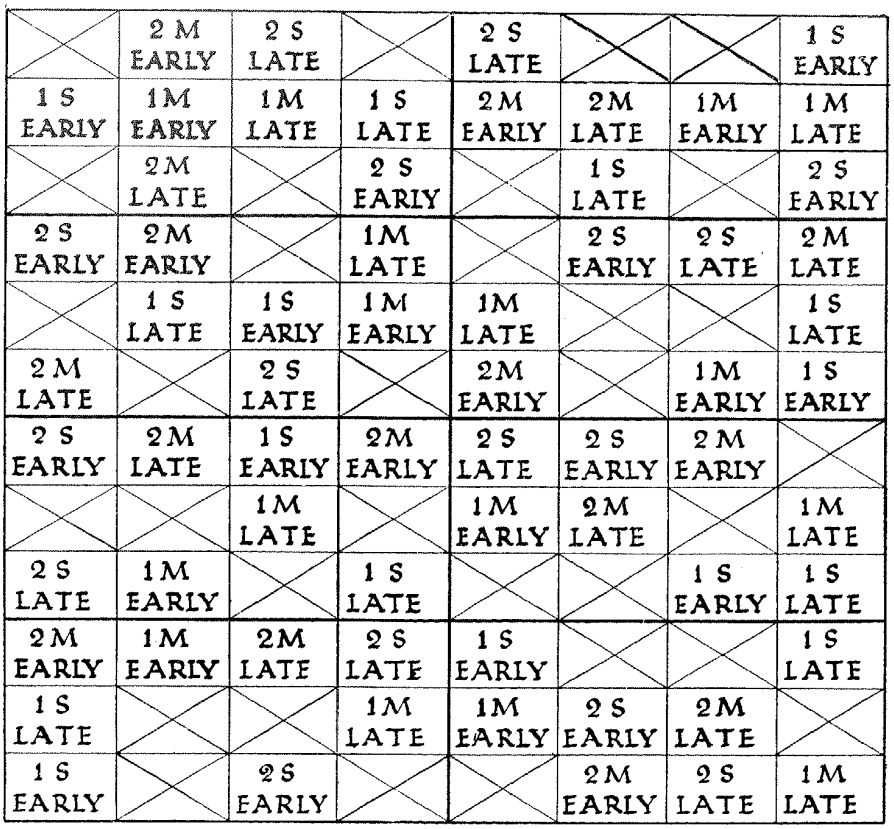
Field Experiments in Economics

Igor Asanov

April 11, 2018



This handout[1][2] summarizes the lectures slides. Please note that the handout is not very useful if you do not attend the class. The handout is also not a substitution for the book. The course is built around Glennerster and Takavarasha book: "Running Randomized Evaluations: A Practical Guide".

Homepage: http://www.igorasanov.com/teaching.html

Literature:

! Glennerster and Takavarasha, "Running Randomized Evaluations: A Practical Guide", 2013, Princeton University Press.

- Duflo, Glennerster, and Kremer, "Using randomization in development economics research: a toolkit", 2008, Chapter 61 in T. Paul Schultz and John Strauss, eds., Handbook of Development Economics, Vol. 4, Amsterdam. Elsevier. pp. 3895-3962.
- Halpern, D., "Inside the Nudge Unit: How small changes can make a big difference" , 2016. Random House.
- List, J. and Gneezy, U., "The why axis: hidden motives and the undiscovered economics of everyday life.", 2014. Random House.
- Stock and Watson, "Introduction to Econometrics", 2015, Pearson/Addison Wesley.
- Angrist and Pischke "Mastering 'Metrics: The Path from Cause to Effect", 2015, Princeton University Press.

Software:

We will use **R** for most of the exercises: Here is the link on the homepage of **R**.

In the class I use **RStudio** as a front end and I would recommend you to install it too - it greatly simplifies workflow in **R**.

[1] © Igor Asanov. This handout as derivative of "Running Randomized Evaluations: A Practical Guide" and it is shared with kind persmission of Rachel Glennerster, http://runningres.com.

[2] The cover picture is adopted form Fisher, R.A., 1926 "The arrangement of field experiments. In Breakthroughs in statistics." and used under Open Government Licence, HMSO.

Contents

1 Introduction

1.1 Economic Methodology

- **Theory:** Various types of tautology.
- **Empirical studies – Observational studies:** Various types of statistical analysis without intervention in the data generating process.
- **Experiment:** Purposeful intervention in the data generating process.

1.2 Theory

A theory is a tautology.

Properties:

- Internal correctness
- Testable
- Simple

A theory allows to make models.

Like a good map, a good model provides a **simple**, and, hence, **inaccurate and imprecise** representation of the world.

"**A good model** in economic theory, like a good fable, **identifies a number of themes and elucidates them**. **We perform thought exercises that are only loosely connected to reality** and that have been stripped of most of their real-life characteristics. However, in a good model, as in a good fable, **something significant remains**." Rubinstein (2006)

Theory makes a predictions – How to test them?

Samuelson and Nordhaus, (1985) Principles of Economics, p. 8:

"**Economists** ... cannot perform the controlled experiments ... because **they cannot easily control other important factors**. Like astronomers or meteorologists, they generally **must be content largely to observe.**"

→ Economists use observational studies without possibility to control for *all* important factors.

Why do we want to have control?

1.3 Fundamental Problem of Casual Inference

Road Not Taken

"Two roads diverged in a yellow wood,

And sorry **I could not travel both**

And be one traveler, long I stood

And looked down one as far as I could

To where it bent in the undergrowth;

...

Two roads diverged in a wood, and

I took the one less traveled by,

And that has made all the difference."

Robert Frost, 1920

Holland, (1986) Statistics and Causal Inference

Casual effect of $T_i = 1$ on unit i (relative to $T_i = 0$): $Y_{1,i} - Y_{0,i}$

The problem: It is impossible to observe the value of $Y_{1,i}$ and $Y_{0,i}$ on the same unit, therefore, it is impossible to observe the effect of $T_i = 1$ on unit i.

1.4 Observational Research

What is the effect of entrepreneurial education on entrepreneurship?

- Does entrepreneurial education increase start-up rate?
- Are those start-ups more efficient?
- Is entrepreneurial education cost-effective? Does investment of €1 in entrepreneurial training return more than €1 to the economy?

1.4.1 Qualitative Impact Evaluation

Methods: Direct observation, open-ended interviews...

Assumptions:

1. Evaluator have a good understanding of contrafactual
2. Participants report the information in an objective manner
3. Evaluator summarizes the information in an objective manner

Advantages: Richness of information.

Disadvantage: Subjective, no comparison group, absence of quantitative measures.

Example: Did you get entrepreneurial skills during the training? (European Commission, 2015).

Can you predict your mark at the exam?

1.4.2 Before and After Comparison

Method: Compare the outcomes before and after the program.

Assumptions:

1. Stability of Environment
2. Stability of subjects characteristics
3. No rebound effect

Example: Assessing the impact of entrepreneurship education programmes: a new methodology (Fayolle et al. 2006).

Method: Compare the response on the questionnaire about attitudes towards entrepreneurship before and after the course on entrepreneurship.

Findings: Increase in entrepreneurial intentions.

Questions:

1. Can we attribute the changes only to the program?
2. Can we say that the subjects would not change their attitudes without the program?
3. Can we say that subjects were in the best condition of the course?

1.4.3 Multivariate Regression

Method: OLS regression with treatment and "control" variables that can explain outcomes of the program.

$$Y = \beta_0 + \beta_T T + \sum_{i=1}^{k} \beta_{Ci} C_i + u$$

Assumptions:

1. Strict exogeniety, $E(u_i|T_i = t) = 0$
2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$

Example: Entrepreneurship among business graduates: does a major in entrepreneurship make a difference?(Kolvereid and Moen, 1997).

Method: OLS with dependent variable as a start-up rate.

Findings: Major in entrepreneurship has a positive association with start-up probability.

1. Is major in entrepreneurship is exogenous to the willingness to start-up rate? Omitted variable bias?
2. Could it be that those who decided to study entrepreneurship wanted to start the business anyways? Selection bias?
3. Is the entrepreneurial intentions normally distributed? Heteroskedasticity?

1.4.4 Regression Discontinuity Design

Method: Exploit cut-off(threshold) rule

Key Assumption:

1. $E[Y_{0,i}|x_i]$ and $E[Y_{1,i}|x_i]$ continuous on x_i and x_o.

Drawback: We estimate the effect close to cut-off.

Example: Can Entrepreneurial Activity be Taught? Quasi-Experimental Evidence from Central America (Klinger and Schündlen, 2011).

Method: Exploit the fact that a number of applicants take the training program based on the score of the application that describes their business idea.

Findings: Business training significantly increases the probability that workshop participant starts a business or expands an existing business.

Questions:

1. Can we say that other variables did not drastically change for those who were admitted to the program? What about confidence about their business idea?
2. Can we say that the program was effective for those whose score is very high?

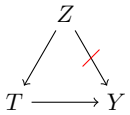
1.4.5 Instrumental Variable

Method: Use variable Z – the instrument – that affects the variable of interest T but does not lead to change in outcome Y (aside from indirect route via T) .

Assumptions:

1. Relevancy, $cor(Z, T) \neq 0$
2. Exogeniety, $cor(Z, u) = 0$

Examples of instrument: Lottery, weather conditions, proximity to the university.



Example: The impact of entrepreneurship education on entrepreneurship skills and motivation (Oosterbeek et al., 2010).

Method: Distance to the university as an instrument to estimate the effect of the entrepreneurship program.

Findings: Program *negatively* affects the intentions to become entrepreneur.

Questions:

1. Can we say that student attended the entrepreneurial class because they were enrolled in one university but not another?
2. Can we say that students choose the university only because of the distance to it?

1.4.6 Meta-Analysis

Method: Aggregate the information across studies.

Assumptions:

1. Outcomes and "treatment" variables are "identical".
2. Studies use similar specifications.

Example: Examining the formation of human capital in entrepreneurship: A meta-analysis of entrepreneurship education outcomes (Martin et al., 2013)

Method: Meta-analysis of literature on results of entrepreneurial education.

1. Are the outcome and "treatment" variables "identical" across studies?
2. Do the studies use the same specifications?

Table 1: Frequency of Variables

Entrepreneurship outcomes	K
Nascent behavior	1
Start-up	6
Entrepreneurship performance	9
Success (duration)	1
Success (financial)	8
Personal income from owned business	1

Table 2: Meta-analysis of Entrepreneurial Education on Entrepreneurship Outcomes

	Effect of Education on Entrepreneurship Outcomes				
	W. mean	SD	K	N	95% CI
Overall	0.159	0.096	13	10,524	0.107-0.211
Without large samples	0.207	0.050	10	2806	0.176-0.238
Start up	0.124	0.082	6	6706	0.058-0.190
Performance	0.166	0.125	9	5790	0.084-0.248

1.5 Field Experiments

1.5.1 Why to Randomize?

Method: Randomly assign people to the treatment and control groups in the naturally occurring environment.

Why to randomize?

Random assignment guarantee that treatment T is independent from potential outcome Y, hence, we can estimate the **average treatment effect**.

$$T \perp\!\!\!\perp Y \Rightarrow E[Y|T=1] - E[Y|T=0] = E[Y_{1,i} - Y_{0,i}]$$

Example: Growing America Through Entrepreneurship (GATE).

Method: Approx. 4000 people are randomly assigned to the groups that receives free entrepreneurship training (treatment) and not (control).

Table 3: Treatment Group

Statistic	N	Mean	St. Dev.
Age	2,094	42.032	10.291
Highest grade completed	2,094	14.389	2.208
Share of Females (♀)	2,094	0.528	0.499
Share of Asian	2,094	0.046	0.210
Share of White	2,094	0.552	0.497
Share of Black	2,094	0.305	0.460
Share of Unemployed	2,094	0.363	0.628

Table 4: Control Group

Statistic	N	Mean	St. Dev.
Age	2,103	42.727	10.308
Highest grade completed	2,103	14.515	2.240
Share of Females (♀)	2,103	0.543	0.498
Share of Asian	2,103	0.043	0.226
Share of White	2,103	0.553	0.507
Share of Black	2,103	0.304	0.470
Share of Unemployed	2,103	0.327	0.684

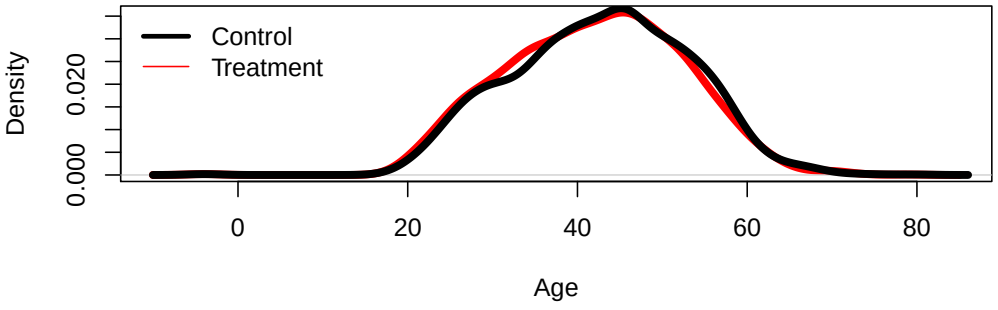


Figure 1: Age Density Plot

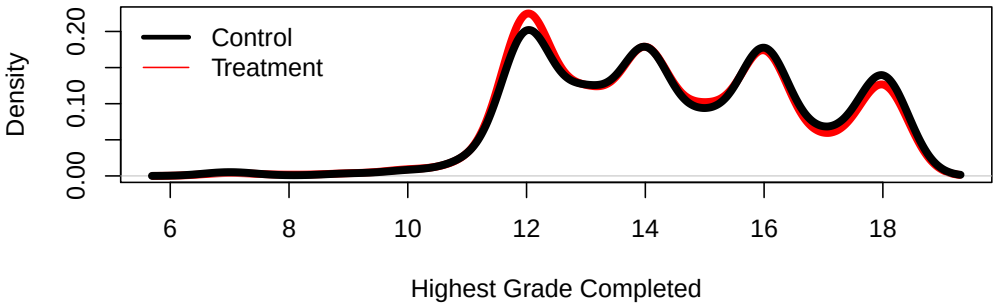


Figure 2: Grade Density Plot

Table 5: Outcomes of GATE project after 6 month

	Dependent variable:			
	Business Plan	Bussines Loan	Start-up	Household Income
Treatment	0.128***	−0.003	0.019	−1,264.170
	(0.017)	(0.008)	(0.021)	(1,417.316)
Constant	0.372***	0.063***	0.331***	43,775.440***
	(0.012)	(0.006)	(0.015)	(1,006.023)
Observations	3,438	3,447	2,052	2,743
Adjusted R^2	0.016	−0.0002	−0.0001	−0.0001
Note:				*p<0.1; **p<0.05; ***p<0.01

1.5.2 Limitations

- Uncontrolled parameters
- Expensive
- Long
- Generalizability?
- Not always appropriate
- Ethical Reasons

When is field experiment inappropriate?

1. **Impact evaluation is not needed.**
   For instance: Textbooks are not going be used.
2. **Macroeconomic questions.**
   How to evaluate the effect of exchange rate or interest rate?
3. **General Equilibrium Theory.**
   If we compare the effect in the whole system, it is hard to imagine a comparison group.

1.5.3 Ethical Considerations

Problem:

Syphilis inoculation project in Guatemala 1946-1948:

- 696 subjects (men in the Guatemala National Penitentiary, army barracks, men and women in the National Mental Health Hospital).
- Prostitutes with the disease were used to infect subjects, but also direct inoculation.
- Subjects then received penicillin.

Belmont Report, 1979:

1. Boundaries Between Practice & Research

2. Basic ethical principles:
   (a) Respect for person
   (b) Benefice
   (c) Justice
3. Applications:
   (a) Informed consent
   (b) Assessment of risk and benefits
   (c) Selection of subjects

Also, consult with your Institutional Review Board.

1.6 Summary

- **Economic Methodology:**
  Theory, observational research, experiment.
- **Fundamental problem of casual inference:**
  Absence of contrafactual
- **Observational research:**
  Requires strong assumptions.
- **Field Eeperiment:**
  Random assignment guarantee that treatment T is independent of potential outcome Y.

1.7 Exercises

1. **Methods I:** Which research methods do we have in economics?
2. **Methods II:** What is the fundamental problem of casual inference? Solutions?
3. **Observational Research I:** What do we have to assume when we use qualitative methods? What are the advantages of this method? Disadvantages?
4. **Observational Research II:** What do we have to assume when we use before and after comparison?
5. **Observational Research III:** What are the assumptions of multivariate regression analysis?
6. **Install R and Rstudio.**
   Download R from https://www.r-project.org/ and install it.
   Download RStudio https://www.rstudio.com/ and install it.
7. **Observational Research III: Crime rate.**

```r
# Install and call the package 'Ecdat'.
install.packages("Ecdat")
require(Ecdat)

# Use the data 'Crime'
data(Crime)
`?`(Crime)

# Plot the crime rate with respect to number of policemen per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
    pch = 18, data = Crime)


# Estimate simple linear relation between crime rate and number of policemen
# per capita.
lmcrime <- lm(crmrte ~ polpc, data = Crime)
summary(lmcrime)


# Plot the crime rate with respect to number of policemen per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
    pch = 18, data = Crime)
abline(lmcrime, col = "red")
```

How do you interpret this result? What are the policy implications?

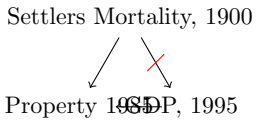
8. **Observational Research III: Health Condition.**
   - Call the package `Ecdat` and use the data `DoctorAUS`. What is the data set about?
   - What is the influence of number of doctor visits (`doctorco`) on the number of illness in past 2 weeks (`illness`)? Visualize this relation.
   - Estimate a simple linear relation between the number of doctor visits and the number of illness in past 2 weeks. Visualize regressions output.
   - How do you interpret this result? What are the policy implications?

9. **Observational Research IV:** What do we have to assume when we use regression discontinuity design?

10. **Observational Research V: Effect of institution on economic development.** The Colonial Origins of Comparative Development: An Empirical Investigation. (Acemoglu et al. , 2001).

	Dependent Variable	
	GDP per Capita, 1995	
Variables	OLS	**IV**
Property Rights Protection, 1985-1995	0.52***	0.95***
	(0.06)	(0.16)
Observations	64	64

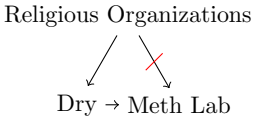
The instrumental variable specifications use settlers mortality in 1990.



	Dependent Variable	
	Meth Lab Seizures per 100,000	
Variables	OLS	**IV**
Alcohol Prohibition	2.01***	4.99**
	(0.60)	(1.69)
R-squared	0.17	0.14
Observations	889	889

11. **Observational Research V: Relation between meth lab seizures and alcohol prohibition in Kentucky (U.S.A).** Breaking Bad: Are Meth Labs Justified in Dry Counties? (Fernandez et al. , 2015).

The instrumental variable specifications use religious organization membership for 1936 as instruments.



12. **Observational Research VI:** What do we have to assume when we use meta-analysis?

13. **Presentation:** List, J. a., 2011. *Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off. Journal of Economic Perspectives*, 25(3), pp.316.

    McKenzie, D. and Sansone, D., 2017. Man vs. machine in predicting successful entrepreneurs: evidence from a business plan competition in Nigeria.

2 Asking Right Questions

2.1 Questions for Field Experiments

1. **Strategic Questions.**

   What do we want to achieve?

2. **Descriptive Questions.**

   What are the needs?

3. **Process Questions.**

   How well is the program being implemented?

   - Methodology
   - Cost-effectiveness

4. **Impact Questions.**

   Did it work?

2.2 Needs Assessment

Questions to asses the needs:

- Whom we target?
- What problems do they face?
- Why?
- What are the existing solutions?
- Which problems are left?

Methods to asses the needs:

- Information from other programs, Literature review
- Qualitative interviews
- Quantitative surveys

When descriptive needs assessment can sufficient:

- No "real" problem.
- The problem has low priority
- Different cause of the problem?
- Insufficient conditions to run the program

2.3 Process Evaluation

2.3.1 Methodology

☐ **Assess operations on paper.**

Articulate tasks that you plan to perform e.g. provide a course, give handouts, send e-mails, provide pills...

☑ **Follow paper trails.**

Check paper records e.g. documents about attendance in the course,signs on paper if people got pills, handouts ...

☑✓ **Assess operations in field.**

Check correspondence of paper records and on-the ground check e.g. randomly visit the course and check if number of students correspond to number written in paper records.

2.3.2 Cost-effectiveness

1. Gather the information about alternative programs.
2. Compare the outcomes of the programs under different scenarios.
   - Program going to bring to low benefits
     → Do not implement
   - Program is good anyways
     → Perhaps, better to investigate other program with uncertain outcomes

Cost-effectiveness: Example

What is effect of different programs on test scores?

Table 6: Effect of different programs on test scores

	Test Score (Standard Deviations)		
Program	Lower Bound	Mean	Upper Bound
Micronutrients	-0.13	-0.071	-0.011
School Meals	0.011	0.039	0.067
Unconditional Cash Transfers	0.021	0.079	0.14
Conditional Cash Transfers	0.083	0.15	0.22
Contract Teachers	0.11	0.16	0.22
Scholarships	-0.026	0.19	0.41

Source: www.aidgrade.com

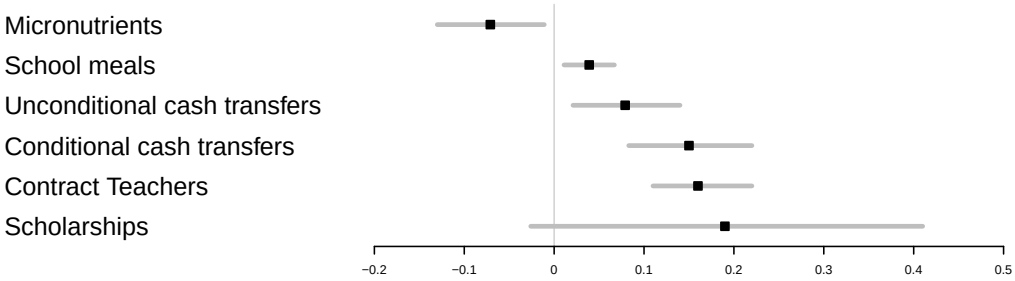


Figure 3: Impact of program on test score.

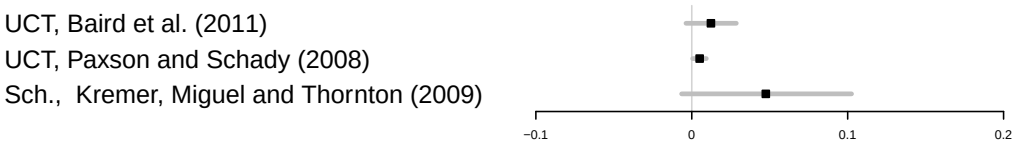


Figure 4: Impact of program on test score per dollar.

2.4 Impact Evaluation

2.4.1 Questions That Need Impact Evaluation

Questions:

- **What is the impact of the program?**
- **Which elements matter the most?**

  Example.

  Effect of Cash Transfer on Education (Akresh et al. , 2013).

  - Randomized experiment in rural Burkina Faso that estimate **the impact of alternative cash transfer delivery mechanisms on education.**
  - **Treatments:** Conditional and unconditional cash transfer to parents.
    → Condition to get the cash – school attendance.

Table 7: The Effect of Conditional and Unconditional Cash Transfer

	Dependent Variable	
	French Reading Test Score	School Attendance
Conditional Cash Transfer	0.196**	0.134***
	(0.90)	(0.049)
Unconditional Cash Transfer	0.003	0.067
	(0.084)	(0.043)
Observations	7.733	7.818

- **Is the approach scalable?**

  Approach 1:

  - Iodine capsules during pregnancy
  - Average cost per dose 0.51-0.56$
    → 7.5% of increase in the total educational attainment (Field et al., 2009).

  Approach 2:

  - Rural electrification
  - Cost of electrification ≈ 4.50$ per month
    → Associated completed schooling increase 20-40% (Khandker et al., 2009).

- **Which type of program to implement?**

  E.g. high or low scholarships to increase attendance.

- **Shall we address only one problem?**

  E.g. Cash transfer and access to school.

- **Do results from one context translate to another?**

  E.g. Is the effect of conditional cash transfer on test score in Bangladesh relevant for Germany?

- **What is the process behind?**

  E.g. Cash transfer → attendance → test score.

2.4.2 Priority of the Impact Evaluation Questions

1. **How influential is the program?**

- **Popularity of the program.**

  How popular is the program? Is it commonly used?

- **Expandability of the program.**

  Is it likely that the program will be expanded?

- **Costs of the program.**

- **New knowledge generation.**

  E.g. Number of studies show that bed nets reduce Malaria rate. Does it translate to the test scores?

- **Theory Driven**

  E.g. Does paying higher wages to teachers increase students test score? Theory: Fairness concerns.

Example: Enhancing the Efficacy of Teacher Incentives Through Loss Aversion (Fryer, Levitt, List and Sadoff, 2012).

- **Problem:** Incentivize teachers to increase student achievement, though financial incentives are ineffective.
- **Idea:** Exploit loss aversion (Kahneman and Tversky, 1979)
- **Main treatments.**
  - "Gain" treatment - teachers receive the monetary reward that depends on performance of their students at the end of the year.
  - "Loss" treatment – teachers get 4000$ but they must return the difference between 4.000$ and their final reward if their students perform below average at the end of the year.

How influential is this program?

Results

Table 8: The Effect Of Treatment on Test Scores

	Dependent Variable	
	Thinklink Math Scores	ISAT/ITBS math Score
Loss	6.866**	6.867**
	(2.677)	(3.269)
Gain	1.263	0.228
	(2.888)	(3.402)
Observations	2311	21444

2. **Can the question be answered?**

   E.g. Can we test the outcomes of fixed vs. floating exchange rate? Can we test lobbying outcomes? Can we gender-based violence?

3. **Do we have sufficient sample size?**
4. **Is the context representative?**

   Typically, we want that the program results will translate into another context. Though, sometimes we can provide a proof-of-concept evaluation.
5. **Is the program at right maturity to evaluate?**

   Avoid the programs that is completely new or will be substituted by the new one.
6. **Do we have the right field partner?**

2.5 Summary

To provide proper randomized evaluation we have to answer on the next set of questions:

1. **Strategic Questions.**
2. **Descriptive Questions.**
3. **Process Questions.**
4. **Impact Questions.**

2.6 Exercises

1. **Assignment:** Think of the field experiment **YOU** could realistically do?
2. **Questions for the field experiments:** What are the 4 sets of questions that we need to ask before running randomized control trial?
3. **Need Assesment:** How to asses the needs? How would you do this for your experiment?
4. **Process Evaluation I:** Which 3 methodological steps you have to have in mind when provide process evaluation? How would you do it for your experiment?
5. **Process Evaluation II:** How to asses a cost effectiveness of the program? How would you do it for your experiment?
6. **Process Evaluation III:** Suppose you want to asses if one can decrease teachers absenteeism by giving the bonus to the teachers that did not miss class. How would you provide process evaluation?
7. **Impact Evaluation I:** Suppose you want investigate how to increase test score in schools. What methods/program can you use?

   - Which method/program is more scalable?
   - Which type of program will you implement?
   - Shall we adress multiple problems?
   - Where else will your result apply?
   - What is underlying process?
8. **Impact Evaluation II:** Priority. How would you prioratize among the programs?
9. **Impact Evaluation III:** Priority. Suppose you can investigate if bonus contracts or delivering teachers by bus affect test score. Which program will you choose? Why?

10. **STAR Project: Class Size and Test Score.**

    - Install and call the package `AER` and use the data `STAR`. What is the data set about?
    - Make a variable `Score` as a sum of math and and reading score for each grade e.g. `$score1<-STAR$read1+STAR$math1`
    - What is the relation between class size in the first grade (`star1`) and test score (`Score`)? Make a boxplot for each grade using function `plot`.
    - Estimate the relation between class size and test score. Interpret the results.
    - Include in the regressions for the first grade students gender (`+gender`). Interpret the results. Is it different from previous results?
    - Include in the regressions students teacher's career ladder level in the first grade (`+ladder1`). Is it different from previous results?
    - Plot relation between (1) gender and class size in the first grade;(2) teachers career ladder level and class size in the first grade.

11. **Presentation:** Dhaliwal, I. et al., 2012. *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries : A General Framework with Applications for Education.* , p.69.

3 Randomizing

3.1 Opportunities to Randomize

3.1.1 What can be randomized?

1. **Randomize across people.**

    - Access
    - Encouragement
      →

2. **Randomize over time.**

    - Access

Time	Group A	Group B	Group C
Year 1			
Year 2			
Year 3			

    - Encouragement? → Perhaps, too complicated.

Time	Group A	Group B	Group C
Year 1			
Year 2			
Year 3			

3.1.2 When is it possible to randomize?

1. New...
    - Program Design
    - Programs
    - Services
    - People
    - Location

Example of New Program.

"...the solution to poverty is to abolish it directly by a now widely discussed measure: the guaranteed income."

— Martin Luther King Jr.

Goal: Guarantee minimum income without administrative costs.

Critics: Reduce incentives to work.

An Experimental Study of the Negative Income Tax. (First study: Ross, 1970)

In experiments that began in 1968 in U.S.A (NJ) households are randomly assigned to treatments with different guaranteed level of income and level of negative income tax.

Results (Munnel, 1986): (1) Reduced work effort, especially among women (though, generally, not as dramatic as expected); (2) Increased rate of breakup.

2. Subscription
    - Oversubscription
    - Undersubscription

Example of Undersubscription.

The Role of Information and Social Interactions in Retirement Plan Decisions... (Duflo and Suez, 2012)

Treatment: Provide 20$ for attending information fair about retirements plans for randomly selected group of employees in randomly selected departments.

Table 9: The effect of Sending an Invitation Letter

	Dependent Variable		
	Fair att.	TDA	TDA
Letter	0.138***	-0.0446	
	(.019)	(.0402)	
Letter in Department	0.90***		0.0568**
	(.022)		(.0257)
Observations	6144	3726	5587

3. Timing
    - Rotation
    - Admission in Phases

Time	Group A	Group B	Group C
Year 1			
Year 2			
Year 3			

Time	Group A	Group B	Group C
Year 1			
Year 2			
Year 3			

4. Admission Cutt-offs

Assignment

- ■ 5% of applicants with the best score to the **100%** credit.
- × 5% of applicants with worst score are **excluded.**
  Rest of the applicants are **randomly assigned to one of the 3 tracks**:
  - △ 70% credit.
  - ▫ 30% credit.
  - ◇ Nothing

Table 10: When is it possible to randomize?

Opportunity	Description
New Program Design	We know the problem but no agreement about the solution
New Programs	When a program is new and being pilot-tested.
New Services	When an existing program offers a new service.
New People	When a program is being expanded to a new group of people
New Locations	When a program is being expanded to new areas.
Oversubscription	When program can not serve all interested people.
Undersubscription	When not everyone who is eligible for the program takes it up.
Rotation	When the program's outcomes are to be shared by rotation.
Admission in phases	When the admission to the program is going to be in phases.
Admission cutoffs	When the program has a merit cutoff.

3.2 Level of Randomization

A. Individual-level of Randomization

1. Select Study Sites

2. Select Eligible people.

3. Randomize access.

4. Survey people in the both treatments. ☑

B. Group-level of Randomization

1. Randomly Select Study Sites

2. Apply Eligible Criteria e.g. cities with population less than 1,000,000.

3. Among Eligible Randomly Assign districts to treatment

4. Survey Random sample of people in both groups. ☑

How to choose unit of randomization?

- **Unit of Measurement**

  Unit of randomization should be equal or higher than unit of measurement. E.g. Useless to randomize the training on workers level when we want to see the effect of training on firms profit
- **Spillovers**
- **Attrition**
- **Compliance**
- **Statistical Power**
- **Feasibility**

Spillovers:

- **Physical.** Migration increase pollution.
- **Behavioral.** Training for workers → peers imitate.
- **Informational.** People may get to know from colleagues that it worth going for the fair.

- **Market-wide.** Older workers lose their jobs because firms receive fingernail incentives for hiring young people.

! This can be a problem since randomized evaluation requires that the outcome of one person is independent from the group where s(he) is located.

How to deal with spillovers?

→ Choose the level of randomization to limit spillovers. For instance, randomize at level of cities to avoid market-wide spillovers.

! Choose the level of randomization not too high as well not too low.

What matters for this choice?

- Untreated individuals in treated group.
- Untreated units near treated units.
- Use two level randomization if you want measure spillovers.

Heckman, 1991.

"As-1: There is no effect of Randomization on participation decision

AS-2: If there is effect of participation decisions, either (a) the effect of treatment is the same for all participant or (b) if agents differ in their response to treatments, their idiosyncratic responses to treatment do not influence their participation decisions."

See for further discussion Heckman and Smith, 1995, Deaton, 2010.

Attrition. Data is missing from some of the people in the sample. E.g. People refuse to answer on they survey in the control group. **How to deal with it?**

- Higher level of randomization
- Incentive surveys.

Compliance. People drop out from the program.

How to deal with it?

- By program staff. Make sure that program staff believe in the running program.
- By Participants.

Low rate of attrition and compliance should be taken very seriously since it introduce selections bias and invalidate the whole experiment.

Statistical Power. All the things equal larger the number of units for randomization higher is statistical power.

Feasibility.

- **Ethics:** Is randomization ethical?
- **Politics:** Is it permitted? Is it fine for community?
- **Logistics:** Can we carry out the program?
- **Costs:** Do we have the money? Is it the best way to use the money?

3.3 Which aspect to randomize?

1. **Simple treatment lottery.** When to use?
   - Program limited in scale or piloted.
   - Oversubscription
   - We want to measure the effect in the long-run.

2. **Treatment lottery around a cutoff.** When to use?
   - When there is admission cutoff
   - *!* If this design can help to answer relevant policy question.

3. **Phase-in design.** When to use?
   - Everybody must get access to the program
   - Anticipation of the program do not change the behavior of people
   - Interest in average effect of the program over years.

4. **Rotation.** When to use?
   - Resources are limited but expect to increase.
   - Outcome of interest is during the program.
   - No interest in a long run effect.
   - When we want to measure seasonal effects.

5. **Encouragement.** When to use?
   - Open access to the program but under subscribed
   - Open access but application takes time and effort

- When we can assume that encouragement will not affect the outcomes of interest

3.4 Mechanics of Simple Randomization

The ingredients of random assignment:

1. A list of eligible units
2. Number of randomization cells
3. Allocation fractions
4. A randomization device
5. Initial data on randomization units.

1. **A list of eligible units:** How to get the list of eligible units?
   - Local governments
   - School registers
   - Resource appraisals
   - Census with basic need assessment
   - Revealed need

   ...

   To make the sample representative use random sampling.

2. **Number of randomization cells.**
   Typically two: treatment and control(comparison) group. Depends on research question.

3. **Allocation fractions.**
   Equal fractions (50%, 50%) typically maximize statistical power.
4. **A randomization device:**
   - Mechanical devices e.g. coins, cards, dice.
   - Published random tables e.g. RAND table, www.random.org
   - Computerized number generation e.g. =rand() in excel, uniform() in STATA, rnorm() or sample() in R.

Steps in Randomization

1. Order the list of eligible units randomly
2. Allocate the units into different groups
3. Randomly choose which group will receive which treatment.

! After randomization check for balance on observable.

3.5 Stratification and Re-Randomization

Suppose we do not achieve balance. Why is it a problem?

Example: Birth Control Pills

- Suppose we test birth control pills.
- By chance in treatment group only women and in control only men
- Our evaluation shows no effect of birth controls or even slightly positive effect on birth rate.

→ We conclude that Birth control pills do not work or even can increase birth rate.

What to do?

3.5.1 Stratification.

Steps to perform stratification:

1. Divide the pool of eligible units into sub-lists based on chosen characteristics
2. Do simple random assignment for each sub-list
3. Randomly pick which cell is treatment and which is comparison.

Drawback: We shall use only binary variables or transform continuous variables into discrete.

Alternatives: Max-min t statistic re-randomization, Propensity score matching.

3.5.2 Max-min t-statistic Randomization.

Steps to perform max-min t-statistic re-randomization:

1. Do simple randomization multiple times e.g 1000 times
2. Calculate t-statistic on the variables that you want to balance
3. Find maximum t-statistic in each of randomization
4. Find randomization where maximum t-statistic is minimal
5. Choose this randomization

3.5.3 Pairwise Matching.

Steps to Perform Pairwise matching:

1. Make pairs based propensity score matching using the variables that you want to balance on.
2. Randomly pick on subject which will go to treatment and which will go to comparison in each pair.

3.6 Simulation

Let's simulate simple data generating process:

```
n <- 100  #Sample size!
sd <- 10
mean <- 0
id <- c(1:n)  #Subjects ID

# Suppose we have 1 dummy, 4 continuous variables, and some noise e
set.seed(12567239)  # Set seed for Repoducibility!
d1 <- round(runif(n, 0, 1))
x1 <- rnorm(n, mean, sd)
x2 <- rnorm(n, mean, sd)
x3 <- rnorm(n, mean, sd)
x4 <- rnorm(n, mean, sd)
e <- rnorm(n, mean, sd)

# Here is our Data Generating Process:
y <- 10 + 2 * d1 + 2 * x1 + 2 * x2 + 3 * x3 + 4 * x4 + e
df <- as.data.frame(cbind(id, y, d1, x1, x2, x3, x4))

head(df)

##   id          y d1         x1          x2          x3         x4
## 1  1  -20.62275  1  -9.421806 -14.9117072   0.6502851   1.341417
## 2  2 -146.59094  0 -19.330254 -21.2415739 -10.5720369 -10.564112
## 3  3  -87.55020  0  -7.094161  -1.8207241  -0.2229564 -18.692582
## 4  4   55.21210  0   2.828126  -5.4205977  -0.7703064  13.498108
## 5  5  -48.96167  1  -9.543756   0.7112626  -6.8553521  -7.377356
## 6  6  -35.94191  0 -22.657923  -4.5767202 -19.0765148  21.319104
```

3.6.1 Simple Randomization.

Let's make simple randomization:

```
set.seed(12567239)  # Set seed for Repoducibility!
df$T <- sample(c(rep(1, n/2), rep(0, n/2)), n, replace = FALSE)
str(df$T)

##  num [1:100] 0 1 1 1 0 1 0 1 1 0 ...

table(df$T)
```

```
##
##  0  1
## 50 50
```

Now, let's make experimental intervention based on simple randomization and estimate the treatment effect.

We set the effect size to 12.

```
eff <- 12
df$yT <- df$y + ifelse(df$T == 1, eff, 0)
lm1 <- lm(yT ~ T, data = df)
summary(lm1)

##
## Call:
## lm(formula = yT ~ T, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.828  -34.936   -2.424   40.152  113.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.814      7.759   2.425   0.0171 *
## T             -2.577     10.972  -0.235   0.8148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.86 on 98 degrees of freedom
## Multiple R-squared:  0.0005627,Adjusted R-squared:  -0.009636
## F-statistic: 0.05518 on 1 and 98 DF,  p-value: 0.8148
```

What happens if we set the effect to zero?

```
eff <- 0
df$yT <- df$y + ifelse(df$T == 1, eff, 0)
lm1 <- lm(yT ~ T, data = df)
summary(lm1)

##
## Call:
## lm(formula = yT ~ T, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.828  -34.936   -2.424   40.152  113.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.814      7.759   2.425   0.0171 *
## T            -14.577     10.972  -1.329   0.1871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 54.86 on 98 degrees of freedom
## Multiple R-squared:  0.01769,Adjusted R-squared:  0.007669
## F-statistic: 1.765 on 1 and 98 DF,  p-value: 0.1871
```

3.6.2 Max-min t-statistic Randomization

Let's make randomization based on max-min t-statistic balacning on **baseline outcome variable** and other correled with outcome variables:

```
TT <- function(d) {
    l <- NULL
    d$T <- NULL
    n <- length(d$y)
    d$T <- sample(c(rep(1, n/2), rep(0, n/2)), n, replace = FALSE)
    yp <- summary(lm(d$y ~ d$T))$coefficients[2, 3]
    d1p <- summary(lm(d$d1 ~ d$T))$coefficients[2, 3]
    x1p <- summary(lm(d$x1 ~ d$T))$coefficients[2, 3]
    x2p <- summary(lm(d$x2 ~ d$T))$coefficients[2, 3]
    l <- append(d$T, c(yp, d1p, x1p, x2p))
    l
}

dd <- as.data.frame(replicate(10000, TT(df)))
colMax <- function(X) apply(X, 2, max)
dd <- rbind(dd, colMax(abs(dd[c(length(df$y) + 1:4), ])))
df$TT <- dd[c(1:length(df$y)), which.min(dd[length(dd$y) + 5, ])]
```

```
eff <- 12
df$yTT <- df$y + ifelse(df$TT == 1, eff, 0)
lmTT <- lm(y ~ TT, data = df)
summary(lmTT)

##
## Call:
## lm(formula = y ~ TT, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -153.433  -36.151   -2.797   38.680  122.249
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.842      7.799   0.877    0.383
## TT             9.367     11.030   0.849    0.398
##
## Residual standard error: 55.15 on 98 degrees of freedom
## Multiple R-squared:  0.007306,Adjusted R-squared:  -0.002824
## F-statistic: 0.7212 on 1 and 98 DF,  p-value: 0.3978
```

3.6.3 Pairwise Matching Randomization.

Now, let's make randomization based on pairwise matching on **baseline outcome variable** and other correled with outcome variables:

```
require(nbpMatching)
df.dist <- gendistance(df[, c("id", "y", "d1", "x1", "x2")], idcol = 1)
df.mdm <- distancematrix(df.dist)^0.1
df.match <- nonbimatch(df.mdm)


head(df.match$matches)

##   Group1.ID Group1.Row Group2.ID Group2.Row Distance
## 1         1          1        20         20 4.977600
## 2         2          2        18         18 5.897635
## 3         3          3        21         21 4.905253
## 4         4          4        53         53 4.866457
## 5         5          5        98         98 4.882453
## 6         6          6        39         39 4.992965

df.assign <- assign.grp(df.match$matches)
df$TM <- as.factor(df.assign$treatment.grp)
df$pair <- as.factor(df.assign$Distance)
```

Now, let's make experimental innervation based on pairwise matched randomization and estimate an effect. Again, we set the effect to 12.

```
eff <- 12
df$yTM <- df$y + ifelse(df$TM == "B", eff, 0)
lmM <- lm(yTM ~ TM + pair, data = df)
summary(lmM)

##
## Call:
## lm(formula = yTM ~ TM + pair, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -131.57  -12.71    0.00   12.71  131.57
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             16.131     27.911   0.578   0.5659
## TMB                     16.949      7.817   2.168   0.0350 *
## pair4.35247598854743    47.577     39.083   1.217   0.2293
## pair4.48225011086149    99.885     39.083   2.556   0.0138 *
## pair4.52786145740364   -36.733     39.083  -0.940   0.3519
...
```

What happens if we set effect to zero?

```
eff <- 0
df$yTM <- df$y + ifelse(df$TM == "B", 0, 0)
lmM <- lm(yTM ~ TM + pair, data = df)
summary(lmM)

##
## Call:
## lm(formula = yTM ~ TM + pair, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -131.57  -12.71    0.00   12.71  131.57
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             16.131     27.911   0.578   0.5659
## TMB                      4.949      7.817   0.633   0.5296
## pair4.35247598854743    47.577     39.083   1.217   0.2293
## pair4.48225011086149    99.885     39.083   2.556   0.0138 *
## pair4.52786145740364   -36.733     39.083  -0.940   0.3519
...
```

3.6.4 Power of Pairwise Matching I

Now, let's see how often we reject the null-hypothesis at 5% level of significance using different methods of randomization.

We make the function that:

1. Simulate simple data generating process.
2. Make experimental intervention based on different methods of randomization
3. Estimate the effect
4. Return p-values for each intervention

```
require(nbpMatching)

simR <- function(n, sd, mean, eff) {
    d1 <- NULL
    x1 <- NULL
    x2 <- NULL
    x3 <- NULL
    x4 <- NULL
    e <- NULL
    y <- NULL
    id <- c(1:n)
    d1 <- round(runif(n, 0, 1))
    x1 <- rnorm(n, mean, sd)
    x2 <- rnorm(n, mean, sd)
    x3 <- rnorm(n, mean, sd)
    x4 <- rnorm(n, mean, sd)
    e <- rnorm(n, mean, sd)
    noise <- rnorm(n, mean, sd)
```

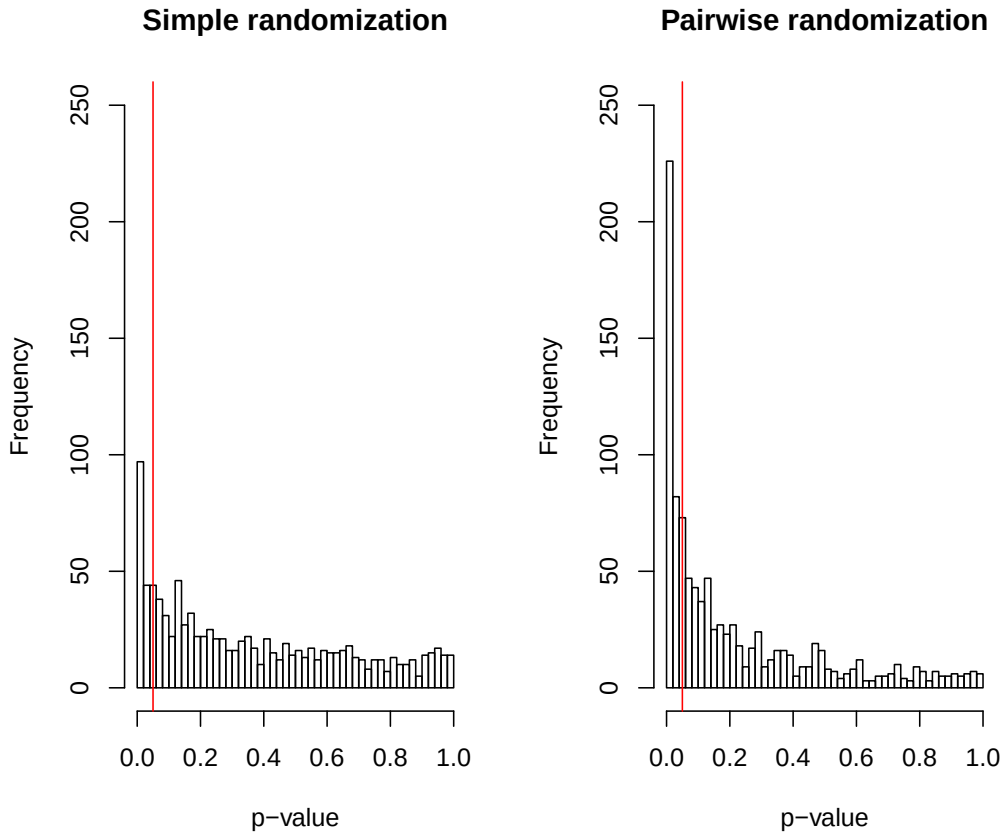
```
    y <- 10 + 2 * d1 + 2 * x1 + 2 * x2 + 3 * x3 + 4 * x4 + e
    df <- as.data.frame(cbind(id, y, d1, x1, x2, x3, x4))

    # Simple Randomization
    df$T <- sample(c(rep(1, n/2), rep(0, n/2)), n, replace = FALSE)
    df$yT <- df$y + ifelse(df$T == 1, eff, 0)



    # Pairwise matched randomization
    df.dist <- gendistance(df[, c("id", "y", "d1", "x1", "x2")], idcol = 1)
    df.mdm <- distancematrix(df.dist)^0.1
    df.match <- nonbimatch(df.mdm)
    df.assign <- assign.grp(df.match$matches)
    df$TM <- as.factor(df.assign$treatment.grp)
    df$pair <- as.factor(df.assign$Distance)
    df$yM <- df$y + ifelse(df$TM == "B", eff, 0)



    lmT <- lm(yT ~ T, data = df)
    pT <- summary(lmT)$coefficients[2, 4]
    lmTM <- lm(yM ~ TM + pair, data = df)
    pTM <- summary(lmTM)$coefficients[2, 4]

    as.data.frame(cbind(pT, pTM))
}


simR(n, sd, mean, eff = 12)

##          pT       pTM
## 1 0.3422138 0.4233231
```

Simulate the analysis 1000 times:

```
eff <- 12
pR <- replicate(1000, unlist(simR(n, sd, mean, eff)))
pR[c(1, 2), c(1:5)]

##          [,1]      [,2]      [,3]        [,4]       [,5]
## pT  0.2555272 0.6300877 0.1849095 0.634435607 0.17613324
## pTM 0.1573919 0.8938946 0.3873995 0.008412453 0.05425026

mean(pR[1, ] < 0.05)

## [1] 0.18

mean(pR[2, ] < 0.05)

## [1] 0.385
```

Let's plot the distribution of p-values

```
par(mfrow = c(1, 2))
hist(pR[1, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 250), main = "Simple randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
hist(pR[2, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 250), main = "Pairwise randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
```

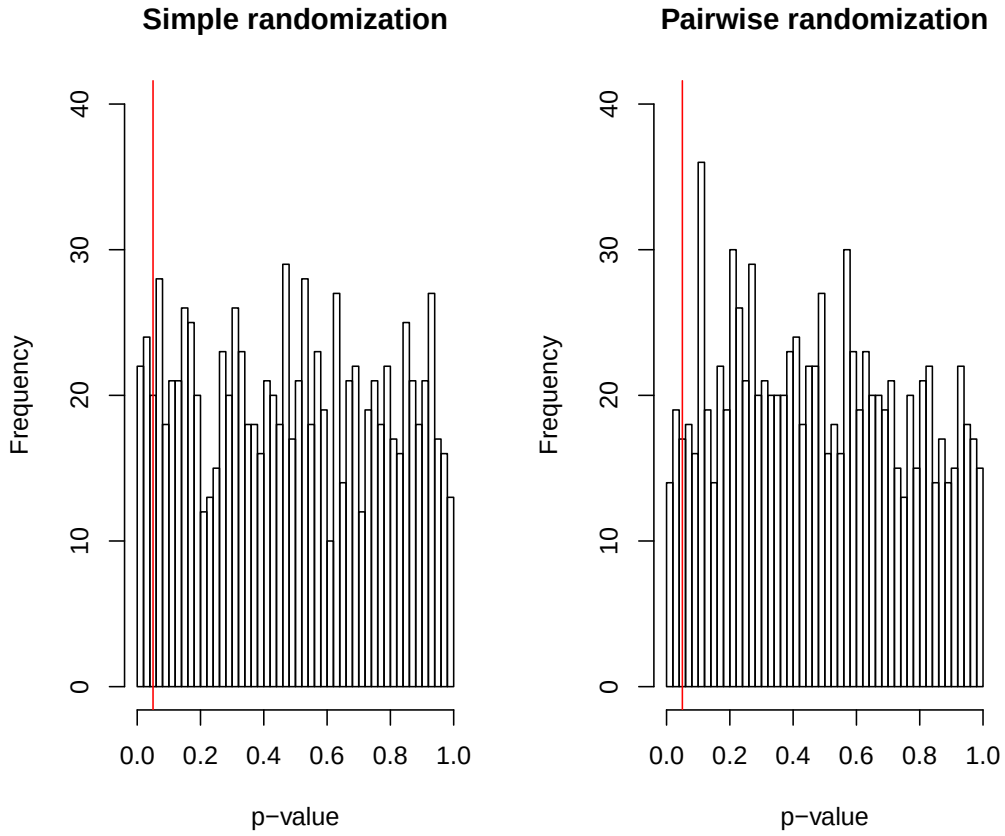


Can it be that we just increase the chance of FALSE positive?

Let's set effect size to 0 and see:

```
eff <- 0
pR <- replicate(1000, unlist(simR(n, sd, mean, eff)))
mean(pR[1, ] < 0.05)

## [1] 0.053

mean(pR[2, ] < 0.05)

## [1] 0.042
```

```
par(mfrow = c(1, 2))
hist(pR[1, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 40), main = "Simple randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
hist(pR[2, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 40), main = "Pairwise randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
```

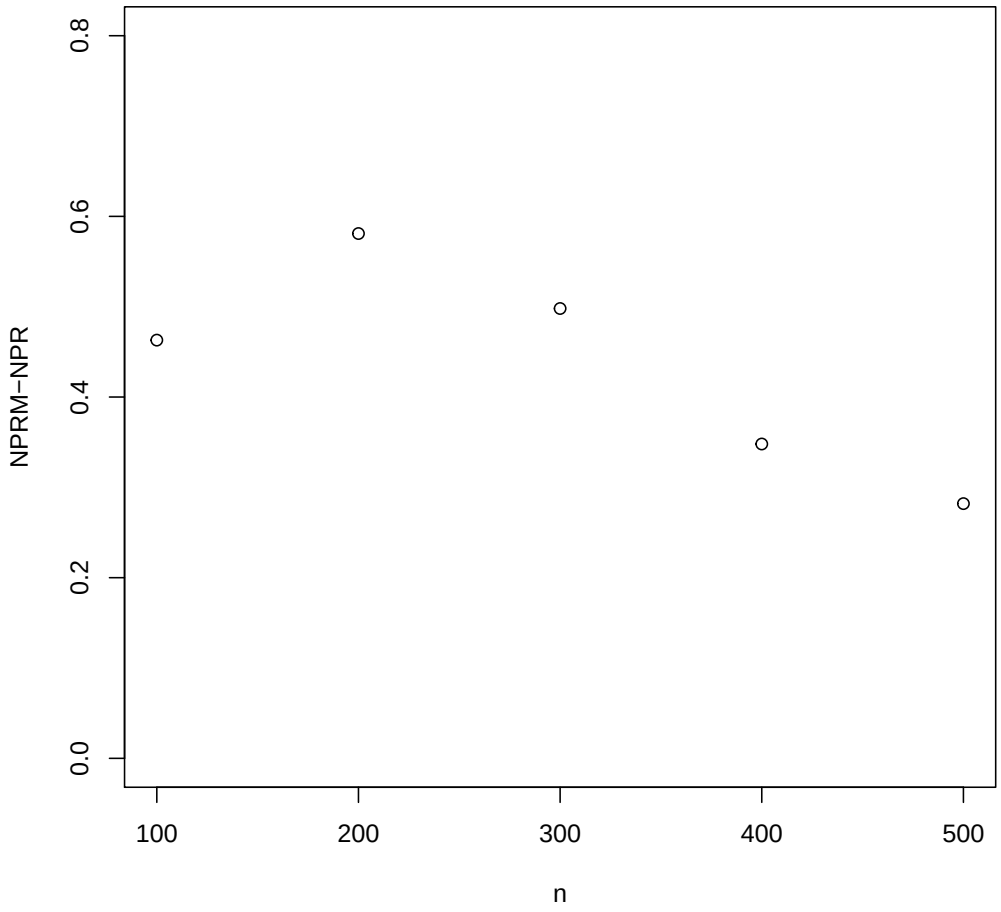


We would like to know how the effect of matching depends on sample size. We make a function that calculate the difference between simple randomization and pairwise randomization in number of cases when p-value is lower than 0.05.

```
Diff <- function(n, sd, mean, eff, nsim) {
    pR <- replicate(nsim, unlist(simR(n, sd, mean, eff)))
    mean(pR[2, ] < 0.05) - mean(pR[1, ] < 0.05)
}
```

```
d1 <- Diff(100, sd, mean, eff = 12, 1000)
d2 <- Diff(200, sd, mean, eff = 12, 1000)
d3 <- Diff(300, sd, mean, eff = 12, 1000)
d4 <- Diff(400, sd, mean, eff = 12, 1000)
d5 <- Diff(500, sd, mean, eff = 12, 1000)
```

```
plot(c((1:5) * 100), c(d1, d2, d3, d4, d5), ylim = c(0, 0.8), ylab = "NPRM-NPR",
    xlab = "n")
```



3.6.5 Power of Pairwise Matching II

Let's return to the birth control example. In case of birth control pills we know that the pills effectively prevent pregnancy only among women. Thus, the effect of treatment conditioned on gender and our experimental intervention will take the following form:

```
df$yT <- df$y + df$d1 * ifelse(df$T == 1, eff, 0)
```

We make a new function to see how often we can reject the null-hypothesis at 5% level of significance.

```
require(nbpMatching)
simRD <- function(n, sd, mean, eff) {
    d1 <- NULL
    x1 <- NULL
```

```
    x2 <- NULL
    x3 <- NULL
    x4 <- NULL
    e <- NULL
    y <- NULL
    id <- c(1:n)
    d1 <- round(runif(n, 0, 1))
    x1 <- rnorm(n, mean, sd)
    x2 <- rnorm(n, mean, sd)
    x3 <- rnorm(n, mean, sd)
    x4 <- rnorm(n, mean, sd)
    e <- rnorm(n, mean, sd)
    noise <- rnorm(n, mean, sd)
    y <- 10 + 2 * d1 + 2 * x1 + 2 * x2 + 3 * x3 + 4 * x4 + e
    df <- as.data.frame(cbind(id, y, d1, x1, x2, x3, x4))

    # Simple Randomization
    df$T <- sample(c(rep(1, n/2), rep(0, n/2)), n, replace = FALSE)
    df$yT <- df$y + df$d1 * ifelse(df$T == 1, eff, 0)



    # Pairwise matched randomization
    df.dist <- gendistance(df[, c("id", "d1", "x1", "x2")], idcol = 1)
    df.mdm <- distancematrix(df.dist)^0.1
    df.match <- nonbimatch(df.mdm)
    df.assign <- assign.grp(df.match$matches)
    df$TM <- as.factor(df.assign$treatment.grp)
    df$pair <- as.factor(df.assign$Distance)
    df$yM <- df$y + df$d1 * ifelse(df$TM == "B", eff, 0)



    lmT <- lm(yT ~ T, data = df)
    pT <- summary(lmT)$coefficients[2, 4]
    lmTM <- lm(yM ~ TM + pair, data = df)
    pTM <- summary(lmTM)$coefficients[2, 4]

    as.data.frame(cbind(pT, pTM))
}


simR(n, sd, mean, eff = 12)


##          pT        pTM
## 1 0.4393146 0.05091229
```

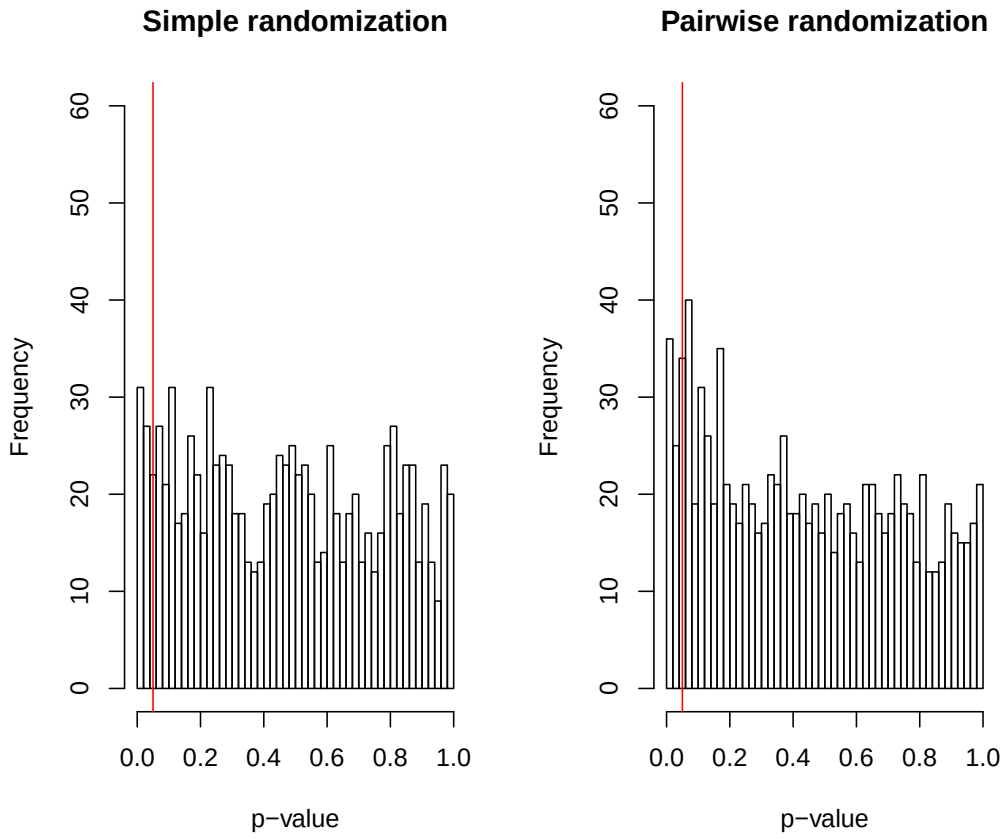
Simulate the analysis 1000 times:

```
eff <- 12
pRD <- replicate(1000, unlist(simRD(n, sd, mean, eff)))
mean(pRD[1, ] < 0.05)

## [1] 0.07
```

```r
mean(pRD[2, ] < 0.05)

## [1] 0.076
```

Plot the distribution of p-values.

```r
par(mfrow = c(1, 2))
hist(pRD[1, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 60), main = "Simple randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
hist(pRD[2, ], breaks = 50, xlim = c(0, 1), ylim = c(0, 60), main = "Pairwise randomization",
    xlab = "p-value")
abline(v = 0.05, col = "red")
```

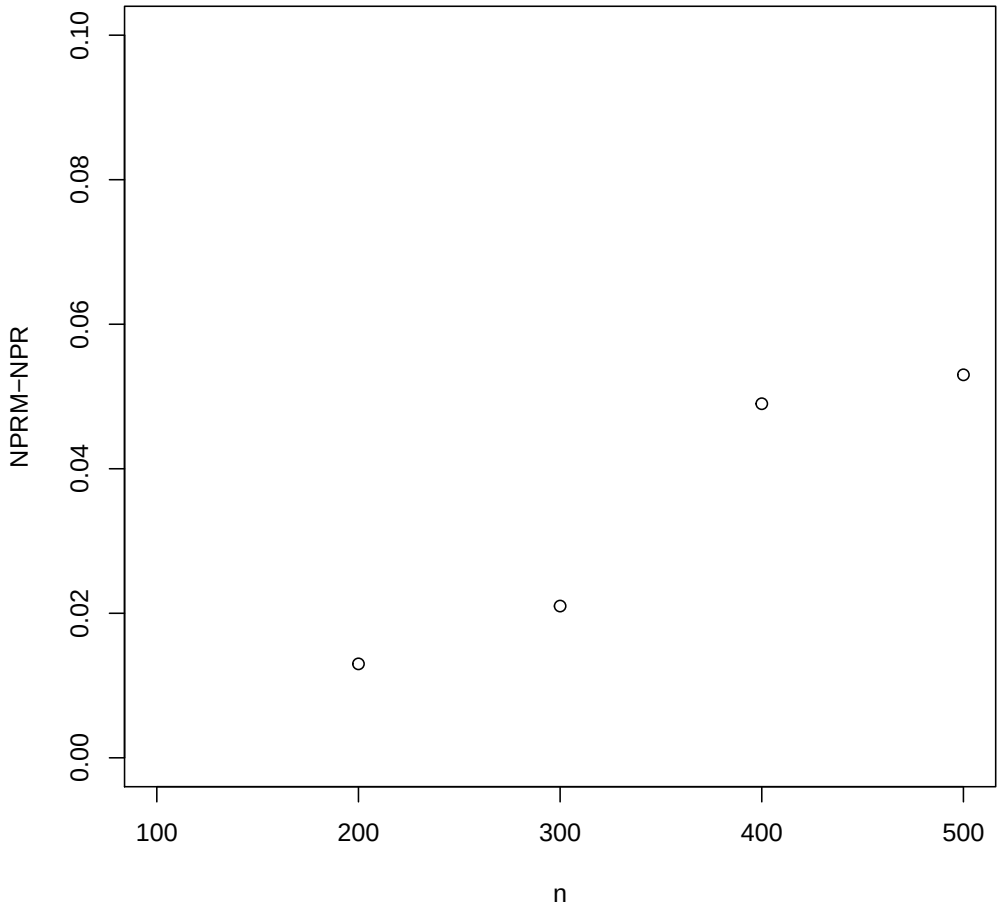


We would like to know how the effect of matching depends on sample size. We make a function that calculate the difference between simple randomization and pairwise randomization in number of cases when p-value is lower than 0.05.

```r
DiffD <- function(n, sd, mean, eff, nsim) {
    pR <- replicate(nsim, unlist(simRD(n, sd, mean, eff)))
    mean(pR[2, ] < 0.05) - mean(pR[1, ] < 0.05)
}
```

```
d1D <- DiffD(100, sd, mean, eff = 12, 1000)
d2D <- DiffD(200, sd, mean, eff = 12, 1000)
d3D <- DiffD(300, sd, mean, eff = 12, 1000)
d4D <- DiffD(400, sd, mean, eff = 12, 1000)
d5D <- DiffD(500, sd, mean, eff = 12, 1000)
```

```
plot(c((1:5) * 100), c(d1D, d2D, d3D, d4D, d5D), ylim = c(0, 0.1), ylab = "NPRM-NPR",
    xlab = "n")
```



Example: Household Bargaining and Excess Fertility: An Experimental Study in Zambia.(Ashraf et al., 2014)

Problem: "Unwanted" birth

→ decrease in female schooling and labor force participation.

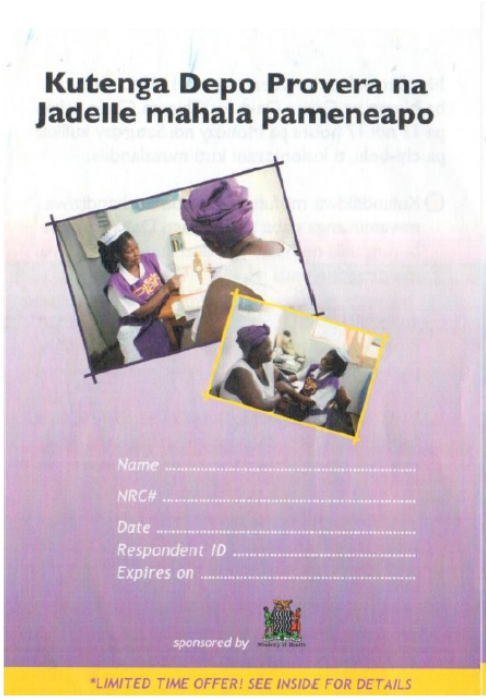
Remedy: Increase birth control by women?

Treatments: Provide a voucher to get free access to contraceptives to women or to couple: To get access women either have to come for consultation about family planning alone or with a husband.

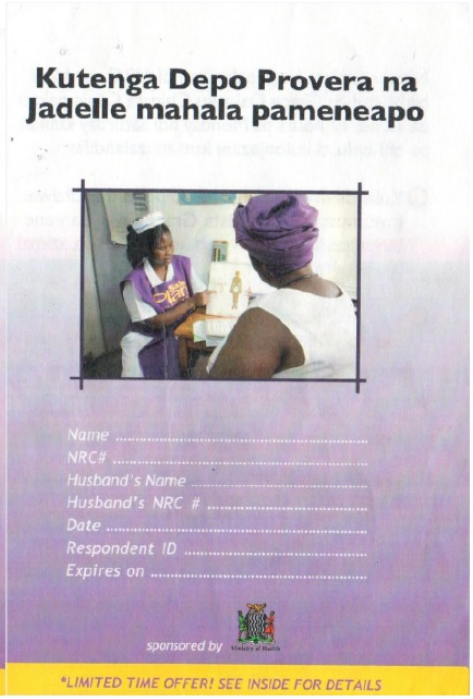
Table 11: Descriptive Statistic on some variables that were balanced in recruited sample

	Indiviudal			Couple			
Variable	Mean	SD	N	Mean	SD	N	P-value
Using any method at baseline	0.844	0.0223	527	0.855	0.0160	498	0.622
Using injectable at baseline	0.194	0.0253	527	0.219	0.0181	498	0.317
Using pill at baseline	0.292	0.0283	527	0.279	0.0203	498	0.643
Husband's ideal number of children	4.204	0.148	378	4.433	0.105	372	0.122

Individuals



Couples



Source: Ashraf et al., 2014

Randomization: Individual level, minmax t statistic method for randomization balancing on:

- Using injectables
- Using pills
- Desire to have kids
- Number of kids
- Wife's eduction
- Wife's age

Findings:

- "... Women given access with their husbands were 19% less likely to seek family planning services, 25% less likely to use concealable contraception, and **27% percent more likely to give birth.**"
- "... women given access to contraception alone report a lower subjective well-being..."

Table 12: Descriptive Statistic on some variables in final sample

	Indiviudal			Couple			
Variable	Mean	SD	N	Mean	SD	N	P-value
Using any method at baseline	0.841	0.0259	377	0.869	0.0184	366	0.280
Using injectable at baseline	0.202	0.0300	377	0.221	0.0214	366	0.511
Using pill at baseline	0.297	0.0337	377	0.306	0.0240	366	0.791
Husbands ideal number of children	**4.168**	0.148	374	**4.435**	0.105	368	**0.0721**

3.7 Best Practice in Randomization

1. **Report the random assignment details:**
   - Method of randomization
   - Variables used for balancing
   - Balancing Criteria
2. **Report Practical Details:**
   - Who did randomization?
   - Randomization device
   - Public or Private?
3. **Avoid re-randomization to achieve balance**
4. **Carefully choose variables for balance:**
   (a) Baseline of outcomes variable
   (b) Variables that shall affect the outcome
   (c) Variables important for subgroup analysis.
5. **Consider statistical power**
6. **Make randomization reproducible** (set seed when randomizing).
7. **Consider attrition problem**

3.8 Summary

- Randomize across (1) subjects or over (2) time: Access or encouragement.
- We can randomize when
  1. Something is new
  2. Subscription issue
  3. Issue with timing
  4. Admission cut-offs
- Choose (individual or group) level of randomization based:
  1. Unit of Measurement
  2. Spillovers
  3. Attrition
  4. Compliance
  5. Statistical Power
  6. Feasibility
- Carefully choose aspect of program to randomize.
- Perform randomization in clear and transparent way.
- Consider balancing if sample size is small.

3.9 Exercises

1. **Randomization:** What can we randomize?
2. **Opportunity for Randomization I:** When can we randomize?
3. **Opportunity for Randomization II:** How can we randomize if there is undersubscription?
4. **Opporuntiy for Randomization III:** Suppose we randomize around cut-off, what are the disadventages?
5. **Assignement:** Which method of randomization do you want to use for your project?
6. **Level of Randomization I:** At which level can we randomize?
7. **Level of Randomization II:** What shall we consider when we choose level of randomization?
8. **Level of Randomization III:** Suppose we evaluate using field experiment migration policy that helps to hire foreigners in national basketball team. We find no difference in the performance (neither measured in personal scores, passes, nor fauls) of migrants as opposed to citizens. Shall we reccomend to stop using this policy?
9. **Level of Randomization IV:** Suppose we evaluate the policy that allows organizing trade unions using field experiment. We observe positive effect of this policy on wages: Wages increase in the organizations where policy allows trade union. Can we claim that this policy was welfare enhancing?
10. **Level of Randomization V:** Suppose we want to evaluate migration policy in the countries with ineffective bureaucracy. About what shall we worry in the design of our experiment?
11. **Recap of Statistics:**
    - **Level of significance** – Predefined probability of rejecting the null hypothesis, despite it being true.
    - **P-value** – Probability of drawing a sample that is at least as averse to the null hypothesis as our data given that the null hypothesis is true.

```r
library(Ecdat)
data(Workinghours)
# Make sample of 'hours' of the same size
s1 <- sample(Workinghours$hours, replace = TRUE)
mean(s1)
mean(Workinghours$hours)
# Make 1000 samples of 'hours' of the same size
x <- replicate(1000, mean(sample(Workinghours$hours, replace = TRUE)))

## Check normality
hist(x)
plot(density(x))
qqnorm(x)
shapiro.test(x)

## Estimate probability of hours to be below 1150
sum(x < 1150)/length(x)
mean(x < 1150)
t.test(Workinghours$hours, mu = 1150, alternative = "greater")

# Make 1000 samples of 'hours' of the same size for women with kids and
# without
x1 <- replicate(1000, mean(sample(subset(Workinghours, child5 != 0)$hours, replace = TRUE)))
x2 <- replicate(1000, mean(sample(subset(Workinghours, child5 == 0)$hours, replace = TRUE)))

## Check normality
```

```r
hist(x1 - x2)
plot(density(x1 - x2))
qqnorm(x1 - x2)
shapiro.test(x1 - x2)

## Calculate confidence intervals
quantile(x1 - x2, c(0.025, 0.975))
t.test(hours ~ child5 == 0, data = Workinghours)
```

What if the variable is binary?

```r
library(Ecdat)
data(Workinghours)
attach(Workinghours)
# Check sample distribution
hist(owned)

# Make 1000 samples of 'ow' of the same size
ow <- replicate(1000, mean(sample(owned, replace = TRUE)))

## Check normality
hist(ow)
plot(density(ow))
qqnorm(ow)
shapiro.test(ow)

mean(ow < 0.66)
t.test(owned, mu = 0.66, alternative = "greater")

# Check sample distribution
hist(child5)

# Make 1000 samples of 'ow' of the same size
ch <- replicate(1000, mean(sample(child5, replace = TRUE)))

## Check normality
hist(ch)
plot(density(ch))
qqnorm(ch)
shapiro.test(ch)

quantile(ch, c(0.005, 0.995))
t.test(child5, mu = 1, conf.level = 0.99)
```

12. **Level of Randomization VI:** You design a program where unemployed people have to go every month two the unemployment office to discuss their job application process. What can be a problem? What can you do to reduce it?
13. **Level of Randomization VII:** You make a program where you can randomize on the level of cities or individuals. What would you take into account?
14. **Which aspect to randomize?:** What can be a problem with
    - Treatment lottery around a cutoff?
    - Phase-in design randomization?
    - Randomization using rotation?

- Randomization of encouragement?

15. **Mechanics of randomization:** What are the ingredients of random assignment?

16. **Balancing I:** Why do we want to achieve a balance? What methods can we use?
17. **Balancing II:** How to perform stratification?

18. **Balancing III:** Suppose you study the effect of new unversity program, you plan to stratify your sample on the next three variables: Gender, age, distance to the university. How will you do this? What can be the problem?

19. **Balancing IV:** How to perform min max t-statistic re-randomization? Pairwise randomization?
20. **Balancing V:** You expect low rate of compliance, which method for balancing will you choose: Stratification or pairwise randomization?

21. **Balancing VI:** Suppose you investigate the effect of negative income tax. Which variables for balancing using minmax t-statisitc rerandomization will you choose? Why?
22. **Assignment:** Which variables will you use to balance your sample?
23. **Presentations:**

    Crépon et al., 2012. *Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment.*

    Bruhn, Miriam, and David McKenzie. 2009. *In Pursuit of Balance: Randomization in Practice in Development Field Experiments.* American Economic Journal: Applied Economics, 1(4): 200-232.

    Baird, Sarah, et al. *Designing experiments to measure spillover and threshold effects.* IZA WP6681 (2012).

4 Outcomes

4.1 Outcomes and Indicators

"There is nothing mysterious about questioning. It is no more than obtaining needed information from subjects."

— The CIAs 1983 Human Resource Exploitation Training Manual

- **Outcome** – a change or impact of the program we are evaluating. e.g. equality, democracy
- **Indicator** – an observational signal used to measure outcomes e.g. number of people who get a job by race, gender; votes share
- **Instrument** – the tool we use to measure indicators e.g. call-back for interview, exit-poll
- **Variable** – the numeric values of indicators
- **Respondent** – the person or group of people we interview, test, or observe to measure the indicators.

4.2 Data Sources

1. Administrative data
   - Basic data collected on everyone
   - Random sample of individuals
2. Collecting your own data
   - Survey
   - Nonsurvey

I. Respondents:

1. Who is subject to treatment?
2. Who is representative?
   → Use random sampling!
3. Who knows information we need?
4. Who is unlikely to manipulate information?
5. Who will be most efficient in reporting data?

II. Enumerator:

1. Same people interview treatment and control group.
2. Enumerators must differ from program staff
3. Enumerator characteristics matters e.g. gender, language
4. Plan how to deal with cheating and shirking:
   - Inform about checking of surveys
   - Provide back-check: 10%-15% of surveys; more in the beginning.
   - Provide clear definitions

   \+ Use digital devices to check for suspicious patterns e.g GPS, time of submission.

III. Time

1. Baseline survey if
   - Small sample size
   - Individual-specific outcomes matters e.g cognitive abilities
   - You want to have a balance
   - Plan to provide subgroup analysis or use controls

     ! We must get information on controls before the implementation of program
2. Beginning of survey depends on (a) novelty and (b) lag effects
3. Frequency of surveys.
   → Find a balance between taking fine grained picture of program and costs (both from evaluates and respondents).
4. End of surveys
   → Consider attrition rate.

4.3 Assessing Outcomes Measures

Criteria:

1. Logical Validity
2. Measurable
   (a) Observable
   (b) Feasible
   (c) Detectable
3. Precision
   (a) Exhaustive Indicator
   (b) Exclusive Indicator
4. Reliability
   (a) Collect the data in identical manner across treatments!
   (b) Align incentives for good reporting
   (c) Is the question socially desirable? → Use proxy indicator

4.4 Field Testing Outcomes Measures

- Have we chosen right respondents?
- Do your instrument pickup variation?
- Is the plan appropriate given the context?
  1. Administrative data can be unreliable
  2. Survey might be to long
  3. Incorrect recall period
  4. Time and place to survey
  5. Understanding of question depends on context

Example: Does conversational interviewing reduce survey measurement error? (Schober and Conrad, 1997)

Problem: People interpret the questions on their own **in the standardized interview** that result in misunderstanding.

What if interviewer can explain the question to respondent, use **conversational interviewing?**

Study:

- Interviewers provide either **standardized** or **conversational** interview on the same set of questions.
- The correct answer on the questions can be determined since questions use fictitious scenarios.

Findings:

- **Straightforward questions:** No difference between two methods in accuracy of answering (about 98%).
- **Complicated questions:** Only 28% of correct answers for standardized interview, whereas 87% for conversational one.

4.5 Nonsurvey instruments

4.5.1 Direct Observation

A.Random spot checks

When are these useful? When subjects have incentives to hide information

Limitations:

- Expensive(Large number of observations)
- Phenomena must be quickly observed

B. Mystery Clients

When are these useful? Antisocial or illegal activity

Limitations: People may change behavior in response to mystery clients

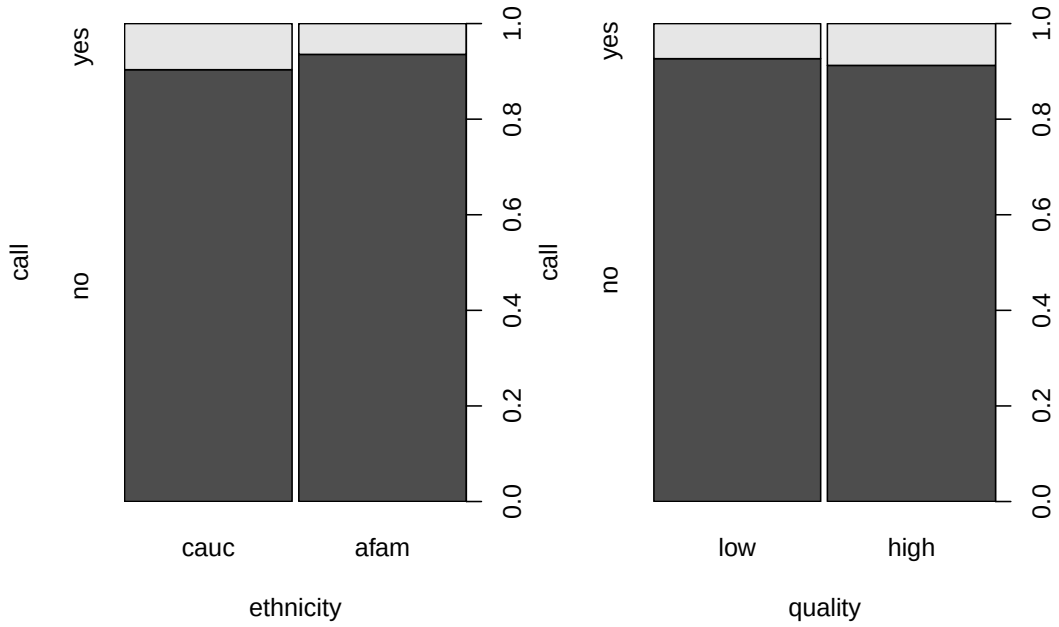
Example: Are Emily and Greg More Employable Than Lakisha and Jamal?(Bertrand and Mullainathan, 2004)

Problem: Discrimination at the job market

Method: Send fictions CV randomly vary 'name and quality; observe rate of call-back.

Table 13: Determinants of call-back

	Dependent variable:
	Call-back
African Sounding Name	-0.032^{***}
	(0.008)
High Quality	0.014^{*}
	(0.008)
Constant	0.089^{***}
	(0.007)
Observations	4,870
R^2	0.004
Adjusted R^2	0.004
Note:	*p<0.1; **p<0.05; ***p<0.01



C. Incognito enumerators(ride-alongs)

When are these useful? We want to observe the whole process

Limitations: Enumerator may affect the behavior

D. Observer group interaction

When are these useful? We want to know group behavior

Limitations: Enumerator may affect the behavior

4.5.2 Nondirect Observation

A. Physical tests e.g check materials, check speed.

When are these useful? Data in objective manner

Limitations: Physical tests measure one specific outcome and can be expensive

B. Biomarkers e.g. HIV test, saliva test.

When are these useful? Objective data about health conditions

Limitations: Expensive, logistically complicated, ethical issues

C. Mechanical tracking devices e.g. GPS unit.

When are these useful? Mechanical devices can overcome the problem of distance

Limitations: Can change the behavior, can brake.

D. Spatial demography e.g. GPS readings, satelite images.

When are these useful? Allows make use of distance in analysis

Limitations: Not travel time.

E. Games e.g. trust game, public good game.

When are these useful? Test theories about response to different incentives

Limitations: Generalisability?

Example: Democratic institutions and collective action capacity (Fearon et al., 2009).

Method: Use public good game to measure the effect of democratic governance institution introduction.

Findings: Higher levels of cooperation in public good game in treated communities.

F. List randomization.

When are these useful? Elicit answers on sensitive questions

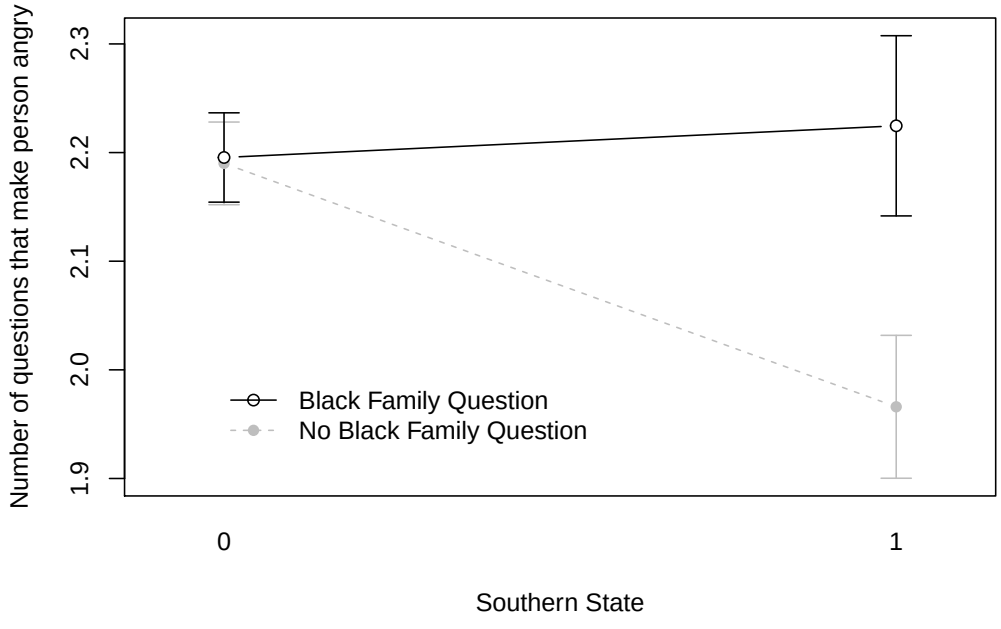
Limitations: Only aggregate measure.

Example 1: The 1991 National Race and Politics Survey (U.S.A.)

Question

Now I'm going to read you four things that sometimes make people angry or upset. After I read all (three/four), just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

1. "the federal government increasing the tax on gasoline;"
2. "professional athletes getting million-dollar-plus salaries;"
3. "large corporations polluting the environment;"
4. **"a black family moving next door to you."**

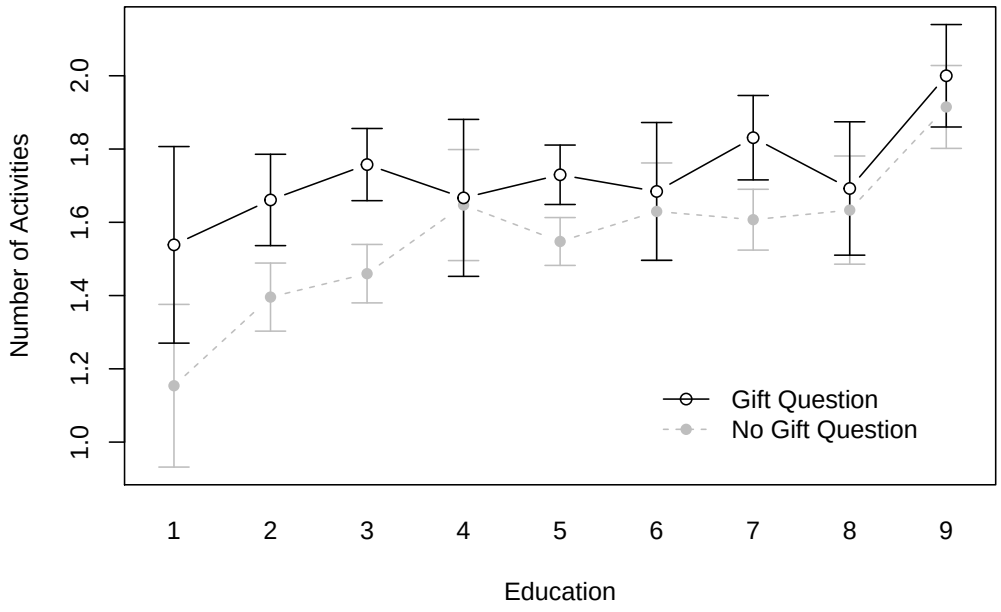


Example 2: The 2012 Mexico Elections Panel Study (U.S.A.)

Question

I am going to read you a list of four activities that appear on this card and I want you to tell me how many of these activities you have done in recent weeks. Please don't tell me which ones, just HOW MANY.

1. See television news that mentions a candidate
2. Attend a campaign event
3. **Exchange your vote for a gift, favor, or access to a service**
4. Talk about politics with other people



G. Endorsement Experiment.

When are these useful? Elicit answers on sensitive questions

Limitations: Only aggregate measure.

Example Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan(Blair et al., 2015)

Questions. Control without "by ISAF"; Treatment with "by ISAF"

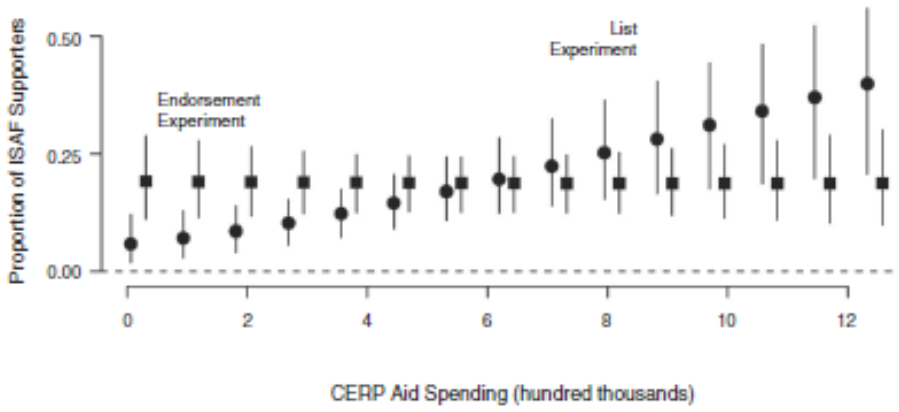
A recent proposal **by ISAF** calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

Questions.

I'am going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you

broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please do not tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

- Karzai Government;
- National Solidarity Program;
- Local Farmers;
- **ISAF**



H. Vignettes.

When are these useful? Elicit unstated biases

Limitations: Expensive, Only aggregate measure.

I. Implicit Association Test.

When are these useful? Elicit unstated biases

Limitations: Approximation

J. Standardized Tests.

When are these useful? Knowledge of large number of people

Limitations: Only intermediate outcome

K. Data Patterns to Check for Cheating.

When are these useful? Check for cheating by participant or enumerators

Limitations: Suspicious patterns is not necessary cheating

L. Network Information.

When are these useful? Measure peer effect, spillovers et.c.

Limitations: Typically, we can get only unprecise picture of social network.

4.6 Summary

- Outcomes, indicators, variables, respondent
- Data source
  - Administrative data
  - Survey data: Respondent, Enumerator, Time
- Assess and Field Test Outcomes Measures

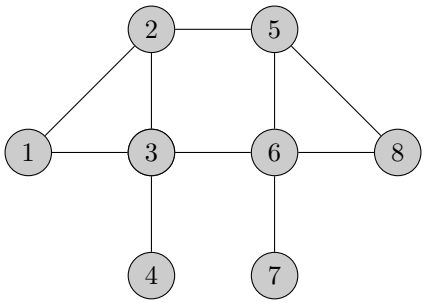


Figure 5: Recorded Social Network

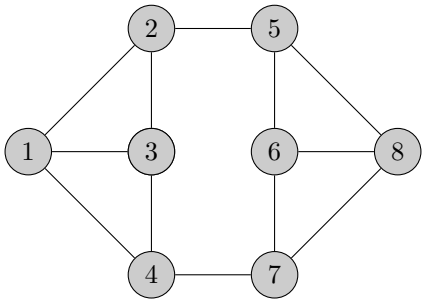


Figure 6: True Social Network

4.7 Exercises

1. **Outcomes and Indicators:** What is the difference between outcome, indicator, variable, respondent?
2. **Data Sources I:** What are the main two data sources we can use?
3. **Data Sources II:** How to make sample representative?
4. **Data Sources III:** Suppose you make an evaluation of the program that aims to reduce racial discrimination in two states: Alabama and Wyoming. You use program stuff to provide a survey about racial attitudes. What can be the problems? How would you deal with it?
5. **Data Sources IV:** How to deal with potential cheating of enumerators?
6. **Data Sources V:** Why is it useful to provide a baseline survey?
7. **Data Sources V:** You are evaluating advertisement campaign for the presidential elections. In your case the advertisement campaign starts 516 days before elections. You have already provided a baseline survey. When will you start to provide other surveys? How frequently?

Source: http://fivethirtyeight.com/

7. **Field Testing Outcomes Measures:** Why is it important to field test a survey?
8. **Nonsurvey instruments I:** Which four nonsurvey measures that use direct observation you know?
9. **Nonsurvey instruments II:** You want to elicit cheating **attitudes**, dishonesty among politicians. What can you use?
10. **Nonsurvey instruments III:** You want to elicit level of cooperation but you do not have direct measure of it. What can you use instead?
11. **Assignment:** Which outcomes will you use in your project?

Table 14: POLLING ACCURACY A YEAR BEFORE THE ELECTION

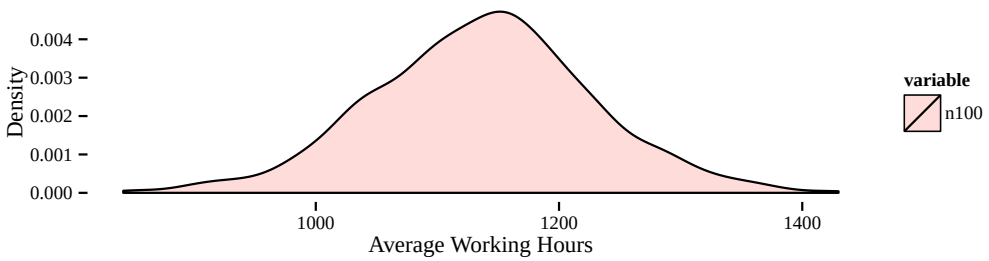
ELECTION	AVERAGE GOP POLL LEAD	GOP ELECTION MARGIN	ABSOLUTE ERROR
1964	-50.3	-22.6	27.7
1992	+21.0	-5.6	26.1
1980	-15.5	+9.7	25.2
2000	+11.9	-0.5	12.4
1984	+7.2	+18.2	11.0
1988	+18.0	+7.7	10.3
2008	-0.3	-7.3	6.9
1956	+22.0	+15.4	6.6
1944	-14.0	-7.5	6.5
2004	+8.7	+2.5	6.2
1996	-13.0	-8.5	4.5
1960	+3.0	-0.2	3.2
2012	-2.8	-3.9	1.0
1948	-3.8	-4.5	0.7
Average			10.6

5 Power Analysis

5.1 Sample Variation

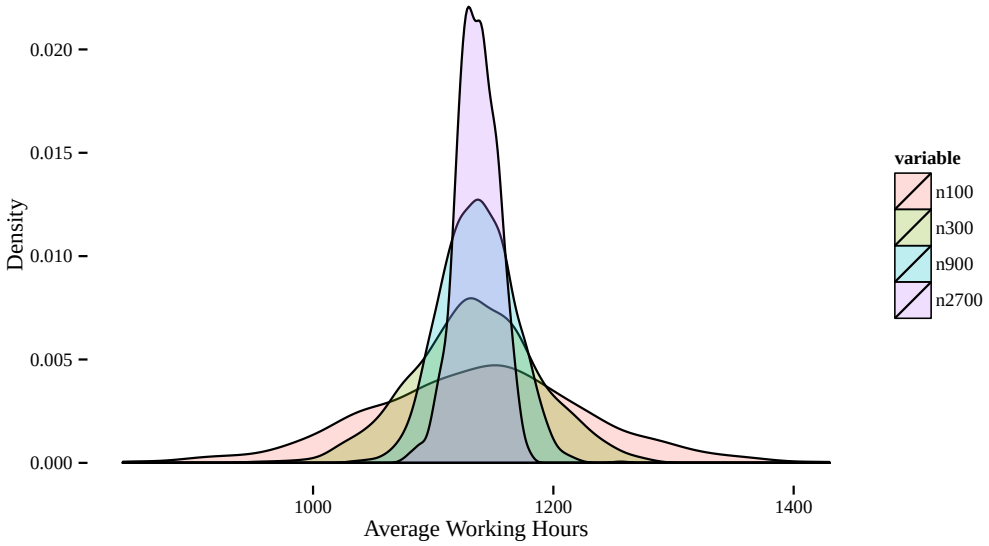
```
library(Ecdat)
data(Workinghours)
# Make sample of 'hours'
s1 <- sample(Workinghours$hours, 100, replace = TRUE)
str(s1)

##  int [1:100] 2495 0 0 1566 40 1584 1175 1944 0 2000 ...

# Make 1000 samples of 'hours' of size 100
n100 <- replicate(1000, mean(sample(Workinghours$hours, 100, replace = TRUE)))
```

```
require(ggplot2)
require(ggthemes)
require(reshape2)
ggplot(melt(data.frame(n100)), aes(x = value, fill = variable)) + geom_density(alpha = 0.25) +
    theme_tufte() + xlab("Average Working Hours") + ylab("Density")
```




```
n300 <- replicate(1000, mean(sample(Workinghours$hours, 300, replace = TRUE)))
n900 <- replicate(1000, mean(sample(Workinghours$hours, 900, replace = TRUE)))
n2700 <- replicate(1000, mean(sample(Workinghours$hours, 2700, replace = TRUE)))
```

```
ggplot(melt(data.frame(n100, n300, n900, n2700)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Working Hours") +
    ylab("Density")
```



Standard Deviation

$$sd = \sqrt{\frac{\sum_{n}^{i=1}(x_i - \bar{x})^2}{n-1}}$$

```
s100 <- sample(Workinghours$hours, 100, replace = TRUE)
s300 <- sample(Workinghours$hours, 300, replace = TRUE)
s900 <- sample(Workinghours$hours, 900, replace = TRUE)
s2700 <- sample(Workinghours$hours, 2700, replace = TRUE)
sqrt(sum((s100 - mean(s100))^2)/99)

## [1] 868.2644

sd(s100)

## [1] 868.2644

sd(s300)

## [1] 878.3049

sd(s900)

## [1] 889.2721

sd(s2700)

## [1] 876.5436
```

```
sd100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE)))
sd300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE)))
sd900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE)))
sd2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700, replace = TRUE)))
```

```
ggplot(melt(data.frame(sd100, sd300, sd900, sd2700)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Standard Deviation of Working Hours") +
    ylab("Density")
```

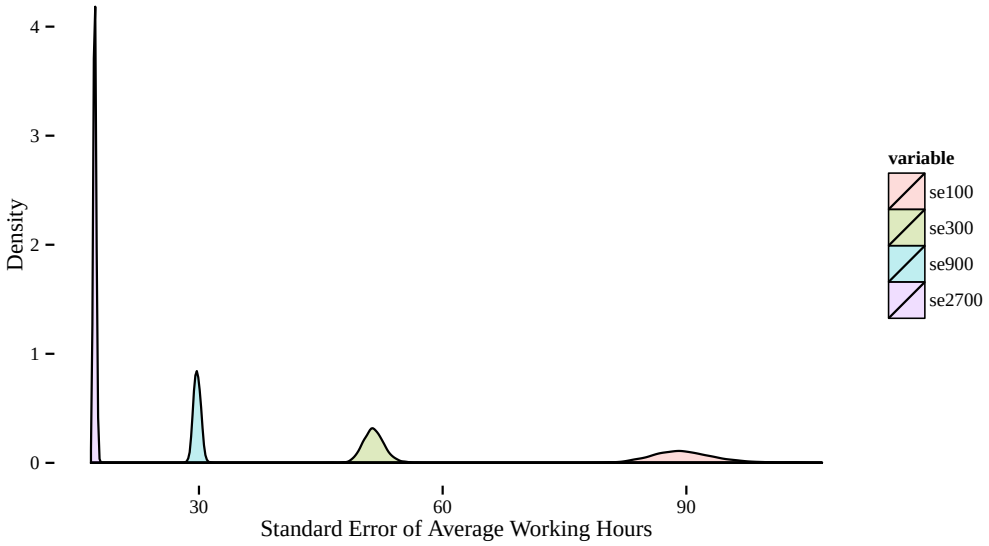
Standard Error of the Mean

$$se = \frac{sd}{\sqrt{n-1}}$$

```
sd(s100)/sqrt(length(s100) - 1)

## [1] 87.26386

sd(s300)/sqrt(length(s300) - 1)

## [1] 50.79368

sd(s900)/sqrt(length(s900) - 1)

## [1] 29.65888

sd(s2700)/sqrt(length(s2700) - 1)

## [1] 16.87222
```

```
se100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE))/sqrt(99))
se300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE))/sqrt(299))
se900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE))/sqrt(899))
se2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700, replace = TRUE))/sqrt(2699))
```

```
ggplot(melt(data.frame(se100, se300, se900, se2700)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Standard Error of Average Working Hours") +
    ylab("Density")
```



Central Limit Theorem

If $Y_1, \ldots, Y_n$ i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of $\bar{Y}$ approximates a normal distribution

Confidence intervals

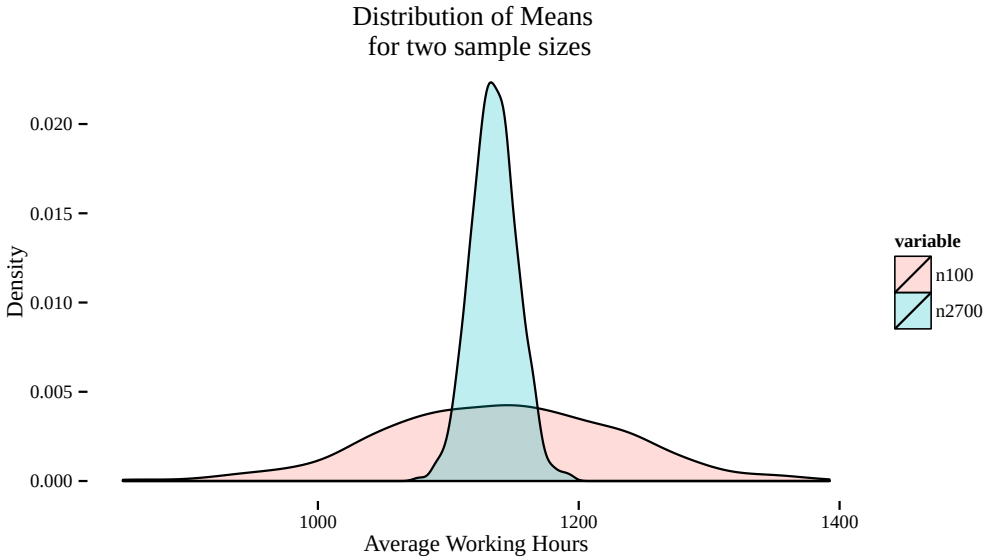
Upper 95% limit $= \bar{x} + 1.96 \cdot se$

Lower 95% limit $= \bar{x} - 1.96 \cdot se$

```
mean(s100) - 1.96 * sd(s100)/sqrt(length(s100) - 1)

## [1] 1019.993

mean(s100) + 1.96 * sd(s100)/sqrt(length(s100) - 1)

## [1] 1362.067

mean(s2700) - 1.96 * sd(s2700)/sqrt(length(s2700) - 1)

## [1] 1136.188

mean(s2700) + 1.96 * sd(s2700)/sqrt(length(s2700) - 1)

## [1] 1202.327
```

Table 15: Error Types

2*		Null Hypothesis, H_0	
		Valid	Invalid
2*Judgement of H_0	Reject	False Positive (Type Error I)	Correct Inference
	Fail to Reject	Correct Inference	False Negative (Type Error II)



Note: The exact confidence interval is slightly different since we have to correct for the fact that we sample from larger sample.

5.2 Hypothesis Testing

1. Formulate Null-Hypothesis e.g. Difference in means is zero. $H_0 : E(Y) = \mu_{Y,0}$
2. Formulate alternative hypothesis:
   - $H_1 : E(Y) \neq \mu_{Y,0}$. (Two-sided test)
   - $H_1 : E(Y) > \mu_{Y,0}$. (One-sided test)
   - $H_1 : E(Y) < \mu_{Y,0}$. (One-sided test)

What can happen then?

- α, significance level – predefined probability of rejecting H_0 despite it being true, typically 5%.
- k – predefined probability of failing to reject H_0 when it is false.
- (1-k), power – predefined probability of rejecting H_0 when it is false, typically 80%

5.3 Determinants of Power

Returning to the STAR experiment.

Let's estimate the distribution of the mean from the same population twice with different sample sizes:

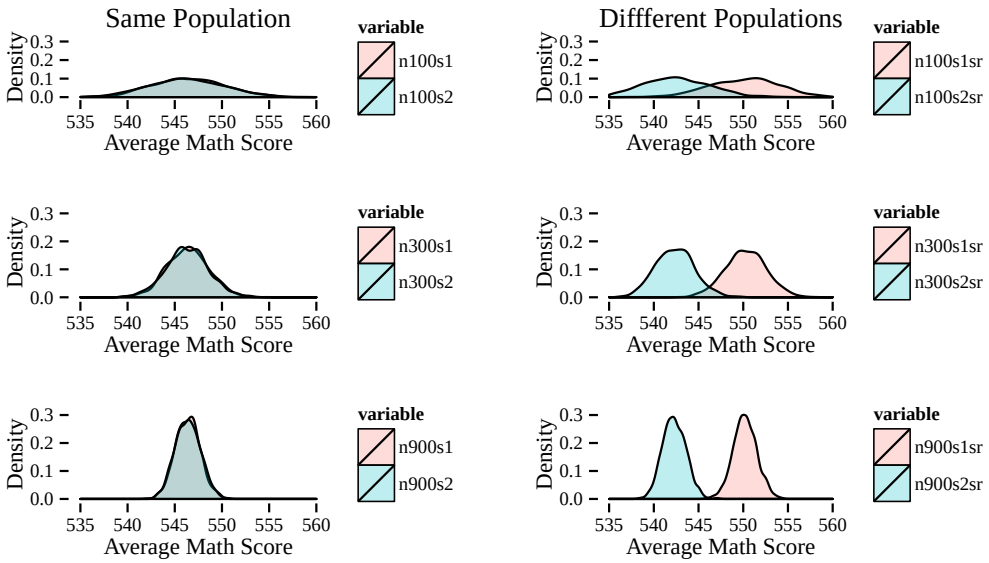
```
require(AER)
data("STAR")


n100s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 100, replace = TRUE)))
n100s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 100, replace = TRUE)))

n300s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 300, replace = TRUE)))
n300s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 300, replace = TRUE)))

n900s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 900, replace = TRUE)))
n900s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small" |
    star1 == "regular")$math1, 900, replace = TRUE)))
```

Now from different populations:

```
n100s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
    100, replace = TRUE)))
n100s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
    100, replace = TRUE)))

n300s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
    300, replace = TRUE)))
n300s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
    300, replace = TRUE)))

n900s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
    900, replace = TRUE)))
n900s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
    900, replace = TRUE)))
```

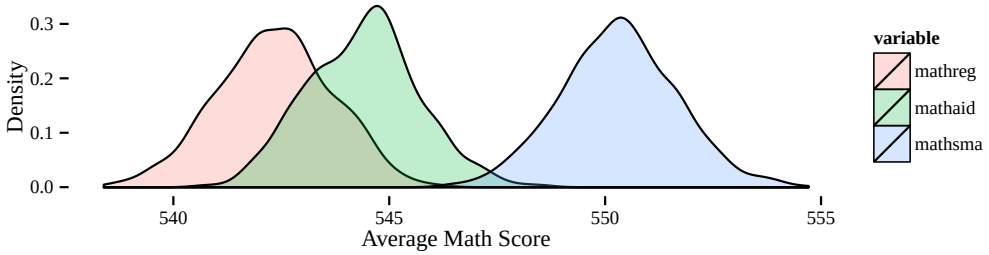
```
plot100 <- ggplot(melt(data.frame(n100s1, n100s2)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560) + ggtitle("Same Population")
plot300 <- ggplot(melt(data.frame(n300s1, n300s2)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560)
plot900 <- ggplot(melt(data.frame(n900s1, n900s2)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560)

plot100sr <- ggplot(melt(data.frame(n100s1sr, n100s2sr)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560) + ggtitle("Diffferent Populations")
plot300sr <- ggplot(melt(data.frame(n300s1sr, n300s2sr)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560)
plot900sr <- ggplot(melt(data.frame(n900s1sr, n900s2sr)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + ylim(0, 0.32) + xlim(535, 560)
```

```
library(gridExtra)
grid.arrange(plot100, plot100sr, plot300, plot300sr, plot900, plot900sr, ncol = 2,
    nrow = 3)
```

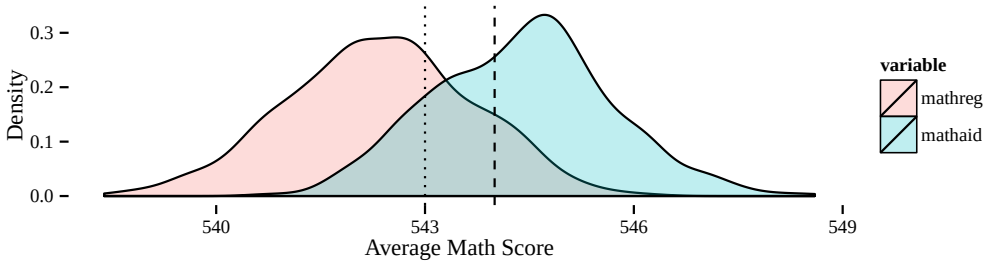



```
tapply(na.omit(STAR)$math1, na.omit(STAR)$star1, mean)

##      regular        small regular+aide
##     542.3272     550.2633     544.4835

mathreg <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
    900, replace = TRUE)))
mathaid <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular+aide")$math1,
    900, replace = TRUE)))
mathsma <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
    900, replace = TRUE)))
```

```
ggplot(melt(data.frame(mathreg, mathaid, mathsma)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density")
```



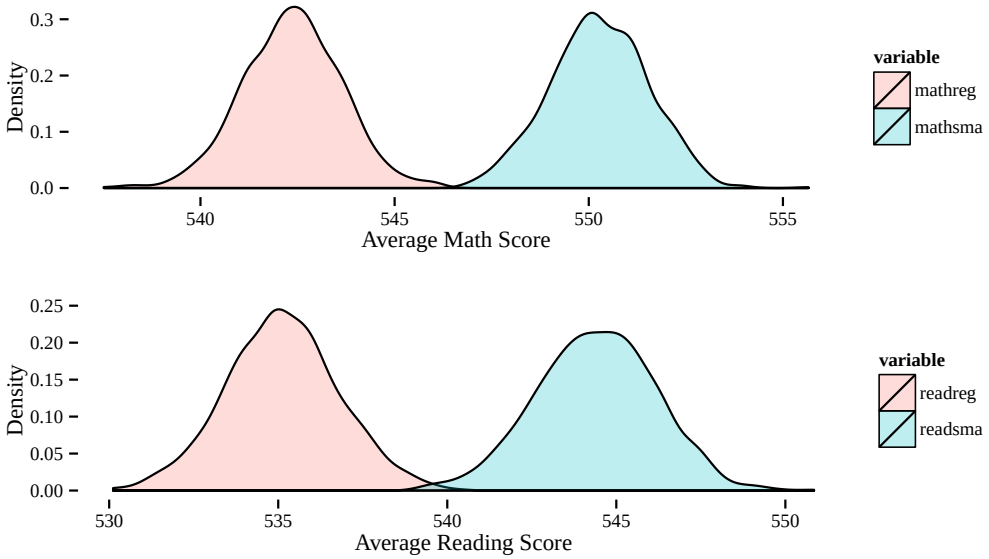

```
ggplot(melt(data.frame(mathreg, mathaid)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density") + geom_vline(xintercept = 543, linetype = 3) + geom_vline(xintercept = 544,
    linetype = 2)
```




```
var(subset(na.omit(STAR))$math1)

## [1] 1555.032

var(subset(na.omit(STAR))$read1)

## [1] 2731.149

mathreg <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
    900, replace = TRUE)))
mathsma <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
    900, replace = TRUE)))

readreg <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "regular")$read1,
    900, replace = TRUE)))
readsma <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 == "small")$read1,
    900, replace = TRUE)))
```

```
plotmath <- ggplot(melt(data.frame(mathreg, mathsma)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
    ylab("Density")
plotread <- ggplot(melt(data.frame(readreg, readsma)), aes(x = value, fill = variable)) +
    geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Reading Score") +
    ylab("Density")
```

```
grid.arrange(plotmath, plotread)
```

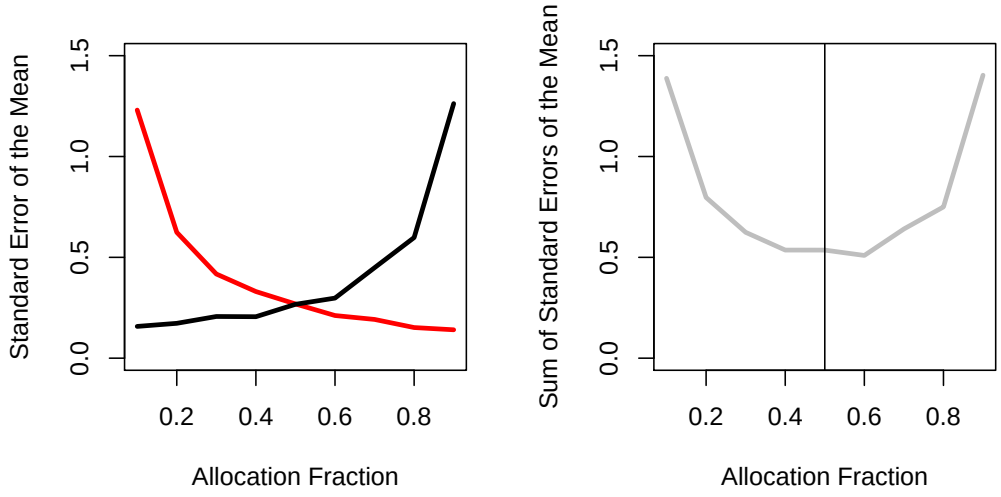



```
powerfraction <- function(P) {
    n <- 300
    sesma <- sd(replicate(100, mean(sample(subset(na.omit(STAR), star1 == "small")$math1,
        n * P, replace = TRUE))))/sqrt(n * P - 1)
    sereg <- sd(replicate(100, mean(sample(subset(na.omit(STAR), star1 == "regular")$math1,
        n * (1 - P), replace = TRUE))))/sqrt(n * (1 - P) - 1)
    as.data.frame(cbind(sesma, sereg, sesma + sereg))
}
powerfraction(0.4)
```

```
##       sesma     sereg       V3
## 1 0.3740136 0.2088594 0.582873
```

```
dd <- sapply(c((1:9)/10), powerfraction)
```

```
par(mfrow = c(1, 2))
plot(c((1:9)/10), dd[1, ], type = "l", col = "red", lwd = 3, xlab = "Allocation Fraction",
    ylab = "Standard Error of the Mean", ylim = c(0, 1.5))
lines(c((1:9)/10), dd[2, ], lwd = 3)
plot(c((1:9)/10), dd[3, ], type = "l", lwd = 3, xlab = "Allocation Fraction",
    ylab = "Sum of Standard Errors of the Mean", col = "grey", ylim = c(0, 1.5))
abline(v = 0.5)
```



5.4 Algebra of Determinants of Power

5.4.1 Individual Level Randomization

$$Y = \beta_0 + \beta_T T + \sum_{i=1}^{k} \beta_{Ci} C_i + u$$

Variance

$$Variance(\hat{\beta_T}) = \frac{1}{P(1-P)} \frac{\sigma^2}{N}$$

, where N – number of observations; σ – standard error term; P – allocation fraction

Critical Value, $\mathbf{t_{\alpha/2}}$

$$\Phi(t_{\alpha/2}) = 1 - \alpha/2$$

, where ϕ is standard normal cumulative distribution function; α – significance level.

We can reject the Null-hypothesis if $\frac{|\bar{\beta_T}|}{SE(\hat{\beta_T})} > t_{\alpha/2}$

Minimum Detectable Effect Size

$$MDE = (t_{(1-k)} + t_{\alpha/2}) \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{N}}$$

Power

$$1 - k = \Phi([MDE \times \sqrt{P(1-P)} \times \sqrt{\frac{N}{\sigma^2}}] - t_{\alpha/2})$$

So, we see:

- More observations(N) more power
- Higher MDE gives higher power
- Lower $t_{\alpha/2}$ or significance level gives higher power
- Smaller variance gives higher power
- Power is maximized if $P = 0.5$

5.4.2 Group Level Randomization

$$Y_{i,j} = \beta_0 + \beta_T T + \sum_{i=1}^{k} \beta_{Ci} C_i + v_j + w_{i,j}$$

, where v_j – shocks at cluster(group) level which we assume to be i.i.d. with variance τ^2, $w_{i,j}$ – shocks at individual level within a cluster $w_{i,j}$, which we assume to be i.i.d. with variance σ^2

If $\rho = \frac{\tau^2}{\tau^2+\sigma^2}$ – intercluster correlation $\Rightarrow$

Minimum Detectable Effect Size

$$MDE = \frac{t_{(1-k)}+t_{\alpha/2}}{\sqrt{P(1-P)}} \times \sqrt{\rho + \frac{1-\rho}{n}\sigma}$$

,where n – cluster size.

5.5 Performing Power Analysis

1. Determine desired power e.g. 80%.
2. Determine Minimum Detectable Effect Size.

$$\text{Standardized Effect Size} = \frac{\bar{Y}_T - \bar{Y}_C}{SD_Y}$$

   - Using "Standard" Effect Size
   - Comparing various MDE sizes to those of interventions with similar objectives
   - Effect size with respect to cost effectiveness
3. Choosing the number of clusters.
4. Non-equal allocation fraction ($P_T \neq P_C$):
   - When the costs (C) of treatments are different. $\frac{P_T}{P_C} = \sqrt{\frac{C_C}{C_T}}$
   - When the MDE varies by treatment
   - When the comparison group is important
   - When the variance is different across treatments (Be cautious in future statistical analysis!)

Argument against Randomized Control Trials:

"...the standard error from the OLS regression is not correct unless the variance in the experimental group is identical to the variance in the control group, which will only be true if the treatment has no effect on the variance, which will not generally be the case particularly if treatment responses are heterogeneous. "

(Deaton, 2009)

Reply:

"Using the standard ordinary least squares variance based on homoskedasticity leads to confidence intervals that are not necessarily justified even in large samples. This point is correct and, in practice, it is certainly recommended to use the robust variance here, at least in sufficiently large samples." (Imbens, 2010)

5. Calculating residual variance
6. Number of repeat samples
7. Temporal Correlation
8. Intrauterine Correlation

5.6 Tools for Power Calculation

Tools:

- G*power
- Optimal Design
- STATA e.g. `sampsi 0.43 0.45, power(0.8) sd(0.05)`
- R e.g.

Good for standard design, but what to do if we have more specific experiment?

Simulation

5.7 Simulation to Determine Power

Let's make a function that generate a data from our experiment. Suppose it is experiment where we try to change the math score. We have number of groups (`Group`), subjects (`id`), two treatments (`Treat`), outcome measure (`Score`) and its standard devation `sdscore` measured in number of surves (`NS`), and some residual variance (`noise`) with standard deviation `sd`.

```
DGP <- function(n, g, NS, meanscore, sdscore, eff, e, sd) {
    Group <- as.factor(rep(1:g, each = n/g * NS))
    id <- as.factor(rep(1:n, each = NS))
    Treat <- rep(rep(0:1, each = NS), n)
    score <- as.vector(replicate(n/2, c(round(rnorm(NS, mean = meanscore, sd = sdscore)),
        round(rnorm(NS, mean = meanscore + eff, sd = sdscore)))))


    noise <- rnorm(n, e, sd)
    Score <- score + noise

    data.frame(Group, id, Treat, Score)
}

head(DGP(100, 10, 4, 5, 1, 0.2, 0.5, 1))

##   Group id Treat    Score
## 1     1  1     0 4.567103
## 2     1  1     0 5.680127
## 3     1  1     0 4.790665
## 4     1  1     0 5.079173
## 5     1  2     1 3.915156
## 6     1  2     1 4.818426
```

Here is the function which run mixed-effect model over our data and return p-value for the treatment:

```
require(lme4)
## SIMULATION LMER ANALYSIS
oneSimulG <- function(n, g, NS, meanscore, sdscore, eff, e, sd) {
    DATA <- DGP(n, g, NS, meanscore, sdscore, eff, e, sd)
    mm.lmer <- lmer(Score ~ Treat + (1 | Group) + (1 | id), data = DATA)


    2 * (1 - pnorm(abs(coef(summary(mm.lmer))["Treat", "t value"])))

}

oneSimulG(10, 2, 4, 2, 1, 2, 0.5, 1)

## [1] 9.190586e-09
```

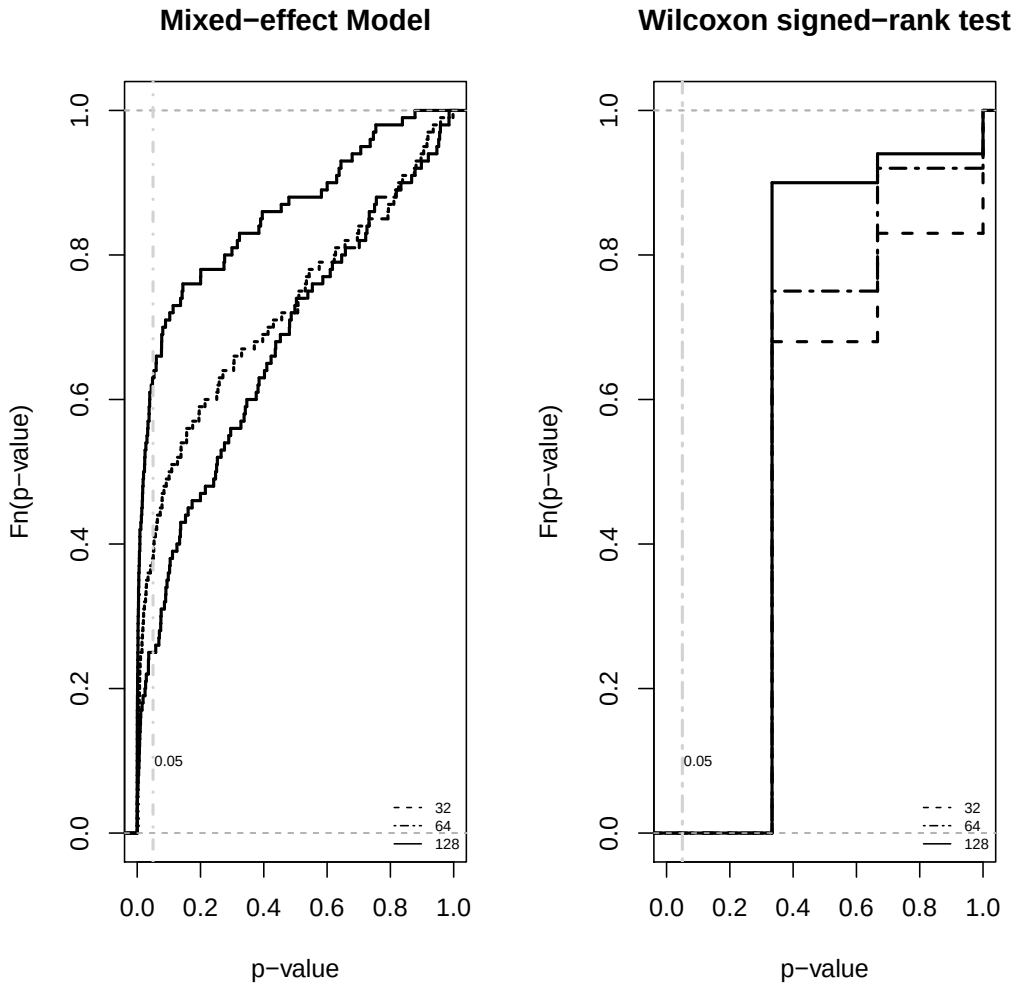
Here is the function which provides the exact Wilcoxon signed-rank test over our data and return p-value for the treatment:

```
require(exactRankTests)
## SIMULATION WILCOX ANALYSIS
oneSimulW <- function(n, g, NS, meanscore, sdscore, eff, e, sd) {
    DATA <- DGP(n, g, NS, meanscore, sdscore, eff, e, sd)
    CONT <- with(subset(DATA, DATA$Treat == 0), tapply(Score, Group, mean))
    TREAT <- with(subset(DATA, DATA$Treat == 1), tapply(Score, Group, mean))
    wilcox.exact(CONT, TREAT)$p.value
}

oneSimulW(10, 2, 4, 2, 1, 2, 0.5, 1)

## [1] 0.3333333
```

Now, let's provide these analysis 100 times for different sample sizes of 32,64, and 128. We would like to know how many times we will get the p-value below 5%.

```
par(mfrow = c(1, 2))

## ECDF LMER####
pvals1 <- replicate(100, oneSimulG(n = 32, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulG(n = 64, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulG(n = 128, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 4)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, legend = c("32", "64", "128"), box.col = NA,
    lty = c(2, 6, 1))
title(main = "Mixed-effect Model")

mean(pvals3 < 0.05)
```

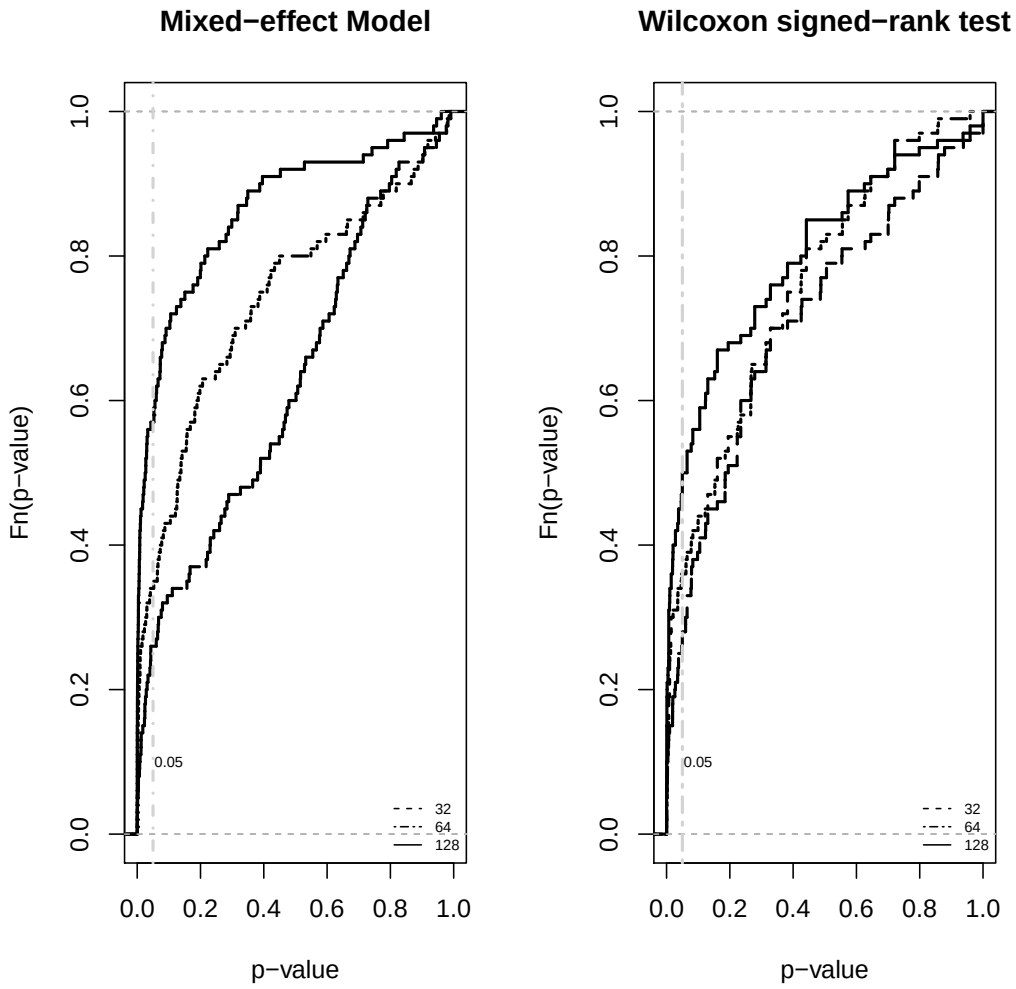
```
## [1] 0.63

## ECDF WILCOX####
pvals1 <- replicate(100, oneSimulW(n = 32, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulW(n = 64, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulW(n = 128, g = 2, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 6)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, box.col = NA, legend = c("32", "64", "128"),
    lty = c(2, 6, 1))
title(main = "Wilcoxon signed-rank test")
```



```r
mean(pvals3 < 0.05)

## [1] 0
```

The experiment is underpowered even with the sample of 128 subjects. Let's increase the group number.

```r
par(mfrow = c(1, 2))
## ECDF LMER####

pvals1 <- replicate(100, oneSimulG(n = 32, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulG(n = 64, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulG(n = 128, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 4)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, legend = c("32", "64", "128"), box.col = NA,
    lty = c(2, 6, 1))
title(main = "Mixed-effect Model")

mean(pvals3 < 0.05)

## [1] 0.57

## ECDF WILCOX####

pvals1 <- replicate(100, oneSimulW(n = 32, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulW(n = 64, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulW(n = 128, g = 8, NS = 2, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 6)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, box.col = NA, legend = c("32", "64", "128"),
    lty = c(2, 6, 1))
title(main = "Wilcoxon signed-rank test")
```



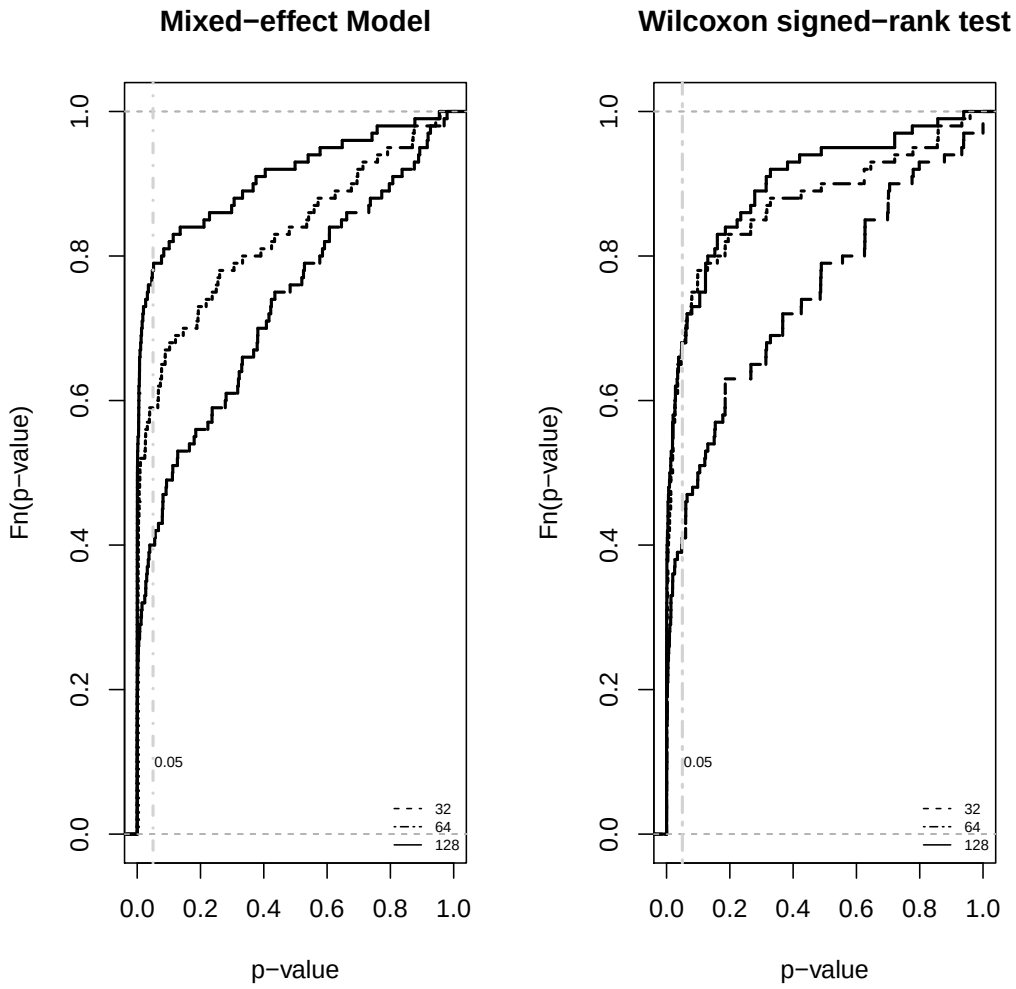

```
mean(pvals3 < 0.05)

## [1] 0.5
```

The experiment is still underpowered but we observe improvemtn in case of Wilcoxon test.

Let's increase the number of surveys (`NS`) from 2 to 4:

```
par(mfrow = c(1, 2))
## ECDF LMER####

pvals1 <- replicate(100, oneSimulG(n = 32, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulG(n = 64, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulG(n = 128, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
```

```
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 4)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, legend = c("32", "64", "128"), box.col = NA,
    lty = c(2, 6, 1))
title(main = "Mixed-effect Model")
mean(pvals3 < 0.05)

## [1] 0.78

## ECDF WILCOX####

pvals1 <- replicate(100, oneSimulW(n = 32, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals2 <- replicate(100, oneSimulW(n = 64, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
pvals3 <- replicate(100, oneSimulW(n = 128, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    eff = 0.4, e = 0, sd = 1))
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 6)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, box.col = NA, legend = c("32", "64", "128"),
    lty = c(2, 6, 1))
title(main = "Wilcoxon signed-rank test")
```




```
mean(pvals3 < 0.05)

## [1] 0.68
```

Let's suppose we had made a pre-survey and get a data on mean IQ level **iq** and its standard deviation **sdiq**. We assume that IQ is a good predictor of math score. We make a function for data generating process with IQ:

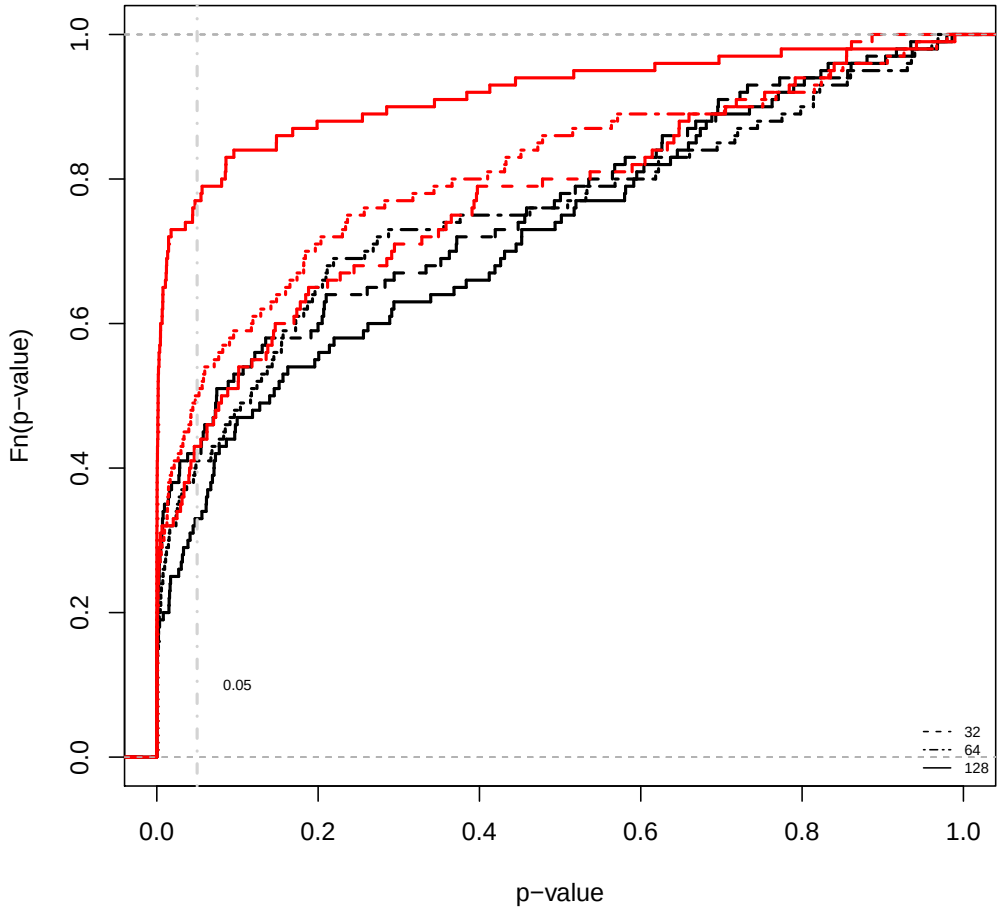
```
DGPiq <- function(n, g, NS, meanscore, sdscore, eff, e, sd, iq, sdiq) {
    Group <- as.factor(rep(1:g, each = n/g * NS))
    id <- as.factor(rep(1:n, each = NS))
    Treat <- rep(rep(0:1, each = NS/2), n)
    score <- as.vector(replicate(n, c(round(rnorm(NS/2, mean = meanscore, sd = sdscore)),
        round(rnorm(NS/2, mean = meanscore + eff, sd = sdscore)))))

    iq <- rnorm(n, iq, sdiq)
    noise <- rnorm(n, e, sd)
    Score <- score + iq + noise

    data.frame(Group, id, Treat, Score, iq)
}
```

And than make function that run regression with IQ as control

```
oneSimulGC <- function(n, g, NS, meanscore, sdscore, eff, e, sd, iq, sdiq) {
    DATA <- DGPiq(n, g, NS, meanscore, sdscore, eff, e, sd, iq, sdiq)
    mm.lmer <- lmer(Score ~ Treat + (1 | Group) + (1 | id), data = DATA)

    mm.lmerC <- lmer(Score ~ Treat + iq + (1 | Group) + (1 | id), data = DATA)
    p <- 2 * (1 - pnorm(abs(coef(summary(mm.lmer))["Treat", "t value"])))
    pc <- 2 * (1 - pnorm(abs(coef(summary(mm.lmerC))["Treat", "t value"])))
    as.data.frame(cbind(p, pc))
}
oneSimulGC(n = 32, g = 8, NS = 4, meanscore = 5, sdscore = 1, ef = 0.4, e = 0,
    sd = 1, iq = 4, sdiq = 4)

##           p        pc
## 1 0.1262276 0.2776844
```

We provide a simulation and see how liklely we will get the p-value lower than 0.05 with (red lines) and without (black lines) IQ level as control variable.

```
pvals1 <- replicate(100, oneSimulGC(n = 32, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$p)
pvals2 <- replicate(100, oneSimulGC(n = 64, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$p)
pvals3 <- replicate(100, oneSimulGC(n = 128, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$p)
pvals1C <- replicate(100, oneSimulGC(n = 32, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$pc)
pvals2C <- replicate(100, oneSimulGC(n = 64, g = 8, NS = 4, meanscore = 5, sdscore = 1,
    ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$pc)
pvals3C <- replicate(100, oneSimulGC(n = 128, g = 8, NS = 4, meanscore = 5,
    sdscore = 1, ef = 0.4, e = 0, sd = 1, iq = 4, sdiq = 5)$pc)
plot(ecdf(pvals1), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), main = NULL,
    xlab = "p-value", ylab = "Fn(p-value)", lwd = 2, lty = 2)
lines(ecdf(pvals2), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6)
lines(ecdf(pvals3), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1)
lines(ecdf(pvals1C), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 2, col = "red")
lines(ecdf(pvals2C), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 6, col = "red")
lines(ecdf(pvals3C), do.p = FALSE, verticals = TRUE, xlim = c(0, 1), ylab = "Fn(p-value)",
    lwd = 2, lty = 1, col = "red")
abline(v = 0.05, col = "lightgray", lwd = 2, lty = 4)
text(0.1, 0.1, expression(0.05), cex = 0.6)
legend("bottomright", cex = 0.6, legend = c("32", "64", "128"), box.col = NA,
    lty = c(2, 6, 1))
```




```
mean(pvals3 < 0.05)
```

```
## [1] 0.33
```

```
mean(pvals3C < 0.05)
```

```
## [1] 0.77
```

5.8 How to design high powered study?

When designing the study...

- Have decent sample size
- Fewer treatment groups
- Randomize at lowest possible level
- Use the design that increase compliance
- Use the design that limits attrition
- Balance sample e.g. stratification

- Choose right allocation fraction

When planning data collection...

- Choose good proxies
- Collect data on control variables
- Collect multilpe observation per person
- Plan data collection to limit attrition
- Limit procedural variation

5.9 Summary

- **Sample variability** requires to estimate how likely that we will find an effect – porvide power analysis
- **Power critically depends** on
  1. Sample size
  2. Effect size;
  3. Significance level
  4. Variance
  5. Allocation Fraction
- Use either standard packages or provide simluation to estimate power of the experiment
- **Consider how to increase the power**

5.10 Exercises

1. **Sample Variation I:** Recall the formula of variance and standard deviation. Will the standard deviation change with sample size? What is advantage of using standard deviation instead of variance?
2. **Sample Variation II:** Recall the formula of standard error of **the mean**. Will the standard error of the mean change with sample size? Why?
3. **Sample Variation III:** Recall the Central Limit Theorem and Confidence intervals. What is the relation between central limit theorem and confidence intervals?
4. **Hypothesis Testing:** Explain type error I and type error II. Explain significance level and power. Why typically in empirical analysis we talk about significance level but not power?
5. **Power I:** What is power of the experiment? What is the typical power level for the experiment? What are the advantages and disadvantages of higher power level?
6. **Power II:** What are the key five factors that determines the power of the study? In which direction?
7. **Power IIIa:** You run an randomized control trial to test adverse effect of a drug with 1000 subjects. You cannot reject the null-hypothesis at any conventional level of sginficance (p-value= 0.1239) that the difference between control and treatment group in adverse effect is diffferent from zero. What is your conclusion?
8. **Power IIIb:** You run a randomized control trial to test adverse effect of a drug with 1000 subjects that assumes 80% power at significance level of 5% if the minimum detectable effect size is 0.4SD. You cannot reject the null-hypothesis at any conventional level of sginficance (p-value= 0.1239) that the difference between control and treatment group in adverse effect is diffferent from zero. What is your conclusion?
9. **Power IV:** If you use group level randomization, is it better to have small or big clusters?
10. **Power V:** One can determine Minimum Detectable Effect Size. Which way would you choose?
11. **Power VI:** When would you choose non-equal allocation fractions?

12. **Power VII:** How to choose residual variance for your study?
    Power VIII: How to design high powered study?

13. **Power IX:** You want to estimate the sample size n needed for the experiment with power 80% and α 5%. You have two treatment groups and you assume the effect size is equal to 2. You have one observation per subject. The outcomes measure Y is equal to zero at baseline with $sd = 10$.You assume no residual noise. Provide a simaltion to estimate sample size:
    - Start to write function that generates data depending on sample size, `DGP<-function(n)`. Use command rep() to generate the treatment dummy `T`, than combine rnorm() and ifelse() to simulate the DGP.
    - Inside of the function estimate an OLS of treatment. Use lm() and safe the model as lm1.
    - Extract the p-value for the treatment dummy. Use `coef(summary(lm1))["T","Pr(>|t|)"]` in last command line of your function.
    - Use replicate() to run estimations 100 times with $n = 10$.
    - Calculate in which number of case p-value is below 0.05. `mean(p10<0.05)`

    Interpret the results. Did you achieve your goal? Why?

```
DGP <- function(n) {

    T <- c(rep(0, n/2), rep(1, n/2))

    y <- ifelse(T == 1, rnorm(n/2, 2, 10), rnorm(n/2, 0, 10))

    lm1 <- lm(y ~ T)

    coef(summary(lm1))["T", "Pr(>|t|)"]
}
DGP(10000)

p10 <- replicate(1000, DGP(1000))

mean(p10 < 0.05)
```

14. **Presentations:** E. Hauer. The harm done by tests of significance. Accident Analysis Prevention, 36:495500, 2004.

6 Threats

6.1 Partial Compliance

What is partial compliance?

- **When the people in the treated group are not treated.** For instance, Fertilizer is not delivered due to the impassable roads in the rainy season.
- **When the people in the treatment group do not complete the course.** For instance, people drop out from the course or farmer sell the fertilizer instead of using it.
- **When the people in comparison group receive the treatment:**
  - People in comparison group were receiving the treatment
  - People in comparison group move to location of the treatment
  - Outside actors provide similar treatment
- When the implementation stuff depart from location or procedures.
- **When the people – *defiers* – exhibit the opposite of compliance.**

How is noncompliance a threat?

- Noncompliance reduce the difference between treatment and control group.
- Noncomplinace reduce comparability between groups – selection bias.
- Defiers can make impossible to estimate the impact of the program.

How to limit partial compliance?

- Make take-up of the program **easy.**
- **Incentivize** take-up of the program. (But incentives may change outcome).
- **Compartmentalize** and routinize field tasks.
- Randomize the **at higher level.**
- Include **the basic program** that everyone receives.

How can we document compliance and identify defiers?

- **Document who receives what treatment**
  - Monitor process of the program
  - Endline survey

  → Monitoring only partially solve the problem and we should take into account **demand effect.**
- **Identify defiers:** Consider theory of change and develop indicators for this.

6.2 Attrition

What is the attrition? Attrition – absence of the data because researcher is unable to collect some or all outcome measures for some people in the sample

- People drop-out of the study e.g. die.
- People participate but can not be measured e.g. do not have time to answer the surveys.
- People refuse to answer on some questions e.g. questions bout illegal or sexual behavior.

How attrition is the threat?

- **Attrition reduce comparability.**
  Example: The weakest students in comparison group drop-out and refuse to answer surveys. Then, we do not have comparison group for weakest students in treated group. → our treatments are incomparable.
- **Attrition lowers statistical power.**

How can we limit the attrition?

- Use design that promise access to the program to all over time e.g. phase-in-design.
- Change the level of randomization
- Improve the data collection:
  - Pilot the data collection.
  - Follow up everyone originally randomized.
  - Do not wait to long for follow up.
  - Improve follow-up by tracking data routinely.e.g. ask if people plan to migrate and ask peers.
  - Choose right time for surveys.
  - Reduce attrition to sub sample e.g. random sample of people who drop-out
  - Provide incentives

6.3 Spillovers

Spillovers:

- Physical
- Informational
- Behavioral
- Market wide

We would like to (1) anticipate the spillovers and/or (2) measure them.

How spillovers are a threat?

→ Spillovers reduce the quality of contrafactual

How to manage spillovers?

- Identify potential spillovers
- Reduce spillovers to comparison group
- Estimate the spillover by measuring outcomes of non-treated units near to treated one e.g. neighbors

6.4 Evaluation-Driven Effects

1. **Hawthorne Effect:** Treatment group works harder than control.
   - Hawthorne works (Western Electric factory), 1924-1932.
   - Treatments: Change in lighting, 5-minutes breaks, providing food.
   - Results: Increased productivity due to the treatment, even if the variable set back to the original condition.

2. **John Henry Effect:** Treatment group compete with the control group.

   The plot
   John Henry is legendary steel driver who he heard that his output is compared to a steam drill and started to work so hard that he died.

3. **Resentment effect:** Comparison group is demoralized or resentful.
4. **Demand effect:** When the participants change their behavior in response to what they think evaluator wants.
5. **Anticipation effect:** Change in the behavior due to the expectation of future treatment.
6. **Survey effect:** Survey can change the behavior

Example: Interviewing changes attitudessometimes (Bridge et al., 1977)

Method: Telephone survey

Treatment: Questions about cancer or burglary prevention.

Results: Attitudes towards health has changed, whereas not for crime.

How Evaluation-Driven Effects are a threat?

- It can undermine generalizability
- It can reduce power
- It can undermine comparability
- Bias estimators

How to limit Evaluation-Driven Effects?

- Use different level of randomization.
- Do not announce the phase-in.
- Make sure that stuff is impartial and identical across groups.
- Measure the Evaluation-Driven Effects.

6.5 Summary

Threats:

- Partial Compliance
- Attrition
- Spillovers
- Evaluation-Driven Effects

6.6 Exercises

1. **Partial Complaince I:** What is partial compliance?
2. **Partial Complaince II:** How is partial compliance a threat?
3. **Partial Complaince III:** How to limit partial compliance?
4. **Attrition I:** What is the attrition?
5. **Attrition II:** How attrition is the threat?
6. **Attrition III:** How can we limit the attrition?
7. **Attrition IV:** You provide a randomized control trial and finilize it with a survey. You have 20 assistants who survey 50 respondents each. You randomly assign whom should they survey. Three of your asstistants get sick and one do not show up at all. The rest provide a survey as it is planned. What can you do in order to make a casual inference from your study?
8. **Spillovers I:** How to manage spillovers?
9. **Evaluation driven effects I:** What are the evaluation driven effects?
10. **Evaluation driven effects II:** How to limit evaluation driven effects?
11. **Assignment:** What can be the threats in your study?
12. **Presentation:** Zwane, Alix Peterson, et al. "Being surveyed can change later behavior and related parameter estimates." Proceedings of the National Academy of Sciences 108.5 (2011): 1821-1826.

7 Analysis

7.1 Basic Analysis

7.1.1 Basic Analysis

Before you start the analysis:

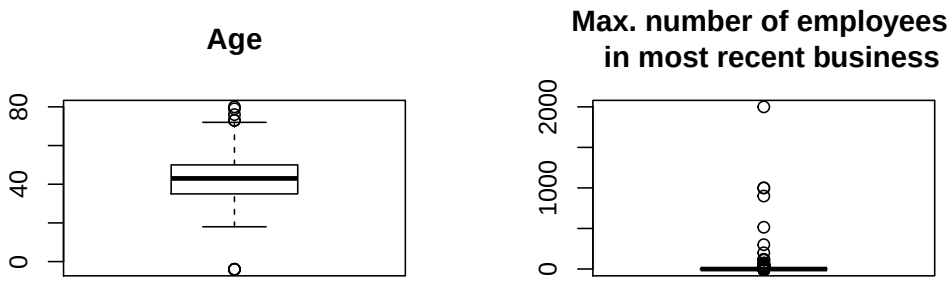
- Correct Errors e.g. inconsistent answers, answers out of range → Ask the questions again or input NA.
- Check for "outliers"
- Check Attrition Rate
- Plot and Describe the data e.g. use function summary(), boxplot(), hist()

Example: GATE project

```
require(sas7bdat)
a<-read.sas7bdat("GATE/application.sas7bdat")
```

```
summary(subset(a,treatment==1)[,c("age",  "gender","race_asian", "cb_maximum_employees" )])

##       age             gender          race_asian      cb_maximum_employees
##  Min.   :-4.00   Min.   :0.0000   Min.   :0.00000   Min.   :  -4.000
##  1st Qu.:34.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:  -1.000
##  Median :42.00   Median :1.0000   Median :0.00000   Median :  -1.000
##  Mean   :42.03   Mean   :0.5277   Mean   :0.04632   Mean   :   1.012
##  3rd Qu.:50.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:   0.000
##  Max.   :79.00   Max.   :1.0000   Max.   :1.00000   Max.   :1000.000
```

```
par(mfrow=c(1,2))
boxplot(a$age, main="Age")
boxplot(a$cb_maximum_employees,main="Max. number of employees \n  in most recent business")
```



7.1.2 Intention to Treat Analysis

Intention to Treat estimate average treatment effects – compare the mean outcomes of those who were randomized to receive the program with those of people randomized to comparison group.

```
require(sas7bdat)
w1<-read.sas7bdat("GATE/wave1.sas7bdat")
```

```
lmbplan<-lm(bpln~treatment, data=w1)
summary(lmbplan)

##
## Call:
## lm(formula = bpln ~ treatment, data = w1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4915 -0.4915 -0.3655  0.5085  0.6345
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36546    0.01241   29.45  < 2e-16 ***
## treatment    0.12600    0.01738    7.25 5.13e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5103 on 3447 degrees of freedom
## Multiple R-squared:  0.01502,Adjusted R-squared:  0.01473
## F-statistic: 52.56 on 1 and 3447 DF,  p-value: 5.134e-13
```

```
t.test(bpln~treatment, data=w1)

##
##  Welch Two Sample t-test
##
## data:  bpln by treatment
## t = -7.2553, df = 3447, p-value = 4.928e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.16005431 -0.09195239
## sample estimates:
## mean in group 0 mean in group 1
##       0.3654642       0.4914676
```

ITT effect has high external validity and therefore policy relevant. Policy-makers want to know the effect of the program not the effect of the program if everybody would attend everyday of the course or study hard et.c.

7.1.3 Including Covariates

```
dd<-merge(w1,a, by="gateid")
```

```
lmbplan<-lm(bpln~treatment.x, data=dd)
lmbplana<-lm(bpln~treatment.x+age, data=dd)
lmbplanae<-lm(bpln~treatment.x+age+years_managerial_experience, data=dd)
lmbplanaeq<-lm(bpln~treatment.x+age+years_managerial_experience+sa_always_finishes_projects+sa_handles_cha
```

```
require(stargazer)
stargazer(lmbplan, lmbplana, lmbplanae, lmbplanaeq, font.size = 'tiny', model.numbers = FALSE, omit.stat=c
```

What if the sample size is smaller?

```
require(dplyr)
dds<-sample_n(dd,100)
```

```
lmbplan<-lm(bpln~treatment.x, data=dds)
lmbplana<-lm(bpln~treatment.x+age, data=dds)
lmbplanae<-lm(bpln~treatment.x+age+years_managerial_experience, data=dds)
lmbplanaeq<-lm(bpln~treatment.x+age+years_managerial_experience+sa_always_finishes_projects+sa_handles_cha
```

Table 16:

	Dependent variable:			
	bpln			
treatment.x	0.126***	0.125***	0.128***	0.125***
	(0.017)	(0.017)	(0.017)	(0.017)
age		−0.001	−0.003***	−0.003***
		(0.001)	(0.001)	(0.001)
years_managerial_experience			0.009***	0.009***
			(0.001)	(0.001)
sa_always_finishes_projects				0.020
				(0.017)
sa_handles_challenges				−0.003
				(0.017)
sa_finds_a_way				−0.024*
				(0.014)
sa_has_clear_goals				−0.037***
				(0.010)
Constant	0.365***	0.418***	0.461***	0.525***
	(0.012)	(0.039)	(0.039)	(0.043)
Observations	3,449	3,449	3,449	3,449
Adjusted R^2	0.015	0.015	0.028	0.035
Note:				*p<0.1; **p<0.05; ***p<0.01

```
require(stargazer)
stargazer(lmbplan, lmbplana, lmbplanae, lmbplanaeq,font.size = 'tiny',  model.numbers = FALSE, omit.stat=c
```

Table 17:

	Dependent variable:			
	bpln			
treatment.x	0.136	0.151	0.156	0.187*
	(0.100)	(0.100)	(0.099)	(0.102)
age		−0.006	−0.010*	−0.007
		(0.005)	(0.006)	(0.006)
years_managerial_experience			0.012*	0.006
			(0.007)	(0.007)
sa_always_finishes_projects				0.040
				(0.111)
sa_handles_challenges				−0.174
				(0.123)
sa_finds_a_way				−0.033
				(0.054)
sa_has_clear_goals				−0.072
				(0.062)
Constant	0.385***	0.642***	0.764***	0.994***
	(0.069)	(0.233)	(0.242)	(0.273)
Observations	100	100	100	100
Adjusted R^2	0.009	0.012	0.030	0.068
Note:				*p<0.1; **p<0.05; ***p<0.01

7.1.4 Subgroup Analysis

```
lmbplan1<-lm(bpln~treatment.x, data=subset(dd, site==1))#Philadelphia
lmbplan2<-lm(bpln~treatment.x, data=subset(dd, site==2))#Pittsburgh
lmbplan3<-lm(bpln~treatment.x, data=subset(dd, site==3))#Minneapolis-St. Paul
lmbplan4<-lm(bpln~treatment.x, data=subset(dd, site==4))#Duluth
lmbplan5<-lm(bpln~treatment.x, data=subset(dd, site==5))#Maine
stargazer(lmbplan1,lmbplan2,lmbplan3,lmbplan4,lmbplan5,font.size = 'tiny',  model.numbers = FALSE, omit.st
```

Table 18:

	Dependent variable:				
	Bussines Plan				
	Philadelphia	Pittsburgh	Minneapolis-St. Paul	Duluth	Maine
treatment.x	0.172***	0.107**	0.125***	0.131	0.056
	(0.034)	(0.044)	(0.028)	(0.081)	(0.046)
Constant	0.370***	0.339***	0.369***	0.326***	0.387***
	(0.024)	(0.031)	(0.020)	(0.056)	(0.031)
Observations	904	482	1,383	167	513
Adjusted R^2	0.027	0.010	0.014	0.010	0.001
Note:					*p<0.1; **p<0.05; ***p<0.01

7.1.5 Interaction Term

```
lmbplan<-lm(bpln~treatment.x, data=dd)
lmbplanint<-lm(bpln~treatment.x*currently_receiving_ui_benefits, data=dd)
stargazer(lmbplan,lmbplanint,font.size = 'tiny',  model.numbers = FALSE, omit.stat=c("ll","f","res.dev","s
```

Table 19:

	Dependent variable:	
	Bussines Plan	
treatment.x	0.126***	0.105***
	(0.017)	(0.020)
currently_receiving_ui_benefits		−0.007
		(0.017)
treatment.x:currently_receiving_ui_benefits		0.058**
		(0.026)
Constant	0.365***	0.368***
	(0.012)	(0.014)
Observations	3,449	3,449
Adjusted R^2	0.015	0.016
Note:		*p<0.1; **p<0.05; ***p<0.01

7.1.6 Multiple Observations

```
require(sas7bdat)
w2<-read.sas7bdat("GATE/wave2.sas7bdat")
w3<-read.sas7bdat("GATE/wave3.sas7bdat")
```

```
w1$wave<-1
w2$wave<-2
w3$wave<-3
zz<-rbind(w1[,c("gateid","treatment", "bpln","wave")], w2[,c("gateid","treatment", "bpln","wave")])
panel<-rbind(zz[,c("gateid","treatment", "bpln","wave")], w3[,c("gateid","treatment", "bpln","wave")])

lmbplanpan<-lm(bpln~treatment+as.factor(wave),data=panel)
summary(lmbplanpan)

##
## Call:
## lm(formula = bpln ~ treatment + as.factor(wave), data = panel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4650 -0.3664 -0.2944  0.6070  0.7246
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.39302    0.00986  39.859  < 2e-16 ***
## treatment          0.07194    0.01035   6.949 3.94e-12 ***
## as.factor(wave)2 -0.11766    0.01217  -9.667  < 2e-16 ***
## as.factor(wave)3 -0.09859    0.01293  -7.628 2.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4892 on 8933 degrees of freedom
## Multiple R-squared:  0.01704,Adjusted R-squared:  0.01671
## F-statistic: 51.63 on 3 and 8933 DF,  p-value: < 2.2e-16
```

7.1.7 Beyond Average Effects

- **Quantile Regression.**
  - **Advantages:** Allows to look at the effect on the whole distribution of outcomes, robust to outliers
  - **Disadvantages:** Ineffective

  See Koenker, R. Hallock, K.F., 2001. Quantile Regression. Journal of Economic Perspectives, 15(4), pp.143156. Available at: http://pubs.aeaweb.org/doi/abs/10.1257/jep.15.4.143.
- **Machine Learning Methods**
  - **Advantages:** Allows to discover hetergenous treatment effect without the problem of multiple-hypothesis testing
  - **Disadvantages:** Ex-post rationalization

  See Athey, S. Imbens, G., 2015. Machine Learning Methods for Estimating Heterogeneous Causal Effects. , pp.19. Available at: http://arxiv.org/abs/1504.01132.

7.2 Corrections

7.2.1 Partial Compliance

- We would like to know what would be the effect if everyone would take the program or drop-out rate will be equal across different subgroups.
- Wald estimator

  Impact on Compliers= ITT/ (Take-up Treatment - Take-up in Comparison)

  Assumption: The difference is attributed to the additional people who take up program in the treatment group.
- Use Instrumental Variable Approach e.g. ivreg()

Example: We can estimate the the probability that the person assigned to treatment will attend the course (c) condition on the distance D to the training center.

First stage: $c_i = \gamma_0 + \gamma_T T_i + \gamma_D D_i + \gamma_T D(T * D)_i + \epsilon_i$

Then we use our prediction to estimate the the chance that the person will write the business plan (B)

Second stage: $B_i = \beta_0 + \beta_c c_i + \beta_D D_i + \epsilon_i$

7.2.2 Attrition

How to "deal" with attrition in analysis:

1. Determine the overall attrition rate
2. Check for differential attrition – difference in attrition rate between treatment and comparison group.
3. Determine the range of the estimated impact given the attrition:
   (a) Using model-based approach e.g. Heckman Selection Procedure
   (b) Use Bounds e.g. Manski-Horowitz or Lee trimming

   See for detailed discussion: Guido W.. Imbens, and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015.

7.2.3 Spillover

To account for spillovers:

1. We can randomize at different levels and use the difference in exposure to treatment to estimate effect of spillover.
2. We have to (1) have a theory how spillovers occur and (2) the non-spillover comparison group.

Example

For instance, if we assume that spillovers depends on distance we can estimate the next regression:

$$Y_i = \beta_0 + \beta_T T_i + \gamma_N N_i + \gamma_{N^T} N_i^T + \epsilon_i$$

,where T_i- Treatment, N_i-Number of people within certain range next to individual i, N_i^T-Number of treated people within certain range next to individual i.

7.2.4 Group Level Randomization

We have to correct for group specific effects and calculate clustered standard errors.

Example

We have to estimate the next regression:

$$Y_{i,j} = \beta_0 + \beta_T T_i + \gamma_C C_i + v_j + \epsilon_i$$

, where v_j - group-level error for group j.

7.2.5 When we used balancing

- If we used **stratification** we have to include dummies for each stratum e.g. age, gender.
- If we used **min-max t-statistic** we have to include variables that we used for balancing as a controls
- If we used **pairwise matching**, we have to include dummies for pairs.

7.2.6 Multiple Outcomes

The problem: Multiple Hypothesis Testing.

What can we do?

- Select key indicator in advance.
- Collapse many indicators into one testable.
- Adjust confidence intervals e.g. Bonferroni correction.

7.3 Pre-analysis Plan

The problem: The data mining problem

What can we do?

- Make pre-analysis plan **to restrict your analysis**

Drawback: The analysis is restricted + limited upside since replication of the study drastically decrease the chance of finding false-positive due to the data-mining

7.4 Summary

- Check for errors prior to analysis
- Provide intention- to- treat estimates
- Include covariates to achieve more precise estimates
- Provide subgroup analysis
- Go Beyond average effects
- Consider corrections due to:
  - Partial compliance
  - Attrition
  - Spillovers
  - Group-level randomization
  - Multiple outcomes
- Make pre-analysis plan

7.5 Exercises

1. **Prior to Analysis I:** What shall you do prior to analysis of the data?
2. **Basic Analysis II:** What is intention to treat estimate?
3. **Basic Analysis III:**How to select control variables?
4. **Basic Analysis IV:** Why to provide subgroup analysis?

5. **Basic Analysis V:** You study in a randomized control trial if response to request about donation vary across people of different ethnicity: In the treatment group you emphasize the importance of donation for others; In the control group you just ask to donate. You provide statistical analysis of the whole sample and do not find any difference between control and treatment group. When, however, you provide an subgroup analysis of 10 ethnic groups, you find that Ashkenazi Jews are less likely to donate if they were treated and you can reject the null-hypothesis at 5% level providing simple t-test for this group (p-value= 0.0467). What will be your conclusion and how would you summarize the results?

6. **Basic Analysis VI:** Interaction term
   - What is interaction term?
   - How do you construct an interaction term?
   - When does interaction matters? Give some example.
   - You estimate the following model:
     $\hat{Testscore}_i = 5 + 8 * Treat + 3 * Female + 2 * Female * Treat$
     - What is predicted test score of untreated female person?
     - What is predicted test score of treated female person?

7. **Corrections I:** What can you do to estimate the effect of treatment in presence of partial compliance? How to estimate the average effect treatment on compliers?
8. **Corrections II:** How to deal with attrition in statistical analysis?
9. **Corrections III:** How to deal with spillovers in the analysis of your data?
10. **Corrections IV:** If you have used ... for balancing how should you provide the analysis?
    - Stratification
    - Min-max t-statistic

- Pairwise matching

11. **Corrections V:** What can be the problem if you measure multiple outcomes? How to deal with it?

12. **Presentations:**
    Olken, Benjamin A. "Promises and perils of pre-analysis plans." The Journal of Economic Perspectives 29.3 (2015): 61-80.
    Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." Journal of Economic Perspectives, 29(3): 81-98.

8 Policy

8.1 Checklist of Common Errors

I. Errors in Design:

- Ignoring spillovers
- Underpowered study:
  - Too small sample size
  - Not clustered standard errors

II. Errors in Implementation

- Unreliable outcome measures e.g. people report change in their behavior but do not change it.
- Collecting data in treatment and comparison group in different manner
- Allowing high level of attrition
- Failing to monitor and document compliance and other threats

II. Mistakes in Analysis

- Dropping noncompliers
- Dropping matched pairs
- Non-Clustered standard errors
- Ignoring the precision estimates e.g. large confidence interval can be uninformative

8.2 Generalizability

- The commonality of generalizability issue
  - Small Non randomized Evaluations
    Problems:
    * Potentiallly Biased Estimator → low internal validity → low external validity
    * Specific context → low external validity
  - National or international non randomized evaluations
    Problem: No causation without manipulation
    Example: High trust level ∼ High Rate of investment (Knack and Keefer, 1997; Willinger et al., 2003; Bohnet et al., 2010 ; Felli et al., 2010;Bottazzi et al., 2011)
    Does it mean that if we increase trust we will increase investment rate?
    → Can be, but not necessary. High level of investment can lead to high trust level.

Table 20: Estimated effect of reducing class size by 7.5 students based on observational and experimental data.

	Estimated Effect
Massachusetts	0.12**
	(0.03)
California	0.14**
	(0.05)
STAR	0.19**
	(0.05)

- Designing randomized evaluations with an eye toward generalizability
- Combining information from randomized and non randomized studies
- Testing whether the results generalize
- Combining testing and theory to asses generalizability

8.3 Comparative Cost-effectiveness Analysis

I. Cost-benefit versus cost-effectiveness analysis.

- Cost-benefit analysis incorporates valuations of multiple outcomes making it possible to argue that program is worth investment.

  But! Value of benefits depends on assumptions.

  Example: Life Value in the US: $9.1 million (Environmental Protection Agency, 2010), $7.9 million (Food and Drug Administration, 2010), $6 million (Transportation Department, 2010),
- Cost-effectiveness analysis leaves the relative valuation of different outcomes up to user

II. Issues to consider when perform cost-effectiveness analysis

- Being comprehensive and consistent about costs across studies
  - Beneficiary costs
  - Transfer costs
- Using discount rates to compare costs and impacts across time
- Compare costs across countries e.g. Purchasing power parity
- Accounting for multiple outcomes

III. Sensitivity Analysis of Cost-effectiveness estimates

- The imprecision of impact estimates
- Sensitivity to changes in context e.g. percentage change, adjust to population density
- Sensitivity to changes in scale

8.4 From Research to Policy Action

Factors to consider when translating research into policy:

- Scaling up discrete packages versus applying general lessons to policy

- Working up with large organizations, including governments, versus small organization

Bridging the gap between research and policy:

- Choose a policy-relevant question
- Feed evidence at the right time
- Transfer knowledge about implementation
- Report evaluation results in a central location:
  - J-pal: www.povertyactionlab.org/evaluation
  - World Bank: www.worldbank.com/dime
  - Innovation Growth Lab: http://www.innovationgrowthlab.org/igl-database-map
  - ...
- Disseminate results in an accessible format
- Thank hard about generalizabilty
- Synthesize general lessons from multiple programs

8.5 Summary

- Check for common mistakes in design, implementation, analysis
- Consider Generalizability.
- Provide cost-effectiveness analysis.
- Communicate you research!

8.6 Exercises

1. **Checklist of Common Errors:** What can be errors in design, in implimentation, in analysis?
2. **Generalizability:** Manski and Garfinkel (1992) argue that *"there is, at present, no basis for the popular belief that extrapolation from social experiments is less problematic than extrapolation from observational data."* Do you agree with this statement? Why? What one can do with this?
3. **Comparative Cost-effectiveness Analysis I:** What are the advantages and disadvantages of cost-effectivness analysis and cost-benefit analysis?
4. **Comparative Cost-effectiveness Analysis II:** You help the CEO of internet-shop company to analyze the effectivness of discount program. Would you provide cost-effectivness analysis and/or cost-benefit analysis?
5. **From Research to Policy Action:** Frances and Gordon (Nature, 1993) provide a psychological experiment and find that letting kids listening to Mozart's Sonata for Two Pianos in D Major K. 448 increases their performance in spatial-temporal tasks.

   In January 13, 1998, Zell Miller, governor of Georgia, propose that state budget would include $105,000 a year to provide every child born in Georgia with a tape or CD of this music piece. Suppose you have been involved in evaluating in randomized control trial the effect of this program on kids' performance in spatial-temporal tasks and find a positive effect.

   What would be your policy conclusions?
6. **Example of Exam Questions I:** What are the assumptions of Before and After Comparison?

- Stability of Environment
- Exogeniety, $cor(Z, u) = 0$
- Stability of subjects characteristics
- No rebound effect
- Outcomes and 'treatment' variables are 'identical'

1. **Example of Exam Questions II:** Here is the R output of estimation of effect of two treatments `star1small` and `star1regular+aide` compared to control:

```
##
## Call:
## lm(formula = score1 ~ star1, data = sample_n(STAR, 650))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -195.453  -68.609   -7.453   60.075  223.672
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1031.453      7.769 132.774  < 2e-16 ***
## star1small          35.005     11.513   3.041  0.00254 **
## star1regular+aide   22.875     11.052   2.070  0.03920 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.89 on 357 degrees of freedom
##   (290 observations deleted due to missingness)
## Multiple R-squared:  0.02659,Adjusted R-squared:  0.02114
## F-statistic: 4.877 on 2 and 357 DF,  p-value: 0.008139
```

   - What is the marginal effect of `star1small` treatment? _____
   - Is the `star1small` treatment more effective than `star1regular+aide`? (a) Yes; (b) No; (c) Unknown

9 Requirements

1. Exam.
   - **Example question (Exam, 24.04.16):** You run the randomized control trial to test if the negative income tax increase divorce rate. You fail to reject the null-hypothesis that there is **no difference** in divorce rate between control and treated group (p-value= 0.85). Write down your conclusion.
2. Seminar work (Pass/Fail). One has to fulfill the seminar work requirements to be eligible for the exam.
   (a) Active participation in the discussion during the seminars.
   (b) **Course Assignment:** Choose one of the three tracks before 3rd of May. Present your work before 12th of July.
   Track D. Design of the field experiment (max. 1 person in group):
      i. Come up with idea
      ii. Statement of the problem
      iii. Breief literature review
      iv. Design of the experiment
      v. Expected results
      vi. Make a presentation (15 min).

   Track R. Reproduction of the field experimental data analysis (max. 2 person in group):
      i. Get the data
      ii. Reproduce the data analysis.
      iii. Compare results with results of the paper
      iv. Make a presentation (15 min).

   Track M. Meta-analysis of field experiments (max. 3 person in group):

i. Choose interesting topic (not yet covered by meta-analysis)
ii. Provide systematic literature search (find all papers and collect them)
iii. Collect outcome measures and studies characteristics
iv. Provide analysis.
v. Make a presentation (15 min).

Useful information

Writing and Explaining Style

! Zinsser, William. On writing well. HarperCollins Publishers, 1991.

- Thomson, William. A guide for the young economist. MIT Press, 2001.
- Strunk, William. The elements of style. Penguin, 2007.
- MOOC. Writing in Science: http://online.stanford.edu/course/writing-in-the-sciences

Data for replication

i. Check **for paper in economics that use field experiment** from top journal that is interesting for you. For instance: American Economic Review, Econometric, Quarterly Journal of Economics, Journal of Monetary Economics, Review of Economic Studies.
ii. Check if the data available.
   A. Check the data on the site of the journal.
   B. Check on the site of **the corresponding author** of the paper.
   C. For Journals from *Elsevier* Publishing company you can check it here: https://datasearch.elsevier.com//
iii. If data is not available, ask yourself, how much this paper is interesting for you?
   - Not very much → 1.
   - Very much → Write **a very polite letter** to the corresponding author.
iv. Get the data and start replication of **all tables from** the paper.

Meta-analysis

- Read Vivalt, E., 2014. How much can we generalize from impact evaluation results?. New York University.
- Check aidgrade.com; http://www.aidgrade.org/wp-content/uploads/AidGrade-Process-Description.pdf; http://www.innovationgrowthlab.org/igl-database-map to search for not covered topics
- For specific reading: https://www.meta-analysis.com/pages/books.php

10 List of Papers

Please read 2-5 papers from the list to get an inspiration for your project.

[1] Kevin J. Boudreau and Karim R. Lakhani. Innovation Experiments : Researching Technical Advance , Knowledge Production and the Design of Supporting Institutions. 2015.

[2] Esther Duflo, Michael Kremer, and Jonathan Robinson. Nudging Farmers to Utilize Fertilizer: Theory and Experimental Evidence from Kenya. *American Economic Review*, 101(6):2350–2390, 2011.

[3] Ernst Fehr and Bettina Rockenbach. Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–40, mar 2003.

[4] Bruno S. Frey and Stephan Meier. Social Comparisons and Pro-social Behavior : Testing in a Field Experiment " Conditional Cooperation ". *The American Economic Review*, 94(5):1717–1722, 2004.

[5] Uri Gneezy and John a List. Putting behavioral economics to work: resting for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384, 2006.

[6] John A. List. Neoclassical Theory versus Prospect theory. 2003.

[7] John a. List. The Nature and Extent of Discrimination in the Marketplace: Evidence From the Field. *Quarterly Journal of Economics*, 119(1):49–89, 2004.

[8] John A List and David Lucking-Reiley. Demand reduction in multiunit auctions: Evidence from a sportscard field experiment. *The American Economic Review*, 90(4):961–972, 2000.

[9] Arno Riedl. Behavioral and Experimental Economics Do Inform Public Policy. *Public Finance Analysis*, 66(1):65–95, 2010.

[10] Joel Slemrod, M Blumenthal, and C Christian. Taxpayer response to an increased probability of audit: Results from a controlledexperiement in Minnesota. *Journal of Public Economics*, 79:455–483, 2001.