Robert Moir: "A Monte Carlo Analysis of the Fisher Randomization Technique"

Experimental Economics 1998

Christian J. Meyer
European University Institute, Department of Economics

March 14, 2016
Topics in Experimental Economics (Schram/Gërxhani)

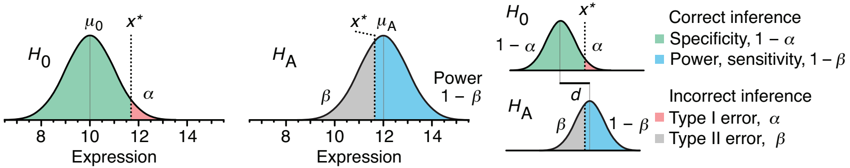
Overview of the Presentation

1. **Introduction**
   - Refresher on Statistical Hypothesis Testing
   - Parametric and Non-parametric Tests
2. **Fisher's Randomization**
   - Randomization or Permutation Tests
   - Example for Fisher's Exact Randomization Test
3. **Monte Carlo Results**
4. **Conclusion**
   - Critique, and Proposals

Statistical Hypothesis Testing

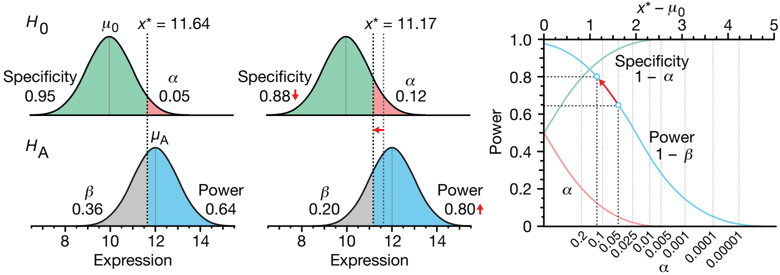
A refresher on how significance, statistical power, and sample size affect correct inference

- Paradigm to analyze data with a hypothesized relationship
  - Trying to find departure from an idealized **null hypothesis** H_0. Contrast with **alternative** H_A for distribution when null is false
  - **Experimental effect** d is difference between the distributions
  - Probability of false positives is called **significance level** or size
  - Probability of detecting the effect is called **statistical power**
  - False positives (Type 1 error, α) vs. false negatives (Type 2 error, β)



Statistical Hypothesis Testing: Trade-offs

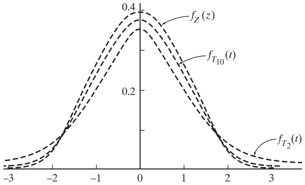
- Compromise: **specificity** (avoiding false positives, $1-\alpha$) vs **power** (avoiding false negatives, $1-\beta$)
  - An increase in power comes at the cost of more false positives



- Effect size and sample size similarly impact power ▸ Illustration
- **Moir studies this trade-off for three different tests**

Student's t-test for Difference of Means

- Before testing, we need to consider statistical assumptions about the observed sample: **independence, distribution**, ...
- Student t-test
  - **parametric** test
  - use for data randomly sampled from a **normally-distributed** population
  - two-sample t-test requires same variance in both
  - normality might be strong assumption
  - simulation evidence on small samples inconclusive



Mann-Whitney-Wilcoxon (MWW) Rank-Sum Test

A common alternative to the t-test to test difference of distributions

- There are two samples
  - H_0: Both samples have the same distribution
  - H_1: Observations in one sample tend to be larger than in the other
  - Requires random samples from population; independence within
- Rank each observation and compare rank totals of both samples; if there is no systematic difference, high and low ranks will be distributed relatively evenly
- Pro
  - **Non-parametric** i.e. distribution of test statistic under H_0 known
  - More efficient than t-test for distributions far from normal
  - Robust to outliers
- Con
  - **Less power** than parametric tests because we discard information

Fisher's Exact Randomization

- Method by R. A. Fisher (1935) for valid hypothesis test
  - without large samples
  - without probability model
  - purely based on physical act of **randomization**
- "Sharp" null: Assignment to treatment has absolutely no effect
- Idea: If null is true, **randomly shuffling around assignment** should produce same test statistic as real data
- How likely it is that we observe an effect "as extreme" as ours? Exact p-value from **number of possible permutations**:
  - In each permutation, calculate test statistic
  - Calculate share of permutations in which test statistic exceeds test statistic from real data

Example for Fisher's ER Means Test: Coffee at EUI

Minutes of concentration

Coffee $Y_i(1)$	*No coffee* $Y_i(0)$
7	0
8	2
11	5
30	9

1. Calculate test statistic: sample average for both groups
   - $\overline{y}_1 = 14$ and $\overline{y}_0 = 4$, difference $d = 10$
2. How many possible ways are there of shuffling around the data?
   - Combination without replacement or "n choose k": $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
   - Here eight values choose sets of four: $\binom{8}{4} = 70$
   - In these 70 combinations, how often is difference in means ≥ 10?
   - Simple counting... turns out 3 times
   - If original assignment was random, p-value $3/70 = 0.043$

Primer on Monte Carlo Simulation Techniques

- Inference often relies on parametric assumptions & asymptotics
  $\rightarrow$ What happens if these are not met and in small samples?
- Monte Carlo studies can characterize **performance of tests**
  - Randomly generate samples with known characteristics and size
  - For each replication (say 10,000), record test performance
- Moir measures performance in two dimensions

1. Size / type 1 error
   - **Nominal** (α) from asymptotic results, i.e. significance we choose
   - **Actual**, i.e. the fraction for which test falls in rejection region
2. Power
   - Fraction of actual replications for which null hypothesis is rejected
     (Note: This requires simulation of alternative hypothesis!)

Simulation Results for Size and Power

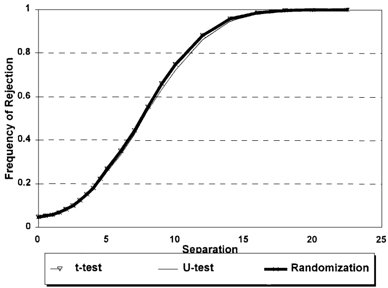
- Simulation: samples relevant for experimental research
  - Small sample (16 observations), constant size
  - Normal errors as baseline
  - Mixture distributions, fat tails, skewed tails
  - Alternative hypothesis for a range of effect sizes
- Two-sided hypothesis tests to identify treatment effect
- General findings
  - **ER means**: Power at least as good as t-test in most cases
  - **t-test**: Lower power than ER means when data is not normal
  - **MWW**: Generally lower power than both ER means and t-test

Discussion of Results for Each Test

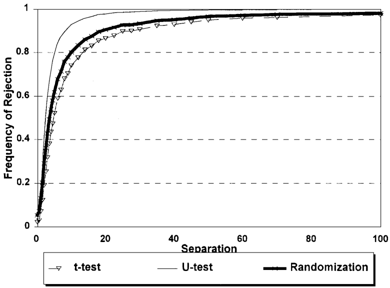
- t-test
  - Departures from normal lead to invalid test statistics (type 1 error higher than should be), particularly with uniform and mixed normal
  - Serious power problems in Cauchy distribution
- MWW
  - Generally lower power than ER means and t-test (expected given it is non-parametric and uses less information)
  - More power than ER means and t-test in Cauchy distribution
- ER means test
  - Always has correct size because distribution is generated from the sample as opposed to asymptotic results
  - Power at least as good as t-test in most cases

Power Graphs for Normal and Cauchy

Normal distribution $N(0, 50)$



Cauchy distribution (median 0)



- NB: Different scale on x-axis!
- Under Cauchy, t-test rejects null less often than other two

Fisher Randomization 75 years Later

- Paper has shown favorable performance of ER means test
- ER means test attractive for fewer and weaker assumptions
  - **Distribution-free** under H_0, i.e. nonparametric
  - No assumptions on sampling from some notional population
  - Exact p-values and **no asymptotics** required
  - Works well in small samples and other “low information” settings
- Lends itself to modern lab experiments and randomized studies
- Today computationally easily feasible
- Today has been expanded to provide confidence intervals, deal with instrumental variables, include covariates, etc.

Thoughts on the **Paper** and **Fisher's ER**

- Monte Carlo might not be the best approach to (re-)introduce Fisher's ER technique to discipline
  - MC typically useful to undertand statistics under realistic data conditions. What is realistic?
  - Except for Cauchy case, hard to see systematic power differences between t-test and ER means test
  - Detailed formal discussion of assumptions, implementation, and assumptions may have been useful
  - Of course not really a fair criticism...
- ER technique has **very sharp null** $H_0 : Y_i(1) = Y_i(0) \; \forall i$
  - Rare that we want to have a purely confirmatory test of this
  - Maybe overly restrictive?
- How "random" is data? Stratification?

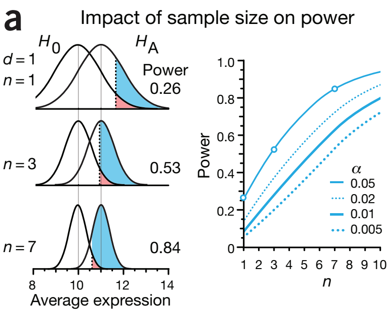
A Monte Carlo Analysis of the Fisher Randomization Technique

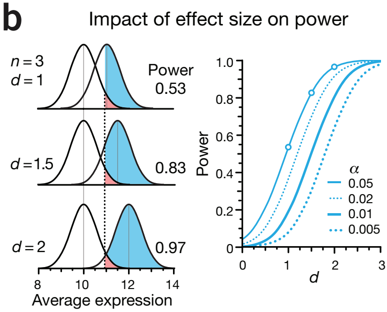
Robert Moir *Experimental Economics* 1998

Christian Johannes Meyer
European University Institute, Department of Economics
✉ christian.meyer@eui.eu
@chrmeyer

EUI Seminar: Topics in Experimental Economics

Sample Size, Effect Size, Statistical Power





Back

Charts on this slide and the previous two slides adapted from: Krzywinski and Altman. (2013). “Points of significance: Power and sample size” *Nature Methods*. doi:10.1038/nmeth.2738.

Monte Carlo Results on Test Size

Error distribution	Test	Level of significance **0.1000** (0.1049)[a]	**0.0500** (0.0536)[a]	**0.0100** (0.0116)[a]
Normal high variance	t-test	0.0973	0.0477	0.0095
	U-test	0.0830	0.0495	0.0075
	ER means test	0.0972	0.0474	0.0095
Uniform	t-test	0.1000	0.0531	0.0122[b]
	U-test	0.0826	0.0520	0.0067
	ER means test	0.0993	0.0508	0.0101
Mixed normal	t-test	0.1028	0.0564[b]	0.0137[b]
	U-test	0.0852	0.0527	0.0073
	ER means test	0.1024	0.0545[b]	0.0111
Sum of uniform + normal	t-test	0.0928	0.0407	0.0060
	U-test	0.0815	0.0509	0.0074
	ER means test	0.0989	0.0494	0.0095
Logistic	t-test	0.1045	0.0483	0.0076
	U-test	0.0878	0.0531	0.0069
	ER means test	0.1060[b]	0.0539[b]	0.0098
Cauchy	t-test	0.0574	0.0209	0.0022
	U-test	0.0832	0.0525	0.0079
	ER means test	0.1027	0.0523	0.0105
Extreme value	t-test	0.0968	0.0429	0.0063
	U-test	0.0873	0.0519	0.0062
	ER means test	0.1046	0.0519	0.0091
Exponential	t-test	0.0957	0.0458	0.0083
	U-test	0.0822	0.0492	0.0075
	ER means test	0.0968	0.0477	0.0095

Back

Monte Carlo Results on Power

Distribution	Comments
Normal[a]	All tests exhibit satisfactory size results. t-test and ER means test: track each other in terms of power, U-test less powerful.
Uniform	t-test: real size > nominal size at 1 percent level of significance. t-test and ER means test: track each other in terms of power, t-test slightly more powerful at 1 percent level (but invalid). U-test: uniformly less powerful.
Mixed normal	t-test: real size > nominal size at 5 percent and 1 percent levels. ER means test: real size > nominal size at 5 percent level, but t-test rejects true null more often than ER means test. t-test and ER means test: track each other in terms of power, t-test slightly more powerful at 1 percent level (but invalid). U-test: uniformly less powerful.
Sum of normal + uniform	All tests exhibit satisfactory size results. t-test and ER means test: track each other in terms of power. U-test: uniformly more powerful at 10 percent and 5 percent levels, but uniformly less powerful at 1 percent level.
Logistic	ER means test: real size > nominal size at 10 percent and 5 percent levels (type I error). t-test and ER means test: track each other in terms of power. U-test: slightly more powerful at 5 percent level, but less powerful at 1 percent level.
Cauchy	All tests exhibit satisfactory size results; however, t-test rejects significantly less often than U-test or ER-test. t-test: uniformly less powerful than ER-test. U-test: more powerful than other tests at 10 percent or 5 percent levels, but less powerful than ER means test at 1 percent level.
Extreme value	All tests exhibit satisfactory size results. t-test and ER means test: track each other in terms of power. U-test: more powerful than other tests at 10 percent or 5 percent levels, but less powerful at 1 percent level.
Exponential	All tests exhibit satisfactory size results. t-test and ER means test: track each other in terms of power. U-test: slightly less powerful at 10 percent level, more powerful at 5 percent level, and considerably less powerful at 1 percent level.

Back