, P., Mukerjee, R. and Dasgupta, T., "Randomization-based Causal Inference from Split-Plot Designs," *The Annals of Statistics*, *in press*.
- Mukerjee, R., Dasgupta, T. and Rubin, D. B.,"Using Standard Tools from Finite Population Sampling to Improve Causal Inference for Complex Experiments," *Journal of the American Statistical Association*, *in press*.